1-2-2005

High-Throughput Inference of Protein-Protein Interaction Sites from Unassigned NMR Data by Analyzing Arrangements Induced By Quadratic Forms on 3-Manifolds

Ramgopal R. Mettu
Dartmouth College

Ryan H. Lilien
Dartmouth College

Bruce Randall Donald
Dartmouth College

High-Throughput Inference of Protein-Protein Interaction Sites from *Unassigned* NMR Data by Analyzing Arrangements Induced By Quadratic Forms on 3-Manifolds

Ramgopal R. Mettu[*] **Ryan H. Lilien**[*,†] **Bruce Randall Donald**[*,‡,§,¶,‖]

January 2, 2005

Abstract

We cast the problem of identifying protein-protein interfaces, using only *unassigned* NMR spectra, into a geometric clustering problem. Identifying protein-protein interfaces is critical to understanding inter- and intra-cellular communication, and NMR allows the study of protein interaction in solution. However it is often the case that NMR studies of a protein complex are very time-consuming, mainly due to the bottleneck in *assigning* the chemical shifts, even if the apo structures of the constituent proteins are known. We study whether it is possible, in a high-throughput manner, to identify the interface region of a protein complex using only *unassigned* chemical shift and residual dipolar coupling (RDC) data.

We introduce a geometric optimization problem where we must cluster the cells in an arrangement on the boundary of a 3-manifold. The arrangement is induced by a spherical quadratic form, which in turn is parameterized by $SO(3) \times \mathbb{R}^2$. We show that this formalism derives directly from the physics of RDCs. We present an optimal algorithm for this problem that runs in $O(n^3 \log n)$ time for an n-residue protein. We then use this clustering algorithm as a subroutine in a practical algorithm for identifying the interface region of a protein complex from unassigned NMR data. We present the results of our algorithm on NMR data for 7 proteins from 5 protein complexes and show that our approach is useful for high-throughput applications in which we seek to rapidly identify the interface region of a protein complex.

A revised version of this paper has been accepted for publication and will appear at *ISMB* (2005) and *Bioinformatics* (2005).

DARTMOUTH COMPUTER SCIENCE TECHNICAL REPORT 2005-530
`http://www.cs.dartmouth.edu/reports/abstracts/TR2005-530/`

[*]Dartmouth Computer Science Department, Hanover, NH 03755, USA.
[†]Dartmouth Medical School, Hanover, NH 03755, USA.
[‡]Dartmouth Chemistry Department, Hanover, NH 03755, USA.
[§]Dartmouth Department of Biological Sciences, 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 the following grants to B.R.D.: National Institutes of Health (GM 65982), National Science Foundation (EIA-0305444 and EIA-9802068).

1 Introduction[1]

As the structural genomics initiative [36] rapidly populates the "space of protein structures," it becomes possible to understand the relationships and interactions between proteins, and further our knowledge of the molecular basis for many biological phenomena. Protein-protein interactions are very important and well-studied in structural biology. Recent advances in solution NMR spectroscopy allow us to directly study the interaction between proteins in solution; indeed, NMR spectroscopy is ideally suited to studying protein-ligand and protein-protein interactions [52]. Even given apo (or, unbound) structural models of the constituent proteins in a protein-protein complex (whose structure is unknown) obtained by either NMR or X-ray crystallography, a key bottleneck known as the *assignment* problem [19, 48, 2, 26, 1, 49, 50, 33, 35, 51] remains before we can make use of the recorded NMR spectra. That is, before we can make use of the NMR spectra, we must *assign* the NMR measurements to the nuclei that the measurements give information about. For example, *nuclear Overhauser effect* (NOE) NMR data provides interatomic distance restraints; in order for these distance restraints to be used in structure determination, we must first assign each restraint to pairs of nuclei in the protein. Current automated computational approaches to studying protein-protein interactions assume that the given NMR data has been assigned. These approaches typically use this NMR data, along with structural models of the constituent proteins, to generate the structure of the protein complex [11, 10, 8, 31]. The assignment process is typically done manually, and is time consuming. For example, the E1N-HPr complex required about 2 years of data analysis [7, 15] to obtain an accurate structural model. Automating the assignment process is an active area of research [26, 24, 32, 48, 53, 54](see [19] for a review of recent work). By avoiding the assignment problem, high-throughput determination of protein-protein interfaces given only *unassigned* NMR data would speed up all current approaches to generating the complex structure (via *docking*, see Section 1.1 below for further discussion). We show that without assignments, some accuracy is sacrificed in the determination of the protein interface, but there are enormous savings in time and cost, making it suitable for high-throughput applications. Furthermore, our approach of using only a *sparse* set of NMR data can be useful in the context of drug design, where a large number of protein-ligand pairs must be screened. Our algorithm uses experiments that require only ^{15}N-labeled samples that can be recorded in about a day of spectrometer time; ^{15}N-labeled samples require considerably less time and expense than ^{13}C samples to prepare. While manual approaches to determining the interface region may be more accurate (using a large suite of NMR spectra recorded for the apo and holo, or complex form, of the protein of interest), in applications such as drug design, a high-throughput algorithm (making use of sparse, unassigned NMR data) that trades some accuracy for time is often highly preferable to slower, data-intensive methods.

In this paper, we present an algorithm that uses the apo structure of a protein in a protein complex and a small number of unassigned NMR spectra to determine which residues are part of the interface region in the complex. By using unassigned NMR spectra we are able to remove the requirement that chemical shifts and NOEs be laboriously assigned to their corresponding atom in each protein. Our algorithm is designed to use an existing structural model of the protein, unassigned *chemical shifts* (i.e., HSQC peaks), amide exchange data, and unassigned *NH residual dipolar couplings* (NH RDCs), which give restraints on the orientation of the backbone NH bond vectors of a protein in solution [45]. Unlike previous work [3] which characterizes the geometry of protein interfaces, we do not assume that the crystal or solution structure of the complex has been solved. In fact, significantly more structures have been solved for proteins in their apo, or free form, rather than in their holo, or complexed form. This due to experimental, as well as inherent, limitations in the size of protein structures that can be solved by NMR or even X-ray crystallography. In practice, it is often more desirable to have a low false-positive rate at the expense of accuracy. Thus, for a protein A, the goal of our algorithm will be to describe the interface region in terms of both an *interaction zone* Z_A and an

