Rapid Protein Structure Detection and Assignment using Residual Dipolar Couplings

Michael A. Erdmann Gordon S. Rule

December 17, 2002

CMU-CS-02-195

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

M.A.E. is with the Department of Computer Science; G.S.R is with the Department of Biological Sciences.

This research was supported in part by Carnegie Mellon University, the Eberly Family Professorship in Structural Biology to G.S.R., and the Pennsylvania Department of Health through the grant "Integrated Protein Informatics for Cancer Research."

Abstract

Motivation: High-throughput structural proteomics requires fast robust algorithms for extracting protein structure from sparse experimental data. Current approaches are too slow. Determining the 3D structure of an unknown protein may require 6–12 months, mainly for data interpretation. Determining ligand induced *changes in structure* of a previously known protein may still require weeks of effort. This second problem is of great interest to drug designers, and is our main focus in this paper. A key step is the *resonance assignment problem*, in which observed NMR peaks must be matched to a protein's atoms.

Contributions: This paper describes two novel procedures, together called PEPMORPH, for inferring structure and assigning resonances: (1) A method for extracting combinatorial protein substructures directly from sparse NMR experiments; (2) A method for matching experimental to known substructures by exploiting the orientational constraint of residual dipolar coupling (RDC). PEPMORPH reverses the traditional approach, in which NMR resonances are assigned prior to structure determination. As a result, PEPMORPH increases the information available during assignment, speeding up the overall process.

Results: We have tested PEPMORPH on a variety of real proteins deposited in the Protein Data Base (PDB), using standard synthetic NMR data with a variety of noise levels, and on one protein (Rho130) using real ^{15}N NOESY data and synthetic RDC data. PEPMORPH assigns a very high fraction of the resonances correctly and flags those resonances that cannot be assigned uniquely because of significant structural change. PEPMORPH runs in $O(n^3)$ time, where n is the number of amino acids in the protein, requiring minutes for moderately sized (20–35kDa) proteins on a 1GHz PC.

Keywords: Protein structure, NMR, residual dipolar coupling, resonance assignment, structural homology.

1 Introduction

The *assignment problem* pervades structural proteomics. The problem is to establish a correspondence between experimentally observed data and known structural building blocks, in our case between observed nuclear magnetic resonances and specific protons in one or more proteins.

Of particular interest to us is the problem of studying conformational changes in known protein structures. For example, in drug-design, the three-dimensional structure of at least one conformation of a protein may be known. The drug designer has a suite of potential drugs, possibly thousands. Each is allowed to bind to the protein, possibly causing a change in the three-dimensional structure and thus the function of the protein. The designer wishes to probe each sample and determine the structure of the resulting protein-drug complex quickly.

There are two basic methods for probing protein structures: Nuclear Magnetic Resonance (NMR) and X-Ray Crystallography. The advantage of NMR is its ability to probe proteins in solution. The advantage of X-ray crystallography is its high accuracy; the disadvantage is its requirement for crystallized structures. These can be difficult to obtain, sometimes requiring months of effort. X-ray is a useful tool for determining structures of wholly unknown proteins, whereas NMR is an essential tool for determining conformational changes in proteins due to protein-protein or protein-ligand interactions (Hajduk et al. 1997).

Our thesis is that proteins reveal much of their three-dimensional structure through two very simple NMR experiments: (i) Measurements of amide-amide proximities from NOESY experiments and (ii) measurements of peptide plane orientations from residual dipolar couplings (RDCs). The NOESY experiment reveals both the interconnectivity of the amino acids and the inherent local dimensionality of substructures of the protein. The RDC experiment provides orientational hash values for distinguishing geometrically similar yet distinct substructures.

We next outline our basic methods and approach. We then review relevant NMR experiments along with related work. A reader new to NMR may wish to interleave the reading of Section 2 with this one. Finally, we describe our approach in detail, then conclude with results.

Our Approach

PEPMORPH models both the experimental NMR data and the known protein as graphs. The graph vertices represent amide protons along with their peptide plane orientations; the graph edges indicate spatial proximity. PEPMORPH matches the two graphs, trying locally to maximize the number of coincident edges while minimizing the orientational differences of the peptide planes. In order to avoid full-blown subgraph isomorphisms, PEPMORPH first extracts structural information from each of the graphs. The structures reflect the natural local dimensionality of the protein. Thus PEPMORPH decomposes the protein into linear, planar, and volumetric regions, represented by combinatorial *polytopes* of graph vertices (see Figure 1). Matching these polytopes is relatively easy, given their geometric simplicity and the constraints imposed by the peptide plane orientations.

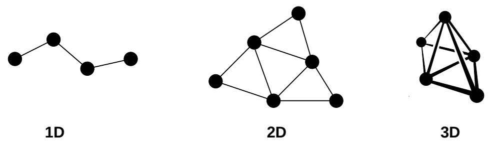


Figure 1: One-, two-, and three-dimensional polytopes. The solid spheres represent amide protons, the edges indicate spatial proximity.

Problem Setup Suppose the protein consists of n amino acids (also called *residues*), numbered $i = 1, \ldots, n$. The drug-designer allows the protein to bind to some target ligand, then probes the resulting complex using two experiments: an amide-amide NOESY experiment and a residual dipolar coupling experiment (see Section 2).[1]

Resonance Data NMR spectra used by PEPMORPH have a basis set of n *resonances*, $\Omega = \{\omega_1, \ldots, \omega_n\}$, one for each amino acid in the protein. Generally these resonances are multi-dimensional, indicating the involvement of multiple nuclei in the NMR spin transfer. Experimentally, some resonances will be missing. For instance, prolines do not have amide protons and thus do not show up in the NMR spectrum. Noise and degeneracy may also lead to missing resonances. A fraction of amino acids may have more than one resonance due to the presence of multiple conformations. Nonetheless, by adjoining special values to our basis set, we can model these spectra as sets of n resonances.