[1]Abbreviations used: NMR, nuclear magnetic resonance; RDC, residual dipolar coupling; HSQC, heteronuclear single-quantum coherence; H^N, amide proton; NOE, nuclear Overhauser effect; SAR, structure activity relation; apo, free or unbound form of a protein in a protein complex; holo, bound or complexed form of a protein in protein complex; SVD, singular value decomposition.

interaction core C_A. We judge the performance of this pair (Z_A, C_A) by examining the accuracy of Z_A and the sensitivity (i.e., percentage of true positives) of C_A. Previous NMR techniques that have utilized prior apo structural information have either required that the experimental data be assigned [8, 31] or that multiple experiments utilizing selective isotopic labeling be performed [39]. We first consider a geometric version of the problem that asks us to cluster the cells of an *arrangement* on a 2-manifold. RDC data arises from the application of a quadratic form to the backbone NH bond vectors (i.e., to points in S^2). This quadratic form, together with the NH vectors induce the arrangement on the protein surface. We define a problem that asks for a subset of the arrangement that is the best candidate for the interface region. In Section 2.3, we give an algorithm that computes the optimal solution to this problem and runs in $O(n^3 \log n)$ time. Then, in Section 3, we give a more practical algorithm for solving this problem that runs in $O(nk^3 + n^3)$ time, where k is a parameter used to grid the the rotation space $SO(3)$ in order to estimate the alignment tensor (see Section 2). In the first phase of our algorithm we use a probabilistic approach to matching residues from the given structural model to the unassigned experimental RDCs; this phase identifies the interaction zone. Then, in the second phase, we use a practical version of our geometric clustering algorithm that, given a size threshold, identifies the interaction core. Instead of explicitly considering the arrangement induced by the protein surface and the given RDCs, this version of the clustering algorithm uses a discretized representation of the arrangement. In Section 4, we apply our algorithm to NMR data for 7 proteins, and show the interaction zones computed by our algorithm are accurate (i.e., identify a large percentage of the interface region), and that our computed interaction cores have high sensitivity (i.e., a very low percentage of false positives).

In this paper, our main contributions are:

1. To formalize the problem of finding a protein interface from *unassigned* NMR data as a geometric clustering problem, by exploiting the computational-geometric properties of RDC physics.
2. An optimal algorithm that runs in $O(n^3 \log n)$ for solving this geometric clustering problem.
3. A practical algorithm running in $O(nk^3 + n^3)$ time to identify the interface region of a protein given unassigned chemical shifts, unassigned RDCs and a structural model of the protein.
4. Testing of our practical algorithm on different combinations of real and simulated NMR data from 7 proteins that shows our algorithm could be useful in high-throughput applications.

1.1 Previous Work

As discussed above, protein-protein interactions are important for understanding many important biological phenomena. NMR allows for the study of proteins in solution, and is ideally suited, as well as widely used, to study protein-protein interactions (see, e.g., [52] for a survey). The majority of techniques to probe protein-protein interactions make use of *assigned* NMR data. Previous NMR techniques that use apo structural information require that the experimental data be assigned [8, 31] or that multiple experiments utilizing selective labeling be performed [39]. The key difference between our work and much of the previous work is that we only make use of *unassigned* NMR data, and seek only to identify the residues involved in the interface region without predicting [11, 10, 31] the structure of the complex. Although we do not compute the structure of the complex, the identified interface residues can be used in a number of ways. First, by running our algorithm on both proteins in the complex, it is possible to constrain the exhaustive searches over rotations and translations typically used in protein-protein docking algorithms. Furthermore, knowledge of the interface residues can be used to model "hot-spots" for mutation studies, or in drug design, where small molecules are identified (or built) to target interface residues in order to disrupt protein-protein interactions [28]. The goal of working with unassigned data is to minimize the need for expensive wetlab time and resources, and to facilitate high-throughput examination of various structural properties of proteins. Recent work [23, 25, 26, 24, 27, 14, 44, 34, 32, 53, 54, 17] makes use of unassigned NMR data, for example, in homology detection and fold determination.

A common approach to studying protein-protein interactions is to *dock* the proteins in the complex. That is, given structural information about the apo forms of the proteins, as well as assigned NMR experimental information such as orientational and distance restraints, docking algorithms [31, 11, 7, 8] compute the translation and rotation that brings the apo structures together to produce the complex structure. Docking protocols usually search over rotations and translations of one protein with respect to the other to minimize various energetic criteria, such as predicted van der Waals and electrostatic energies. Experimental data is used to constrain this search; that is, assigned RDC data and assigned intermolecular NOE data can be used to compute the desired rotation and translation, respectively. Most docking protocols use experimental restraints in conjunction with a grid search to identify the translation and rotation that both fits the experimental data and achieves a low-energy conformation for the resulting complex structure. In all of these techniques, the experimental NMR data must first be assigned. NOE data is particularly hard to assign due to chemical shift degeneracy [7, 8]. Kohlbacher *et al.* [22] score candidate dockings by comparing the unassigned experimental ^{1}H spectrum of the complex against a simulated spectrum for the candidate docking; however they not make use of experimental data to directly identify the interface region. Reese and Dötsch [39] use *selective isotopic labeling* to determine protein-protein interfaces with unassigned chemical shifts. Selective labeling allows a particular residue type to be more readily distinguished in NMR studies. With knowledge of the primary sequence, selective labeling experiments can be used to identify which residues, of a fixed amino acid type, have chemical shifts that are perturbed between the apo and holo form of the proteins in the complex. While this approach only uses unassigned NMR data, the amount of additional wetlab time required is dependent on the primary sequence of the protein, and thus varies from protein to protein.

Another ubiquitous technique in the study of protein-protein interfaces is called *chemical shift mapping* [52]. This technique was developed originally in the context of protein-ligand binding by Fesik and co-workers [43]; in that context, the technique is known as "structure-activity relationship by NMR" ("SAR by NMR"). The main idea for both protein-ligand and protein-protein interfaces, is to observe the change in HSQC spectra (see Section 1.2 below) for the free and complex spectra of the protein. In order to directly identify the interface region from chemical shift perturbations[2], the HSQC must be assigned. McCoy and Wyss [31] use assigned HSQC spectra to identify the interface region, and they use assigned RDCs to compute the relative rotation of the two proteins in the complex. Without assigned HSQC spectra, it is in fact possible, through titration experiments, to identify which (unassigned) HSQC peaks have shifted [38].

In contrast to docking approaches, our algorithm only finds the interface region of the given protein and not the complex structure. Furthermore, we use *unassigned* chemical shifts and RDCs. Compared to the work of [39] which uses selective labeling and unassigned NMR data, our approach is faster and cheaper since the amount of wetlab time is fixed and does not depend on the protein being studied. We do show, however, that selective labeling can optionally be used with our algorithm to improve the accuracy and sensitivity of the results (see Section 4).

1.2 Background

Solution NMR spectroscopy experiments give useful information about various biological and physical geometric properties of the protein being studied. Our algorithm uses experimental data from several high-throughput NMR techniques for the protein complex of interest; in this section, we discuss the information content of this data with respect to our algorithm.

Chemical shifts. Our algorithm uses ^{1}H-^{15}N *Heteronuclear Single Quantum Coherence spectroscopy* (2D HSQC) data [5, pages 411–447]. The HSQC data for a protein consists of a set of peaks which encode the resonant frequency of the amide atoms in each residue. These characteristic frequencies are also commonly

[2] Another NMR technique, *saturation transfer* [30], can also be used to identify which atoms experience changes in their local environment in the complex, but also requires assigned NMR data to directly identify the interface region.

referred to as *chemical shifts*; thus, amide HSQC data for a protein (ideally) is a set of pairs, one pair per residue (except for prolines and the N-terminus), that contain the chemical shifts of the amide proton and nitrogen. We will make use of the property that the chemical shift of a given nucleus changes as its local electronic environment changes. For example, a change in the chemical shift of a given atom between the apo and holo form of a protein indicate a change in the local electronic environment of that atom possibly due to binding or conformational change. Once the identity of these atoms (in the primary sequence) is known, chemical shift information can be useful in studying protein-ligand [43] and protein-protein [52] interactions (see Section 1.1). In this paper, we assume these identities are unknown, (i.e., *unassigned*), and treat the chemical shift peak for a given residue as a unique identifier that indexes into the experimental RDC data (described below). Our algorithm also uses NMR data from either *amide exchange* [13] or *water HSQC* [16, 18] experiments to identify which of the chemical shifts from the given HSQC spectrum is associated with surface, or solvent accessible, residues in the protein. The HSQC experiment together with these experiments to identify solvent accessible residues can be performed in less than a day of spectrometer time.

Residual Dipolar Couplings. Our algorithm also uses *residual dipolar coupling* (RDC) data [29, 41, 45]. Residual dipolar couplings give *global* orientational restraints on internuclear vectors. In this paper, we use NH RDCs, which give orientational information about backbone amide bond vectors. Each residual dipolar coupling D is a real number, where:

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

D_{max} is the dipolar interaction constant, $\mathbf{v}$ is the internuclear vector of interest with respect to an arbitrary substructure frame, and $\mathbf{S}$ is the 3×3 *Saupe* order matrix, or *alignment tensor*, which specifies the orientation of the protein in the laboratory frame (i.e, magnetic field in the NMR spectrometer). $\mathbf{S}$ is a symmetric, traceless, rank 2 tensor, that describes the average substructure alignment between the protein and the (alignment) medium [29]. We note that given a structural model, and the assignment of 5 or more of the recorded RDC values to their corresponding internuclear vectors in the model, it is possible to use SVD to reconstruct the alignment tensor $\mathbf{S}$ [29]. There are a number of techniques to estimate the alignment tensor given *unassigned* RDCs [23, 25, 26, 24, 27, 14, 44, 34, 32, 53, 54]. Many solutions may exist to Equation (1) for the internuclear vector $\mathbf{v}$ given an RDC value D and $\mathbf{S}$; however, given $\mathbf{v}$ and $\mathbf{S}$, we can *back-compute* or *simulate* D (modulo noise, dynamics, crystal contacts in the structural model etc.) in constant time. We note that the number of solutions to Equation (1) can be reduced by recording RDCs for multiple aligning media [46, 47]. Each medium (ideally) gives an unique alignment tensor, and thus for ℓ aligning media, we have ℓ equations for a given NH vector $\mathbf{v}$. The solutions to $\mathbf{v}$ must lie in the intersection of the solutions of these ℓ equations. The functional relationship given by Equation (1) between the recorded residual dipolar couplings and the corresponding internuclear vectors is a *quadratic form*; we note that the constant D_{max} can be folded into the matrix S to be consistent with the standard representation of a quadratic form. Like the HSQC experiment, RDCs can be recorded in about an hour of spectrometer time.

2 A Formal Problem Definition and Its Application

In this section, we formally define a geometric clustering problem on an arrangement on a 2-manifold, where the arrangement is induced by a spherical quadratic form. We first state the problem formally and then discuss its relevance and application to the problem of determining protein-protein interfaces given unassigned NMR data.