NOESY Data The data returned by a NOESY experiment consists of a set of *crosspeak distances* $\mathcal{D}=\{d_{\omega\gamma}\}$, representing rough separations of those amide protons that lie within approximately 5Å of each other. Each crosspeak is indexed by two resonances, ω and γ in Ω, representing the two N-H pairs whose spins generate the crosspeak. Experimentally, some crosspeaks may be missing.

RDC Data The data $\mathcal{R}$ returned by the residual dipolar coupling experiment associates two angles, $({}^{\alpha}\theta_{\omega}, {}^{o}\theta_{\omega})$, with each resonance ω in Ω. These are the angles that the N-C_α and N-C(O) bond vectors make with the magnetic field axis. Again, some angles may be missing experimentally.

We can now formulate

The Assignment Problem:

Compute i from ω_i, for $i = 1, ..., n$, given Ω, $\mathcal{D}$, and $\mathcal{R}$.

PEPMORPH creates an experimental graph G_e to represent the experimental data. The vertices are the basis resonances, labeled with their RDC angles; the edges are the NOESY crosspeaks, labeled with their distances. PEPMORPH creates an analogous graph G_k from a known 3D structure of the protein.

PEPMORPH matches the two graphs G_e and G_k, thereby creating a potential solution to the Assignment Problem. PEPMORPH has three phases:

[1] Only the protein is made NMR-sensitive, not the ligand.

1. First, PEPMORPH decomposes the two graphs G_e and G_k into *polytopes*. Polytopes are transitive closures of 1D, 2D, and 3D simplices. Each simplex is purely combinatorial, formed from the *unlabeled* graph edges. The closure relation is subface-connectivity. Although combinatorial in nature, polytopes capture the local geometry of the protein, and thus define a natural abstraction of the protein fold. This phase runs in time $O(n)$, where n is the number of amino acids in the protein. The reason for the linear time complexity is that the number of crosspeaks per amino acid is bounded by a constant, due to steric constraints.

2. Next, PEPMORPH tries to infer the direction of the magnetic axis used during data collection *as it would appear* in the coordinate frame of the known protein. PEPMORPH accomplishes this by matching experimental to known polytopes. Because of the combinatorial/geometric structure of polytopes, matching polytopes does not involve any general subgraph isomorphisms. This phase runs in time $O(n^3)$, arising from the need to compare pairs of linear-sized polytopes. The output of this phase is an inferred magnetic axis, along with a matching of many of the resonances $\{\omega_i\}$ to their generating residues $\{i\}$.

3. Finally, PEPMORPH extends the assignments found in phase 2, by a combination of techniques. One of these involves embedding the unknown protein into 3D. Another involves performing small minimum cost bipartite graph matchings on subsets of the protein. The time complexity of this phase is $O(n^3)$, because of the embedding complexity and the bipartite matching complexity.

We have tested PEPMORPH on a number of proteins, with varying degrees of noise. For high noise, the polytopes produced in Phase 1 may degenerate to single or double simplices of varying dimensions. Nonetheless, the approach continues to infer magnetic axes well and solve the assignment problem with graceful degradation as noise increases. For very high noise we have implemented additional variations, that optimize assignments over a space of likely magnetic axes.

2 NMR Overview

There are dozens of different NMR experiments (Wüthrich 1986, Cavanagh et al. 1996). At their core these experiments report the proximity of (NMR-sensitive) atoms that are close to each other, either atoms that are separated by specific bonds or atoms that are spatially close. A classic through-bond experiment is the HNCA *J-correlation* which reports correlations (also called *crosspeaks*) between the amide group of one amino acid and both its own alpha carbon and the alpha carbon of the preceding amino acid. This experiment provides inter-residue chemical shifts (resonances) useful for establishing structural connectivities along the backbone. A classic through-space experiment is the NHHN Nuclear Overhauser Enhancement Spectroscopy (NOESY), which reports correlations between pairs of protons, each attached to a backbone nitrogen. For typical 600–750MHz spectrometers, this experiment is sensitive to protons within 5–6Å of each other. For short-range correlations (protons less than 3.5Å apart), the crosspeak intensity provides a good estimate of the inter-proton distance. This estimate becomes increasingly less accurate at greater separations.

In addition to these backbone experiments, numerous other experiments measure correlations within and across the sidechains projecting off the backbone. A full-blown structure determination might use J-correlated spectroscopy to assign backbone and sidechain atoms, then use NOESY experiments to seed a distance geometry routine, refine the resulting structure, and repeat (Crippen and Havel 1988, Guntert 1998, Zimmerman and Montelione 1995, Nilges et al. 1997, Mumenthaler and Braun 1995, Mumenthaler et al. 1997). Automation of the assignment process is exemplified by the programs AUTOASSIGN (Zimmerman et al. 1997) and PASTA (Leutner et al. 1998).

For large proteins, rapid proton-proton relaxation of the excited spins leads to broad overlapping lines with poor NMR signal intensity. Consequently, programs such as AUTOASSIGN and PASTA work well for smaller proteins (200 residues), but not beyond. For larger proteins, one purposefully deuterates the protein's aliphatic protons to decrease the relaxation rate of the remaining amide protons, thus restoring their NMR-sensitivity (Torchia et al. 1988, Grzesiek et al. 1995, Gardner et al. 1997). Since PEPMORPH only requires spectral information from amide protons, it works equally well on large deuterated proteins.

Dipolar Coupling

Dipolar coupling arises because of the influence of one nuclear spin's magnetic moment on another. Dipolar coupling can be detected in two ways, depending on the state of alignment of the protein:

1. NOESY. In isotropic media, proteins tumble quickly and randomly, without preferred orientation. Thus the dipolar coupling does not affect the nuclear resonance frequencies. Instead, a rotation-induced fluctuation of the local magnetic field causes relaxation of the two spins. This relaxation generates crosspeaks in a NOESY experiment, the intensities of which are proportional to $1/d^6$, where d is the separation between the spins. One version, the 4D ^{15}N NOESY, correlates an amide backbone proton $^1\mathrm{H}_N$ and its ^{15}N nitrogen with another $^1\mathrm{H}_N$ proton and its ^{15}N. Applied to deuterated proteins, this experiment is nicely robust, with almost all crosspeaks ($95+\%$) present and resolvable. Using earlier terminology, the pair of nuclei $(^1\mathrm{H}_N, ^{15}\mathrm{N})$ in amino acid i generate a 2-dimensional resonance ω_i. Two of these 2-dimensional resonances then index a NOESY crosspeak.

2. Residual Dipolar Coupling. In partially aligned samples, obtained by placing orienting media into solution with the protein, the dipolar coupling may be detected as an additional coupling between the spins. Examples pioneering this approach include the addition of phospholipid bicells (Tjandra and Bax 1997, Struppe and Vold 1998) or filamentous bacteriophage (Hansen et al. 1998). The magnitude of the resulting *residual dipolar coupling*, or *RDC*, depends on the relative orientation of the bond that connects the two atoms *and* on their inter-atomic distance. Specifically, the residual dipolar coupling of two interacting spin $\frac{1}{2}$ nuclei is:

$$R = -\mathcal{O}\frac{\gamma_1\gamma_2 h}{2\pi^2 d^3}\left(\frac{3\cos^2\theta - 1}{2}\right), \tag{1}$$

where γ_1 and γ_2 are the gyromagnetic ratios of the two nuclei, h is Plank's constant, d is the separation of the two nuclei, θ is the angle between the magnetic field axis and the inter-nuclear vector, and $\mathcal{O}$ is the extent of alignment of the protein molecules.

For nuclei with known separations d, residual dipolar couplings provide very accurate measurements of inter-nuclear orientations since one can measure the RDC induced change in NMR *frequency* with high precision. This approach can be used to measure the orientation of key bond vectors in a protein, such as the N-H, N-C_α, and N-C(O) vectors in each amino acid's peptide-plane (except for proline). PEPMORPH uses the N-C_α and N-C(O) bond vectors. The N-H vector lies in the plane of these two vectors; if noise is a significant issue this redundancy may be used to reduce noise. Three comments:

1. RDC values are indexed by resonances, just as are NOESY crosspeaks.
2. The residual dipolar coupling R may be used to infer an angle θ in the range $[0, \pi/2]$. Thus an RDC value constrains the bond vector to a two-sided cone making angle θ with the magnetic axis, and vice-versa.
3. Experimentally, in order to obtain θ from Equation (1) it is necessary to determine $\mathcal{O}$. This scaling factor is the same for all bond-vector types. One can therefore infer $\mathcal{O}$ from the entire distribution of RDCs in a protein (Clore et al. 1998).

Recently, several research groups have begun to extract structural information from residual dipolar couplings. Some researchers have used residual dipolar couplings to determine the relative orientation of two domains in a protein (Losonczi et al. 1999), others to build small protein structures *de novo* (Hus et al. 2001), yet others to constrain protein fold predictions (Moltke and Grzesiek 1999, Wedemeyer et al. 2002, Rohl and Baker 2001) or to recognize homologous protein folds (Annila et al. 1999). Several of these approaches rely on residual dipolar couplings of bond vectors outside the peptide plane. Others use alignments with multiple media to obtain constraints from several orientation estimates (Al-Hashimi et al. 2000).

The approaches above generally assume that the assignment problem has *already* been solved. Their programs expect matchings between measured resonances (and RDCs) and their generating nuclei. Our goal is to solve this prior problem.

3 Closely Related Work

Kraulis (1994) demonstrated the feasibility of solving the assignment problem for small proteins using a dense network of inter-proton distances coupled with molecular mechanics calculations. Kaptein's group (van Geerestein-Ujah et al. 1995) used graph theory to identify prescribed NOE connectivity patterns,[2] then sequentially connect residues. Donald recently reported a similar technique to compute assignments (Bailey-Kellogg et al. 2000). Both methods have origins in Wand's mainchain assignment strategy (Nelson et al. 1991). Kaptein's SERENDIPITY, Donald's JIGSAW, and our PEPMORPH share the goal of extracting secondary structure directly from sparse NMR data. One difference is that PEPMORPH does not rely on preconceived notions of secondary structure; it simply searches for simplicial clusters of NOEs.

Hus et al. (2002) have reported a method for solving the assignment problem for a known protein structure based on experimental RDCs and chemical shifts. They employed a minimum-cost bipartite graph matching algorithm whose cost function measured the squared difference

[2] "NOE" is often used as shorthand for "NOESY crosspeak".

between observed and known values. They assigned roughly 50% of the residues in ubiquitin using only RDCs, and roughly 90% using both RDCs and a limited number of inter-residue J-connectivities. They concluded that residual dipolar coupling alone is not sufficient to solve the assignment problem.

Extending this reasoning, for larger proteins it is likely that chemical shifts will provide less and less constraint, due to degeneracy. The thesis of our work is that a combination of local structural information from NOESY experiments and global orientational information from RDCs provides a basic foundation for solving the assignment problem that scales robustly to larger proteins.

4 Graph Representations and Problem Formulations

PEPMORPH creates two undirected graphs, each representing the amide protons of the protein, one capturing the experimental information, the other capturing the known structural information.