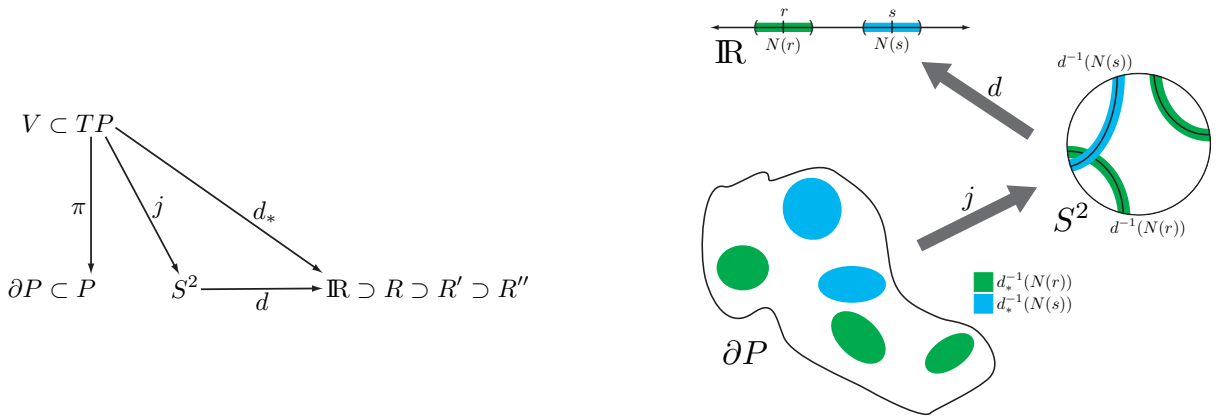

Figure 1: *Left:* Commutative diagram of the mappings used in our problem definition. *Right*: Our clustering problem in an arrangement with the set $R' = \{r, s\}$. Starting with the neighborhood of R', i.e., the intervals $N(r)$ and $N(s)$ in $\mathbb{R}$, we consider the set of orientations (contained in S^2) that are associated with these intervals. These orientations are $d^{-1}(N(r))$ and $d^{-1}(N(r))$, shown as colored green and blue bands, respectively, on the unit 2-sphere. By our definition of d_*^{-1} and $\mathcal{B}(V)$, these sets of orientations are mapped to patches on ∂P, denoted by the colored patches in the figure. Our optimization problem asks us to find the largest set of patches that does not exceed the given diameter threshold c_0.

2.1 An Arrangement Problem on 2-Manifolds

Let P be a semi-algebraic 3-manifold with boundary in $\mathbb{R}^3$ with constant degree, and let ∂P denote the boundary of P, which is a 2-manifold in $\mathbb{R}^3$. Let TP denote the tangent bundle of P; that is, $TP = \{(p, \mathbf{v}) \mid p \in P, \mathbf{v} \in T_pP\}$ where T_pP is the tangent space of $p \in P$. Let $V \subset TP$ be a finite set. Let $\mathcal{B}$ be the mapping $\mathcal{B}((p, \mathbf{v})) = ((p \oplus B_\delta) \cap P) \times (\mathbf{v} \oplus B_{\delta'})$, where B_δ and $B_{\delta'}$ are 3-dimensional balls of radius $\delta > 0$ and $\delta' > 0$, respectively, centered at the origin. Here, $\oplus$ denotes the Minkowski sum, i.e., for sets A and B, $A \oplus B = \{a + b \mid a \in A, b \in B\}$. Note that $\mathcal{B}(V)$ is an arrangement on ∂P.

Let $\pi : TP \to P$ be the map $\pi(p, \mathbf{v}) = p$. Let $d : S^2 \to \mathbb{R}$ be a quadratic form on S^2 with $d(\mathbf{v}) = \mathbf{v}^T\mathbf{S}\mathbf{v}$, where $\mathbf{S}$ is a symmetric, traceless tensor of rank 2. Let $j : TP \setminus 0 \to S^2$ be the map $j(p, \mathbf{v}) = \frac{\mathbf{v}}{\|\mathbf{v}\|}$, where 0 is the zero section of TP. (Remark: The *zero section* of a tangent bundle is simply the set of all elements $(p, \mathbf{v})$ with $\|\mathbf{v}\| = 0$). Let $d_* : TP \setminus 0 \to \mathbb{R}$ be a quadratic form on $TP \setminus 0$ with $d_*(\mathbf{v}) = d(j(p, \mathbf{v}))$; we note that d_* is the *lifting* of d by j. Figure 1(left) gives a commutative diagram of the mappings π, j, d, and d_*. Let the *cost* of $X \subseteq TP \setminus 0$ be defined as

$$c(X) \quad = \quad \max_{x,y \in X} \rho(\pi(x), \pi(y)),$$

where $\rho(p, q)$ is the Euclidean distance between p and q on P. We will also adopt that convention that $\rho(X, Y) = \max_{p \in X, q \in Y} \rho(p, q)$. Let R be an arbitrary, finite set of reals. Define the *neighborhood* of $r \in R$ as $N(r) = (r - \varepsilon, r + \varepsilon)$.

Call a *candidate assignment* $(t, r) \in \mathcal{B}(V) \times R$ *consistent* if $d_*(t) \in N(r)$. The *possible assignments* for $r \in R$ are $d_*^{-1}(N(r)) \cap \mathcal{B}(V)$. Now, given $R' \subset R$, V, and $c_0 \in \mathbb{R}$ we wish to find the largest subset R'' of R' such that

$$c(d_*^{-1}(N(R'')) \cap \mathcal{B}(V) \cap \pi^{-1}(\partial P)) \quad \leq \quad c_0. \tag{2}$$

Note that $d_*^{-1}(N(R'')) \cap \mathcal{B}(V) \cap \pi^{-1}(\partial P)$ represents consistent assignments for R''. Computing this set requires us to take the intersection between the set $d_*^{-1}(N(R''))$ and the arrangement $\mathcal{B}(V)$. By the definition of $\mathcal{B}(V)$, the intersection between $d_*^{-1}(N(R''))$ and $\mathcal{B}(\mathbf{v})$ has the interesting property that for each element $\mathbf{v} \in V$, it contains either all of the set $\mathcal{B}(\mathbf{v})$ or none of it. The set $\pi^{-1}(\partial P)$ serves to constrain the subset of

TP being considered so that its base points are in ∂P. We note that this restriction can be relaxed to include any "shell" with depth γ of P; that is, the set $\pi^{-1}(\partial P)$ can be replaced with the set $\pi^{-1}(\partial P \oplus (B_\gamma \cap P))$. In Section 2.3, we give an algorithm for computing the optimal subset R'' of R'.