The *experimental graph* G_e is a labeled graph of the form $G_e = (\Omega, \mathcal{D}, \theta, d)$. The graph vertices are the resonances Ω; the graph edges are the observed crosspeaks $\mathcal{D}$. The function $\theta : \Omega \to [0, \pi/2]^2$ assigns to each vertex ω the pair of angles $({}^\alpha\theta_\omega, {}^o\theta_\omega)$ measured by the RDC experiment. (We can indicate missing values by adjoining special symbols to the range spaces.) The function $d : \mathcal{D} \to \Re$ assigns to each edge the distance estimate $d_{\omega\gamma}$ inferred from the NOESY crosspeak intensity.

The *known* graph G_k is also a labeled graph, of the form $G_k = (V, E, \Theta, D)$, with slightly different semantics. The graph vertices are the protein's amino acids, that is $V = \{1, \ldots, n\}$. The graph edges consist of all pairs of amino acids whose associated amide protons lie within 5Å of each other, that is, $E = \{(i, j) \in V \mid \parallel \mathrm{H}_i^N - \mathrm{H}_j^N \parallel \leq 5.0\text{Å}\}$. The edge labeling function $D : E \to \Re$ assigns these separations to the edges. The vertex labeling function Θ is slightly more complex than before. Ideally this function should report the angles between the known structure's N-C_α and N-C(O) bond vectors and the magnetic axis. The question is **"What is the magnetic axis?"** The magnetic axis used to collect data on the protein-ligand complex bears no obvious relationship to the coordinates of the known structure. Instead, PEPMORPH *must infer* the orientation of the magnetic axis relative to the known structure. Consequently, the vertex labeling function in G_k is an indexed function. Specifically, if $\mathbf{b}$ is a unit vector in $\Re^3$, we define $\Theta_\mathbf{b} : V \to [0, \pi/2]^2$ by $\Theta_\mathbf{b}(i) = ({}^\alpha\theta_{\mathbf{b}i}, {}^o\theta_{\mathbf{b}i})$, where ${}^\alpha\theta_{\mathbf{b}i}$ is the angle that the bond vector N-C_α at residue #i in the known structure makes with the *line* defined by $\mathbf{b}$, and ${}^o\theta_{\mathbf{b}i}$ is the angle that the bond vector N-C(O) makes with that line. The labeling function Θ is the collection of all possible $\{\Theta_\mathbf{b}\}$.

Given these definitions, one possible reformulation of the Assignment Problem is:

NMR Graph Matching Problem: Find a magnetic axis $\mathbf{b}$ and a one-to-one assignment function $a : \Omega \to V$ that optimizes the cost function:

$$\lambda \sum_{\omega,\gamma \in \Omega} |d(\omega, \gamma) - D(a(\omega), a(\gamma))|^2 \; + \; \mu \sum_{\omega \in \Omega} \parallel \theta(\omega) - \Theta_\mathbf{b}(a(\omega)) \parallel^2 .$$

Here the functions d and D have been extended to all potential graph edges; λ and μ are weighting factors (if $\lambda = 0$ this is an optimization over a family of min-cost bipartite graph matchings).

One difficulty with this formulation is that the NOESY distance function d is very noisy at distances above approximately 3.5Å. We thus prefer to ignore exact distances and instead focus primarily on the presence or absence of crosspeaks. We therefore reformulate the Assignment Problem further as the following optimization:

Assignment Optimization:

$$\min_{\mathbf{b} \in S^3, a \in \mathcal{A}} \sum_{\omega \in \Omega} \| \theta(\omega) - \Theta_{\mathbf{b}}(a(\omega)) \|^2, \tag{2}$$

where $\mathbf{b}$ is a unit vector as before, and $\mathcal{A}$ is some *admissible* collection of (possibly partial) assignment functions $a : \Omega \to V$, perhaps all assignments that are maximal subgraph isomorphisms of G_e and G_k, or all geometrically feasible assignments, etc.

PEPMORPH solves this optimization for embeddable subgraphs of G_e and G_k, namely combinatorial polytopes. The set of admissible assignments $\mathcal{A}$ is small for polytopes, permitting quick optimization. Once PEPMORPH has found assignments for subgraphs, the program then patches these together to create an overall assignment of G_e to G_k.

5 Phase I: Computing Polytopes

Given an undirected graph with vertices V and edges E, a combinatorial k-*simplex* is a set of $k+1$ vertices $v_1, \ldots v_{k+1}$, all in V, such that each pair (v_i, v_j) is an edge in E. k is the *dimension* of the simplex; the simplex embeds naturally into $\Re^k$. PEPMORPH finds all the 0-, 1-, 2- and 3-simplices of the two graphs G_e and G_k. For each graph the function also identifies all k-simplices that are not contained in any $(k+1)$-simplex, for $k = 0, 1, 2$. For example, 0-simplices in G_e that are not contained in any 1-simplex correspond to resonances for which there are no NOESY crosspeaks.

Two simplices of dimension k are said to be *adjacent* if they share a simplex of dimension $k-1$. Adjacency defines a symmetric relation whose transitive closure is an equivalence relation. Imagine partitioning a collection of k-simplices into equivalence classes by adjacency. Each class defines a *k-dimensional polytope*. Intuitively, if one embedded all the simplices into $\Re^k$, the polytopes would be maximal volumes of full dimensionality.

PEPMORPH finds all the 1-, 2-, and 3-dimensional polytopes of the graphs G_e and G_k, considering only those simplices that are not contained within higher-dimensional simplices. It is a remarkable property of proteins that *combinatorial* polytopes constructed using the 5Å cutoff of NMR experiments mirror the natural dimensionality of proteins. Specifically:

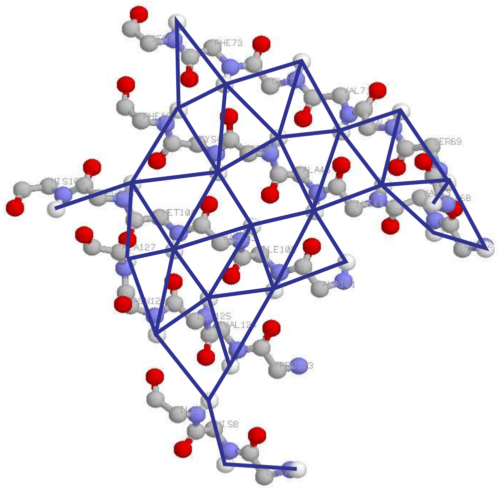
1. Loops in a protein tend to generate 1D polytopes (which we call *corals*).

2. β-sheets tend to generate 2D polytopes (which we call *surfaces*).

3. α-helices tend to generate 3D polytopes (which we call *volumes*).

To the best of our knowledge this simplicial nature of proteins has not been reported before.

PEPMORPH further infers helices from volumes by looking for sequences of at least four strong NOESY crosspeaks. With high probability such sequences form the backbones of helices (see



(a)

Type	Traditional Residues	Extracted Residues
Sheet	7–9, 123–127, 101–106, 42–48, 68–74	124–127, 100–106, 44–50, 69–74
Sheet	208–212, 182–188, 155–160, 224–230, 262–265, 281–283	208–212, 180–188, 154–163, 224–229, 262–265
Helix	17–32	17–33
Helix	53–66	53–68
Helix	89–99	88–99
Helix	111–119	110–120
Helix	135–149	136–151
Helix	167–179	167–179
Helix	196–205	198–207
Helix	215–220	215–219
Helix	(classified as two loops)	238–242
Helix	251–256	251–257
Helix	275–279	276–280
Helix	285–304	284–305

(b)

Figure 2: Polytopes of NOE connectivities, defined purely combinatorially without knowledge of the 3D structure, capture the natural local dimensionality of a protein. Thick lines in **Panel (a)** indicate amide-protons within 5Å of each other in the 5-strand sheet of 5AT1_A. **Panel (b)** compares the structures extracted by PepMorph from protein graphs (see Section 5) to traditional secondary structures in 5AT1_A.

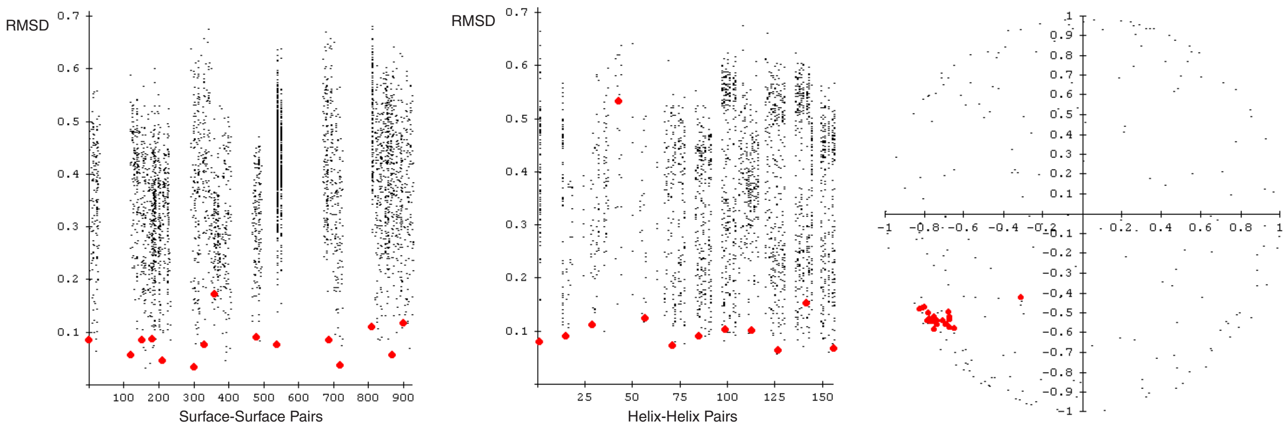
(Wüthrich 1986) for related comments). Panel (a) of Figure 2 depicts the backbone atoms of the 5-strand β-sheet found in the taut form of chain A of aspartate carbamoyltransferase (ATC) (pdb code: 5AT1). The thick lines connect amide protons that lie within 5Å of each other. One can see numerous triangles (2-simplices). When connected these triangles form a large 2D polytope that covers nearly the entire sheet. Panel (b) compares all the surfaces and volumes extracted by PepMorph with the sheets and helices of 5AT1_A.

6 Phase II: Inferring the Magnetic Axis

PepMorph infers the orientation of the experimental magnetic axis as it would appear relative to the known structure by matching simplices, polytopes, and/or inferred secondary structures between the experimental and known graphs. PepMorph will use whatever structures it is able to build, and thus its performance degrades gracefully with increased noise and degeneracy. Ideally, the drug designer will perform an HNCA experiment, revealing sequential backbone segments. At worst, degeneracy may reduce the size of polytopes to just one or two simplices each. PepMorph solves the optimization problem (2) for whatever substructures are available.

An important feature of optimization (2) is the set of admissible assignments $\mathcal{A}$. The specifics of $\mathcal{A}$ depend on the structures being matched:

- When comparing two 2-simplices, $\mathcal{A}$ consists of the six possible pairings of the underlying vertices.