2.2 Application to Protein-Protein Interfaces

We now apply the optimization problem presented above in the context of determining protein-protein interfaces using NMR spectroscopy. As mentioned above, the input to our optimization problem is the manifold P, a quadratic form d, sets R' and V, and a scalar c_0. For a protein A, we view the problem of inferring the interface region of A in a complex with another protein B as an instantiation of the above problem on arrangements as follows. We take the 3-manifold P to be the space-filled structural model of A, and the 2-manifold ∂P to be the solvent-accessible surface of the structural model of A. The set $V \subset TP$ is simply the protein NH bond vectors, from the given structural model of A. We define the arrangement $\mathcal{B}(V)$ slightly differently from above; for an NH vector $\mathbf{v}$ associated with the k^{th} residue along the backbone, we define $\mathcal{B}(\mathbf{v})$ to be the subset of P that contains the van der Waals balls of the atoms in the k^{th} residue. We note that in this definition, the elements of $\mathcal{B}(V)$ can only intersect at their boundaries. In general, one RDC value is measured for each bond of a particular type – e.g., one RDC for every backbone amide bond. For each amide bond, a pair of (H^N, N) chemical shifts (frequencies) is also measured. We let R be the set of RDC values for the backbone amide bond vectors of our protein. We assume that the alignment tensor $\mathbf{S}$ has been estimated; there exist numerous techniques for estimating the alignment tensor from unassigned NMR data [23, 25, 26, 24, 34, 32, 14, 53, 54] (see Sections 1.2 and 3 for discussion on the technique we use in our algorithm). The quadratic form d is defined using $\mathbf{S}$ (see Equation (1)). We take the set R' to be the RDCs associated with amide chemical shifts that are perturbed between the apo and holo form of A. Recall that the unassigned chemical shifts that are perturbed between the apo and holo forms of a protein are associated with residues that are candidates for the interface region. Furthermore, these chemical shifts index into the experimental RDCs, thus we can determine the set R' from the experimental data. In the remainder of the paper, we let $\varepsilon = 1$, thus $N(r) = (r-1, r+1)$ (i.e., that there is 1Hz of error in the experimental RDCs). We take the c_0 to be a user-defined parameter that is given as input (see Section 4 for further discussion).

To solve our optimization problem, we wish to find the subset of the arrangement $d_*^{-1}(N(R')) \cap \mathcal{B}(V) \cap \pi^{-1}(\partial P)$ that minimizes the objective function c (see Figure 1(right)). Intuitively, this geometric optimization problem corresponds to identifying a set of candidate NH bond vectors and their residues that (a) map to, within experimental error, a set of RDCs R'' that is a subset of R' and (b) are clustered on the protein surface. Our problem definition not only accounts explicitly for experimental error in the RDC data, but it also captures the ambiguity in the structural model by representing each NH vector as a cone to model orientational uncertainty and convolving the NH vector with a surface patch on ∂P to model positional uncertainty. (Remark: It is worth noting that our framework allows these surface patches to be defined arbitrarily as long as they are of constant degree.) In Section 2.3 we give an optimal and combinatorially precise algorithm for solving this problem, and in Section 3 we give a practical algorithm along with results on experimental protein NMR data.

2.3 A Clustering Algorithm on Arrangements

In this section, we describe a combinatorially precise algorithm for solving the clustering problem presented in Section 2.1 above. For ease of exposition, let the arrangement $\mathcal{A} = d_*^{-1}(N(R')) \cap \mathcal{B}(V) \cap \pi^{-1}(\partial P)$ and the parameter c_0 be fixed. We note that, then, the sets R', V, and the quadratic form d are fixed as well. Let $|V| = n$. By definition, $\mathcal{A}$ has n generating cells; the complexity of our algorithm is determined by the number of generating cells in $\mathcal{A}$. In fact, for our application (see Section 2.2 above) $\mathcal{A}$ always has generating cells that only intersect at the boundaries, and thus total number of cells in $\mathcal{A}$ (of dimension 3) in this case is n. Since we assume that P, and thus ∂P, has maximum constant degree, the boundaries of the cells of $\mathcal{A}$ are algebraic surfaces that also have constant maximum degree. Our goal is to compute a subset of $\mathcal{A}$ that

minimizes Equation (2). Informally, our algorithm exploits the fact that the arrangement $\mathcal{A}$ can be represented using a *vertical decomposition* [20], and that we can quickly find the extrema of each cell of $\mathcal{A}$.

Our algorithm works as follows. First, we note that given $\mathcal{B}(V)$, we can take the intersection $d_*^{-1}(N(R')) \cap \mathcal{B}(V)$ in $O(n)$ time since we are given V and d, and each cell of $\mathcal{A} \cap \mathcal{B}(\mathbf{v})$ is either equal to $\mathcal{B}(\mathbf{v})$ (for some $\mathbf{v} \in V$) or $\emptyset$. First, we obtain the *vertical decomposition* of $\mathcal{A}$. The vertical decomposition of an arrangement is a essentially a recursively-defined sweep (along each dimension) of the cells of the arrangement. We omit a full description of the decomposition here, see [20] for examples and further references. For an arbitrary arrangement in $\mathbb{R}^3$ of size n, the worst-case complexity of the vertical decomposition is $\Theta(n^3)$ [20]; there is an algorithm to construct the decomposition that requires, in the worst case, $O(n^3 \log n)$ time [6]. We note that with the given decomposition, finding the extrema of the cells of $\mathcal{A}$ requires, in the worst-case $\Theta(n^3)$ time, since that is the worst-case complexity of the decomposition. Now, we can have at most $O(n)$ extrema over all cells of the arrangement, since each cell has constant degree; thus, we have $O(n^2)$ pairs of extrema. For each pair of extrema $p, q \in \mathbb{R}^3$, we check if $\rho(p, q)$ is at most c_0. For each such pair p, q, we construct a ball with diameter $\rho(p, q)$ with p and q on the boundary. Let there be k such balls. If $k = 0$, then we return the $R'' = \emptyset$. Otherwise, we calculate the following *score* on each ball. For each ball s, we compute how many cells of the arrangement lie completely in s; let number this be denoted $\sigma(s)$. This is equivalent to asking how many cells of the arrangement have all of their extrema in s; this can be done in $O(n)$ time. Let $\mathcal{C}$ be the set of all such balls. Let $s^* = \arg\max_{s \in \mathcal{C}} \sigma(s)$, and let $\mathcal{A}^*$ be the subset of $\mathcal{A}$ contained in s^*. The set $\mathcal{A}^*$ can be computed in $O(n^3)$ time, since $\sigma(s)$ can be computed in $O(n)$ time for each $s \in \mathcal{C}$, and $|\mathcal{C}|$ is $O(n^2)$. By definition, each cell of $\mathcal{A}^*$ is also in $\mathcal{A}$. Our algorithm finds the optimal set $R'' \subseteq R'$ such that R'' is the largest set that satisfies Equation (2). We return all triples $(r, \mathbf{v}, \mathcal{B}(\mathbf{v}) \cap \pi^{-1}(\partial P))$ where $r \in R''$, $\mathbf{v} \in V' = V \cap \mathcal{A}^*$, $(r, \mathbf{v})$ is a consistent assignment, and the patches $\{\mathcal{B}(\mathbf{v}) \cap \pi^{-1}(\partial P)\}_{\mathbf{v} \in V'}$ that are contained in the ball (of maximum score) associated with R''. The correctness of our algorithm follows if we can show that every subset with diameter at most c_0 is considered by the scoring phase. It is straightforward to see that the subset of $\mathcal{A}$ that yields the maximum score and has diameter at most c_0 is associated with the subset R'' that minimizes Equation (2). Thus, the following lemma proves the correctness of our algorithm:

Lemma 2.1 *Every subset $X \subseteq \mathcal{A}$ with diameter at most c_0 is contained in one of the balls in $\mathcal{C}$.*

Proof: Fix a subset X and let p and q be the pair of extrema that have maximal distance and let s denote the ball with p and q on its perimeter with diameter $\alpha = \rho(p, q)$. Note that s must contain every cell in X completely; that is, no cell of X lies outside of s, otherwise we could create a ball with diameter greater than $\rho(p, q)$. Furthermore, s is the smallest ball that can contain all of X, since any ball s' with diameter $\alpha' < \alpha$ cannot contain X. Now, s by definition is explicity considered by our algorithm in the scoring phase, and thus is contained in $\mathcal{C}$. ∎

By Lemma 2.1 and the time required to maintain the vertical decomposition data structure for $\mathcal{A}$, we have the following theorem:

Theorem 1 *The set $R'' \subseteq R'$ that minimizes Equation (2) can be computed in $O(n^3 \log n)$ time.*

3 A Clustering-based Algorithm for Identification of Protein Interfaces

The algorithm in Section 2.3 above is exact and combinatorially precise, but requires computation of algebraic surfaces. In this section we give use a practical version of the the algorithm of Section 2.3 to develop an algorithm for finding the interface region of a protein given unassigned RDCs, unassigned chemical shifts, and a structural model. Due to experimental error in the RDCs we make use of a probabilistic method to compute $\mathcal{A}$ rather than computing the intersection directly. We also model the elements of $\mathcal{A}$ using a discrete point set that represents the protein surface, rather than using an analytic representation of ∂P. As before, the input to our algorithm is the set of backbone NH vectors from a 3D structural model of the apo form

of a protein A in the complex, RDCs for the protein, a set of chemical shifts (for surface residues) that are perturbed in the holo form of the protein, and an upper bound on the diameter of the interface region. As a preprocessing step to our algorithm, we note that there is existing software to identify the perturbed chemical shifts (e.g., [38]). In the first phase, we identify the set of NH vectors (i.e., residues) associated with the given perturbed chemical shifts by using unassigned experimental RDCs. We do this by probabilistically filtering the residues in the apo structure based on how well their back-computed RDCs correspond to the experimentally-observed RDC values, eliminating from consideration residues whose back-computed RDC values match poorly. The output of this phase is Z_A. In the second phase, we compute a discrete set of points sampled uniformly from the protein surface, and identify a subset of the candidate residues on the protein surface by clustering the associated patches in this set. We use the algorithm of Section 2.3 to overcome false positive interface residues in Z_A, exploiting the intuition that the interface region of a protein is clustered on the surface of the protein. The output of this phase is C_A.

Let A be the apo form of an n-residue protein in the complex, and let H denote the holo form of the protein in the complex. We use V_A to denote the surface backbone NH vectors from the structure of A. Let R denote the RDC values observed for the NH vectors of the surface residues of A. We define the distance $\delta'(r, d(\mathbf{v}))$ between an experimental RDC r corresponding to an NH vector $\mathbf{v}$ and the back-computed RDC $d(\mathbf{v})$ as $|r - d(\mathbf{v})| / Y_r$, where $Y_r = \sum_{\mathbf{v} \in V_A} |r - d(\mathbf{v})|$. Finally, recall that we defined d to be the quadratic form which maps NH vectors to RDC values under the alignment tensor associated with R. We will use d to back-compute RDC values, each of which can be computed in $O(1)$ time given $\mathbf{v}$ and the alignment tensor. As mentioned in Section 2, there are a number of algorithms for estimating the alignment tensor; we use the algorithm of Langmead *et al.* [26], which works by searching over the anisotropic and orientational parameters of the alignment tensor. This algorithm requires $O(nk^3)$ time, where k is the resolution of a 3×3 grid over $SO(3)$; [24] gives an algorithm that requires $O(n^3)$ and simultaneously assigns the RDCs and estimates the tensor. We use the former algorithm, since the latter requires sparse NOE restraints. We note that the dependence on the grid and its resolution k can be eliminated by using the theory of real closed fields (as shown in [26]), resulting in a fully polynomial-time algorithm for tensor estimation.

Our algorithm first partitions the set R into two sets, M and M'. The set M consists of RDCs that are associated with chemical shifts that have been perturbed, and M' is simply $R - M$. Informally, our goal is to conservatively identify NH vectors that may be part of the interface region. We first compute an estimated alignment tensor using the algorithm of [26], and fix the RDC map d. Then, we construct probability distributions $p(r, \cdot)$ for each $r \in M'$ where for each $\mathbf{v} \in V_A$, $p(r, \mathbf{v}) = 1 - \delta'(r, d(\mathbf{v}))$. Initially, these probability distributions reflect our belief that the RDC r is associated with the NH vector $\mathbf{v}$. Starting with $V' = \emptyset$, we update these distributions as follows. We choose the pair $(r, \mathbf{v})$ that maximizes $p(r, \mathbf{v})$, and add $\mathbf{v}$ to V'; then we set $p(r, \mathbf{v}) = 1$ and $p(r, \mathbf{v}') = 0$ for all $\mathbf{v}' \neq \mathbf{v}$. Then, for all $r' \neq r$, we set $p(r', \mathbf{v}) = 0$ and renormalize the distributions $p(r', \cdot)$. We remove r from M' and repeat this procedure is repeated until M' is empty. The set V' represents the set of NH vectors that we believe are *not* in the interface region, since M' are RDCs associated with chemical shifts that are not perturbed. The output of this phase is the set $V'' = V_A - V' \subseteq V$.