(a) (b) (c)

Figure 3: Inferring the magnetic axis by comparing structures. **Panel (a)** depicts the RMSD of the cost function (2) for all possible surface-surface assignments. Tiny dots indicate RMSDs for all assignments; larger filled-in circles indicate RMSDs of correct assignments. **Panel (b)** shows all helix-helix assignments. **Panel (c)** depicts the x-y projection of all candidate magnetic axes (those whose RMSDs are among the best 5%). Axes corresponding to correct assignments are represented by larger filled-in circles; there is large cluster in the southwest corner.

- When comparing two 3-simplices, $\mathcal{A}$ consists of the twenty-four possible pairings of the underlying vertices.
- When comparing (short) corals, $\mathcal{A}$ consists of all possible linear alignments of subcorals.
- When comparing two simple surfaces[3], $\mathcal{A}$ consists of all possible assignments that combinatorially align the surfaces. These assignments are computed by considering all possible triangle-triangle pairings. The assignment function implied by a given triangle-triangle pairing is computed by expanding outward from that pairing, combinatorially aligning adjacent neighbor triangles, then aligning their neighbors and so forth. — The set $\mathcal{A}$ thus defined is significantly smaller than the set of all possible assignments, $O(n^2)$ vs. $O(n!)$, yet provides a basis for determining the structurally sound surface matchings.
- The definition of $\mathcal{A}$ for volumes is analogous to that for surfaces.
- When comparing inferred backbone segments (as when comparing helices or the results of an HNCA experiment), $\mathcal{A}$ consists of all possible linear alignments of those segments, along with directional ambiguity as appropriate.

For each assignment function in $\mathcal{A}$, PEPMORPH computes an optimal magnetic axis $\mathbf{b}$ using numerical optimization. The result is a set of magnetic axes, one for each possible assignment function, computed in $O(n^3)$ time. Figure 3 shows the results of comparing all surfaces (a) and helices (b) of 5AT1_A with all surfaces and helices of 8ATC_A (the relaxed conformation of ATC).

[3]A surface is *simple* if each edge bounds at most two triangles. The protein surfaces we have encountered thus far have nearly all been simple. PEPMORPH handles nonsimple surfaces by matching triangles of similar area.

Figure 3 indicates that correct assignment functions do indeed yield low costs, but not necessarily *the lowest* cost. PEPMORPH therefore clusters the resulting magnetic axes. Panel (c) of Figure 3 depicts all the magnetic axes produced by the polytope-level optimizations whose RMSDs are among the best 5%. There is a significant cluster of axes in the southwest corner. PEPMORPH computes the centroid of this cluster as its inferred magnetic axis $\mathbf{b}$. As intended, this axis is very near the correct magnetic axis. We note in passing that PEPMORPH deals with significant noise and degeneracy by retaining as needed all clusters of reasonable size.

7 Phase III: Computing the Assignments

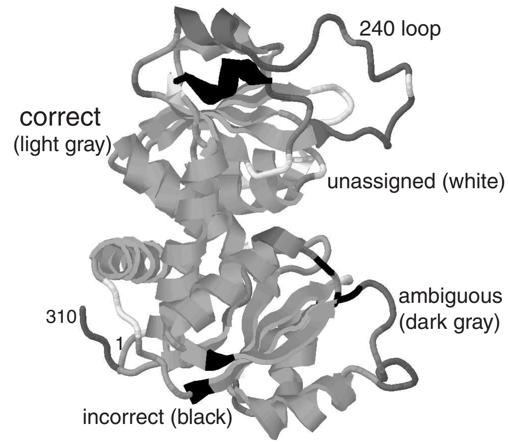
The output of PEPMORPH's Phase II is a magnetic axis that repeatedly appears as the optimum choice when matching substructures. Depending on the level of noise in the experimental data this magnetic axis may be unique or one of a small number of such axes. In Phase III, PEPMORPH repeats the matchings of Phase II, but now holds the magnetic axis fixed at its inferred direction. Generally for large structures this means that an alignment with the lowest RMSD will actually be the correct alignment. Thus PEPMORPH can rapidly assign a large number of the atoms in the secondary structures of the protein. In cases were noise and degeneracy are high, PEPMORPH may only be able to match structures consisting of one or two simplices. In this case there often are many incorrect assignments with low RMSD. Nonetheless, the correct assignments tend to cluster. PEPMORPH looks for large connected sets of mutually consistent assignments. Having found such large connected sets, PEPMORPH then extends the assignments further by looking at unassigned neighbors of assigned atoms, locally optimizing the assignment RMSD in these neighborhoods.

At this stage PEPMORPH has usually matched 50% to 75% of the resonances to their generating nuclei. The unassigned resonances and atoms cluster into small distinct components in the graphs G_e and G_k. PEPMORPH pairs up these components by considering neighboring already assigned atoms. PEPMORPH then focuses on each of the paired components. One of PEPMORPH's strategies is to embed the unknown protein into 3D by using the known structure and the existing assignments as a scaffold, then using distance geometry on the remaining unassigned atoms. Another strategy is to perform local bipartite graph matching. Sometimes this process fully assigns a pair of components. Often however these final steps are not precise enough to produce unique assignments, merely likely assignments. There is usually a good reason for this, either significant noise in the data or significant structural differences. In order to deal with degeneracy, noise, and unassignable structural differences, PEPMORPH actually returns a set-valued assignment function, meaning that resonances may be assigned to zero, one, or more amino acids.

Some of our trial results are reported in Figure 4. The NOESY data for Rho130 was based on a 3D experiment previously performed in the Rule Lab. Other examples are based on the PDB. In many cases, we injected synthetic noise. We used a quadratic error model for NOESY distances, as suggested by the results of (Briercheck and Rule 1998). We randomly deleted crosspeaks, usually on the order of 5%. For RDC data we injected noise either by (i) perturbing θ randomly ± 5 degrees or by (ii) perturbing the measured dipolar couplings randomly $\pm 10\%$ (representing angular errors of 4–15 degrees), as suggested in (Hus et al. 2002). This form of synthetic noise allowed us to explore a wide range of scenarios, ranging from graphs with full secondary structures to graphs whose polytopes consisted of one or two simplices.

Unknown → Known	ΔS	Noise	$\lvert$H$\rvert$	Assign	Good	Unique	Bad	$\Delta\mathbf{b}$	%
1ubq → 1ubq	-	cmap	72	72	72	72	0	0°	100
1ubq → 1ubq	-	synth	72	72	72	68	0	3°	100
Rho130 → Rho130	-	exper	118	111	111	111	0	0°	94
Rho130 → Rho130	-	both	118	109	109	109	0	1°	92
5at1_A → 8atc_A	2.2Å	cmap	298	293	286	231	7	6°	96
(*) 5at1_A → 8atc_A	2.2Å	synth	298	287	275	224	12	4°	92
5at1_A → 8atc_A	2.2Å	high	298	244	206	196	38	9°	69
1fpk_A → 1fpk_B	0.8Å	cmap	297	275	269	253	6	2°	91
1fpk_A → 1fpk_B	0.8Å	synth	297	273	248	244	25	2°	84
1ki4_A → 1ki7_A	0.4Å	cmap	277	259	259	258	0	7°	94
1ki4_A → 1ki7_A	0.4Å	synth	277	270	265	264	5	5°	96
4cts_A → 4cts_B	0.5Å	cmap	414	386	373	325	13	4°	90
4cts_A → 4cts_B	0.5Å	synth	414	321	278	280	43	3°	67

(a)



(b)

Figure 4: Results of assignment trials. **Panel (a):** $|\mathrm{H}|$ is the number of amide protons in the protein. *Assign* is the number of amide protons assigned by PEPMORPH (either uniquely or with multiple targets); *Good* is the number assigned correctly (meaning a target is correct); *Unique* is the number assigned to a unique target (correctly or incorrectly); *Bad* is the number assigned incorrectly. $\Delta\mathbf{b}$ is the angular difference between the inferred and correct magnetic axes; % is the ratio *Good*$/|\mathrm{H}|$ as a percentage. In the *Noise* column, *cmap* stands for *contact map*, meaning that PEPMORPH used rough separations (not exact distances) to compute polytopes, and used accurate RDC data; *exper* means experimentally obtained NOESY data and accurate RDC data; *synth* means standard synthetic errors in both NOESY and RDC data (see page 10); *both* means experimental NOEs and synthetic RDCs; *high* means synthetic errors high enough to destroy almost all secondary structure information in the experimental graph G_e. ΔS is the *structural difference* between the proteins being compared, as measured by the RMS difference between the 3D coordinates of corresponding amide protons.

Panel (b) depicts assignments from row (*), using shading to indicate correctness. *White* = unassigned (includes proline); *Light Gray* = unique and correct assignment; *Dark Gray* = assignment with multiple targets, including the correct one; *Black* = incorrect assignment.

Chain A of ATC is a fairly large protein, yet PEPMORPH assigns the central cores of both its domains very well. To the best of our knowledge, PEPMORPH is the first program able to assign proteins this large using sparse data. As stated in the review of previous work, existing assignment programs tend to fail beyond approximately 200 residues. Programs based on *sparse* NMR data and RDCs have so far been tested primarily on ubiquitin (1ubq), a very small protein.

In the comparisons of Rho130 to itself, due to prolines and missing NMR data the graph G_e actually contains 10 components, 7 of which are isolated resonances. The largest of the remaining components contains 98 resonances. PEPMORPH assigns all 98 resonances directly. It then assigns the remaining 13 resonances using focused bipartite graph matching.

The taut and relaxed forms of ATC are structurally different; in switching between conformations the two domains of chain A rotate about a hinge point. Despite significant motion PEPMORPH is able to assign nearly all the atoms of 5AT1_A to 8ATC_A. Two loops (ca. residues 75–85 and 230–245) cause some difficulty. In fact it turns out that the motion of at least one of these loops, the so-called *240 loop* is a key biochemical change in ATC's catalytic function. It makes no sense to assign the atoms of these loops based solely on NOESY and RDC data. PEPMORPH matches the loops correctly to each other, but does not assign individual atoms.

The 1FPK example compares the two halves of the dimer fructose-1,6-bisphosphatase to each other. The kinase example (1KI4 vs 1KI7) compares complexes of the thymidine kinase from herpes simplex with two different ligands. This would be a typical situation faced by a drug designer while screening potential drug compounds.

The first and last pairs of comparisons in Figure 4(a) provide a good conceptual bracket for the current competence of PEPMORPH. Ubiquitin (1UBQ) is a very small protein with distinct helix and sheet features. PEPMORPH works well on ubiquitin, as one would expect. Citrate Synthase (4CTS) is a large (ca. 45kDa) highly helical protein, giving rise to numerous substructures whose local connectivities and peptide orientations are effectively indistinguishable from one another. PEPMORPH performs quite well in comparing the two chains of 4CTS as high-resolution structures, then degrades gracefully but clearly in the presence of noise. This last example succinctly illustrates the current research frontier in automated interpretation of NMR.

8 Conclusions

This research seeks to understand the roles local structure and global orientation play in solving the assignment problem. We have exhibited a prototype program, PEPMORPH, that uses sparse NMR data in the form of backbone amide-amide NOESY crosspeaks and peptide plane RDCs to perform assignments. We have tested PEPMORPH on several proteins, using a variety of data sources (PDB, experimental, synthetic). Under reasonable noise conditions, PEPMORPH can assign upwards of 85% of the amide protons in a protein. Under perfect conditions, PEPMORPH will assign nearly 100% of the atoms, under extreme noise conditions it will still assign in excess of 50%.