In the second phase of our algorithm, we have a set of NH vectors, i.e., residues, V'', that are potentially part of the interface region. However, recall that for a given RDC value r and a fixed alignment tensor, the set of vectors $\mathbf{v}$ with $d(\mathbf{v}) = r$ is potentially large. Additionally, some of the NH vectors in V'' may have been incorrectly assigned to this set and may actually not be part of the interface region. Thus, there is not a necessarily one-to-one correspondence between the RDCs associated with M and the identified NH vectors V', and it is likely that a number of the identified NH vectors are false positives. To overcome this ambiguity, we use a practical version of the algorithm of Section 2.3 along with c_0 to find the interaction core C_A.

We now describe our implementation of the clustering algorithm of Section 2.3. First, we compute an approximation to ∂P by taking a uniform sample (at a fixed resolution) of ∂P. We make use of the MSMS [40] algorithm for constructing this point set; MSMS runs in $O(m \log m)$ time, where m is the number of atoms in

A. Let $\mathcal{S}_A$ be the point set computed by MSMS; note that $|\mathcal{S}_A| = O(m) = O(n)$. We partition the point set as follows: for each NH vector $\mathbf{v} \in V''$, we let $\mathcal{S}_\mathbf{v} \subset \mathcal{S}_A$ be all points in $\mathcal{S}_A$ that are associated with the same residue as $\mathbf{v}$. This set can be computed $O(|\mathcal{S}_A|)$ time. We then compute the extrema of each set $\mathcal{E}_\mathbf{v} \subseteq \mathcal{S}_\mathbf{v}$, this can be done in $O(|\mathcal{S}_A|)$ as well. Then, we proceed as in the algorithm of Section 2.3 by constructing and scoring balls defined by pairs of the computed extrema. For each pair of extrema $x \in \mathcal{E}_\mathbf{v}$ and, $y \in \mathcal{E}_\mathbf{w}$ with $\mathbf{v} \neq \mathbf{w}$ we construct a ball of diameter $\rho(x, y)$ (only if $\rho(x, y) \leq c_0$) with x and y on the boundary and compute how many of the sets $\mathcal{E}_\mathbf{v}$ lie in this ball. The score for each of the $O(n^2)$ balls can be computed in $O(n)$ time, since there are at most $|\mathcal{S}_A|$ extrema and $|\mathcal{S}_A| = O(n)$. The output of this phase is the ball with maximum score along with the set of NH vectors $\mathbf{v}$ and associated sets $\mathcal{S}_\mathbf{v}$ that lie inside this ball.

We now summarize the asymptotic time complexity of our practical algorithm. During the first phase of the algorithm, we must estimate the alignment tensor; this requires $O(nk^3)$ time. We then compute the set $V_A - V'$, which requires $O(n^2)$ time to construct and update the probability distributions. In the second phase of the algorithm, we must construct the set $\mathcal{S}_A$, which requires $O(m \log m)$ time, where m is the number of atoms in A. The clustering step of the second phase requires $O(n^3)$ time. The overall running time of our algorithm is then $O(nk^3 + m \log m + n^3) = O(nk^3 + n^3)$ time.

4 Results and Discussion

We implemented and tested the algorithm described in Section 3 on 7 proteins from 5 different protein complexes: the apo forms of Pex13P (PDB ID: 1NM7), CAD (PDB ID: 1C9F), ubiquitin (PDB ID: 1D3Z), barnase (PDB ID: 1BNR), barstar (PDB ID: 1BTA), E1N (PDB ID: 1EZA), and HPr (PDB ID: 1HDN) from the CAD-ICAD [37], ubiquitin-CUE [21], barnase-barstar [4], E1N-HPr [15] protein-protein complexes and the Pex13P-Pex14P [12] protein-peptide complex. We then compared the output of our algorithm (the interaction zone and core) against the true (i.e., experimentally-determined) interface region reported in the literature. We report the *accuracy* (the fraction of the actual interface region that was found by our algorithm) of the interaction zone, and the *sensitivity* (the fraction of the output of our algorithm that was part of the interface region for the given protein) of the interaction core. Table 1 shows the results of our algorithm on these proteins.

For our experiments, we used experimental RDC data for a single aligning medium for E1N, HPr, and ubiquitin available from the BioMagResBank (BMRB) [42]. For these proteins, a second set of RDC data for a second aligning medium was simulated. As mentioned in Section 1.2, additional aligning media serve to constrain the solutions for the NH vector orientations. These can be incorporated easily into the first phase of our algorithm. For ℓ aligning media, each RDC r is given one probability distribution per medium, and while constructing the set V', we take the maximum joint probability when matching a set of ℓ RDCs (i.e. ℓ RDCs with the same chemical shift) to a vector $\mathbf{v}$. For the remaining proteins, experimental RDC data is not publicly available; two sets of RDC data for two independent aligning media were simulated for Pex13P, CAD, barnase and barstar. For simulated RDC data, we used a Gaussian error window of 1Hz. Although we have experimental NMR chemical shifts and NH vectors for all residues in the proteins being tested, we only make use of surface NH vectors and chemical shifts. Surface NH vectors can be easily identified from the given structural model, and surface chemical shifts can be identified experimentally using amide exchange data; we used the program MolMol to compute these NH vectors. Solvent accessibility (i.e., percentage of atomic surface area exposed to solvent) and the chemical shift assignment was used to identify chemical shifts associated with residues whose solvent accessibility was at least 40%. The set of surface residues that we used as input in all of our experiments were the residues identified by MolMol as being at least 40% solvent-accessible, as well as any residues in the interface region for that protein. We implemented our algorithm in Matlab (Mathworks Inc, Natick, MA), and ran all of our experiments on a Pentium-4 class processor. Since some of our input data (specifically, simulated RDC data) was generated with a Gaussian error window, the test results in Tables 1 and 2 give the average accuracy and sensitivity over 10 trials for each protein. For our test cases, each execution of our algorithm required about 2 or 3 minutes of CPU time on average.