PEPMORPH has direct applications in drug discovery. It can assist a drug designer in rapidly evaluating the structural changes of potential drug complexes.

A broader impact of this work may be on general protein structure determination. As the PDB is populated with proteins of known structure, structural homology provides the potential for mapping experimental data directly to structures by comparison with *all* proteins in the PDB. To date, such approaches have not been very successful at predicting wholly new proteins. Wanting is a compact representation of proteins that is easily probed by NMR or X-ray. The results of our research, as well as research surveyed in this paper, suggest that structural backbone connectivity coupled with peptide orientation constraints may be the basis for such a representation.

9 References

Al-Hashimi, H.M., Valafar, H., Terrell, M., Zartler, E.R., Eidsness, M.K., and Prestegard, J.H. (2000) Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *JMR*, 143:402–406.

Annila, A., Aitio, H., Thulin, E., and Drakenberg, T. (1999) Recognition of protein folds via dipolar couplings. *JBioNMR*, 14:223–230.

Bailey-Kellogg, C., Widge, A., Kelley, J.J., Berardi, M.J., Bushweller, J.H., and Donald, B.R. (2000) The NOESY Jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *JCB*, 7(3–4):537–558.

Briercheck, D.M. and Rule, G.S. (1998) Effect of deuteration on the accuracy of HN-HN distance constraints. *JMR*, 134:52–56.

Cavanagh, J., Fairbrother, W.J., Palmer, A.G. III, and Skelton, N.J. (1996) *Protein NMR Spectroscopy*. Academic Press, San Diego.

Clore, G.M., Gronenborn, A.M., and Bax, A. (1998) A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *JMR*, 133:216–21.

Crippen, G.M and Havel, T.F. (1988) *Distance Geometry and Molecular Conformation*. Research Studies Press, England.

Gardner, K.H., Rosen, M.K., and Kay, L.E. (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry*, 36:1389–401.

Grzesiek, S., Wingfield, P., Stahl, S., Kaufman, J.D., and Bax, A. (1995) Four-dimensional ^{15}N-separated NOESY of slowly tumbling perdeuterated ^{15}N-enriched proteins. Application to HIV-1 NEF. *J. Am. Chem. Soc.*, 117:9594–9595.

Guntert, P. (1998) Structure calculation of biological macromolecules from NMR data. *Quart Rev Biophysics*, 31:145–237.

Hajduk, P.J., Meadows, R.P., and Fesik, S.W. (1997) Discovering high-affinity ligands for proteins. *Science*, 278:497–499.

Hansen, M.R., Mueller, L., and Pardi, A. (1998) Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat Struct Biol*, 5:1065–74.

Hus, J.-C., Marion, D., and Blackledge, M. (2001) Determination of protein backbone structure using only residual dipolar couplings. *J Am Chem Soc*, 123:1541–1542.

Hus, J.-C., Prompers, J.J., and Brüschweiler, R. (2002) Assignment strategy for proteins with known structure. *JMR*, 157:119–123.

Kraulis, P.J. (1994) Protein three-dimensional structure determination and sequence-specific assignment of ^{13}C and ^{15}N-separated NOE data. A novel real-space ab initio approach. *JMB*, 243:696–718.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G., and Kessler, H. (1998) Automated backbone assignments of labeled proteins using the threshold accepting algorithm. *JBioNMR*, 11:31–43.

Losonczi, J.A., Andrec, M., Fischer, M.W., and Prestegard, J.H. (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. *JMR*, 138:334–342.

Moltke, S. and Grzesiek, S. (1999) Structural constraints from residual tensorial couplings in high resolution NMR without an explicit term for the alignment tensor. *JBioNMR*, 15:77–82.

Mumenthaler, C. and Braun, W. (1995) Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *JMB*, 254:465–80.

Mumenthaler, C., Guntert, P., Braun, W., and Wüthrich, K. (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *JBioNMR*, 10:351–62.

Nelson, S.J., Schneider, D.M., and Wand, A.J. (1991) Implementation of the main chain directed assignment strategy. Computer assisted approach. *Biophys J*, 59:1113–1122.

Nilges, M., Macias, M.J., O'Donoghue, S.I., and Oschkinat, H. (1997) Automated NOESY interpretation with ambiguous distance restraints: The refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *JMB*, 269:408–422.

Rohl, C.A. and Baker, D. (2001) De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc*, 124:2723–2729.

Struppe, J. and Vold, R.R. (1998) Dilute bicellar solutions for structural NMR work. *JMR*, 135:541–6.

Tjandra, N. and Bax, A. (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, 278:1111–4.

Torchia, D.A., Sparks, S.W., and Bax, A. (1988) Delineation of α-helical domains in deuterated staphylococcal nuclease by 2D NOE NMR spectroscopy. *J. Am. Chem. Soc.*, 110:2320–2321.

van Geerestein-Ujah, E.C., Slijper, M., Boelens, R., and Kaptein, R. (1995) Graph-theoretical assignment of secondary structure in multidimensional protein NMR spectra: application to the lac repressor headpiece. *JBioNMR*, 6:67–78.

Wedemeyer, W.J., Rohl, C.A., and Scheraga, H.A. (2002) Exact solutions for chemical bond orientations from residual dipolar couplings. *JBioNMR*, 22:137–151.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotkahara, S., Chien, C.-Y., Powers, R., and Montelione, G.T. (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. *JMB*, 4:592–610.

Zimmerman, D.E. and Montelione, G.T. (1995) Automated analysis of nuclear magnetic resonance assignments for proteins. *Curr Opin Struct Biol*, 5:664–73.