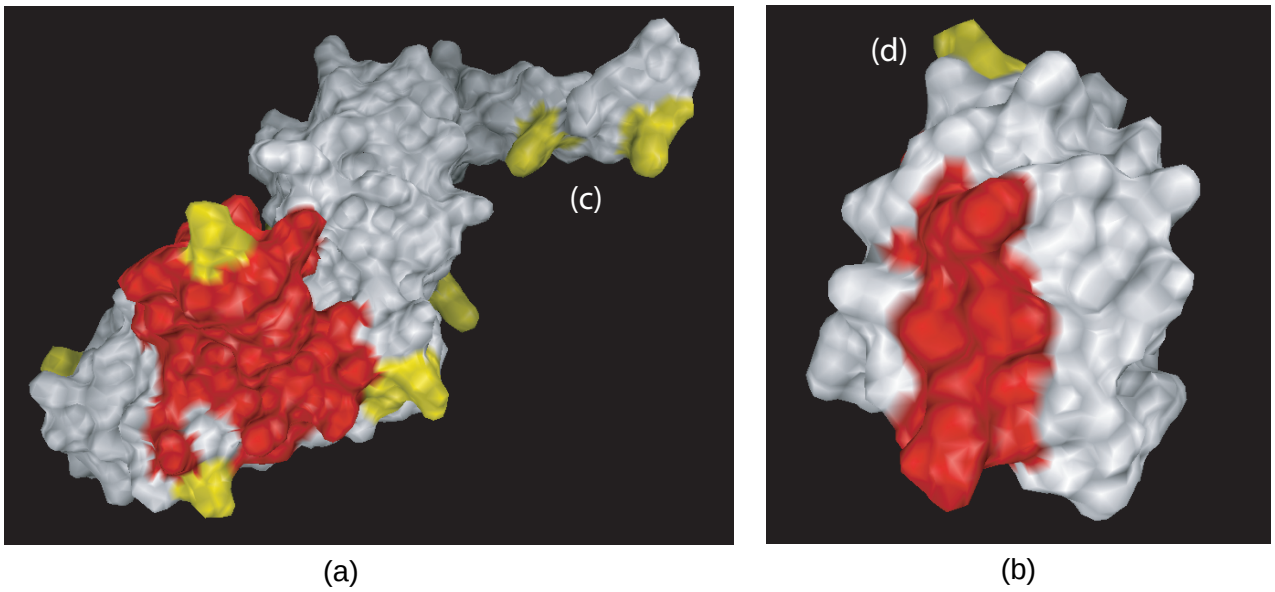

Figure 2: **Sample output of our algorithm for the E1N-HPr complex.** The results of our algorithm for E1N (a) and HPr (b). The interaction core (C_A) is shown red and residues in the interaction zone but not the interaction core ($Z_A \setminus C_A$) are shown in yellow; note that the interaction zone Z_A is simply the union of the yellow and red residues. Selective labelings of NV and EF were simulated for E1N and HPr, respectively; a setting of $c_0 = 35$ Å was used for both proteins. For both E1N and HPr, false positive residues are eliminated by the interaction core. In E1N, the interaction core captures about 85% of the true interface region, and eliminates all false positive residues (for example, the region (c) near the N-terminus); the sensitivity of the interaction core for E1N is 100%. In HPr, the single false positive residue (d) in the interaction zone is eliminated by the interaction core; the core has 98% accuracy and 100% sensitivity.

Protein	Accuracy	Sensitivity
PEX13P	73%	80%
barnase	72%	90%
barstar	77%	100%
ubiquitin	73%	73%
CAD	75%	90%
HPr	88%	100%
E1N	90%	100%

Protein	Core diameter (Å)	Sensitivity
barstar	20	100%
	25	100%
	30	99%
	35	81%
HPr	20	100%
	25	91%
	30	88%
	35	73%

Table 1: **Results without selective labeling.** The table on the left shows the accuracy of the interaction zone and the sensitivity of the interaction core. We took the diameter of core (c_0) to be 20 Å. The table on the right shows the tradeoff between sensitivity and c_0; sensitivity is decreased as c_0 is increased.

The *accuracy* of the interaction zone Z_A is the percentage of true interface residues contained in Z_A. For our test cases, we achieved accuracies between 73% and 90%. The *sensitivity* of the interaction core C_A is the percentage of C_A comprised of interface residues; we achieved sensitivities of between 73% and 100%. Accuracy and sensitivity results are reported for each protein in Table 1(left). A key feature of our algorithm is the ability to choose the diameter threshold c_0 for the interaction core. Figure 2 gives a visualization of the interaction zone and interaction core for a sample output of our algorithm on the E1N-HPr complex. With a conservative value (i.e., significantly smaller than the interface region itself), we are able to achieve very high sensitivity at the expense of decreased accuracy. That is, when c_0 is small, the second phase of our algorithm returns a small number of residues, but they are all guaranteed to be in the interface region. As we increase c_0, the size of the interaction core increases, but these residues are not all necessarily guaranteed to be in the interface region. Table 1(right) shows the tradeoff between the sensitivity of the interaction core and c_0 for two representative proteins. However, the accuracy of the interaction core (i.e., percentage of the true interface region contained in the core) decreases as the core diameter decreases. For example, for barstar, the core accuracy decreases from 86% to 77% when c_0 is decreased from 30 Å to 25 Å. We note that this feature of our algorithm is important in applications such as drug design and protein-protein docking, since users can treat c_0 as essentially a confidence parameter, setting it conservatively for obtain high sensitivity. For example, the docking study of [10] found that in some cases, distance restraints between just a single pair of residues are sufficient to significantly constrain the relative rotations and translations of the two proteins in the complex. It is thus possible to run our algorithm on both of the proteins of a complex and use the computed interaction cores to constrain the docking process *a priori*, reducing the time spent searching rotations and translations by existing approaches [11, 10, 8, 31]. Furthermore, if c_0 is set conservatively, it is likely that the remaining interface residues are nearby; in our test cases, all interface residues that were not in the interface core were all within about 10 Å from the core.

Selective labeling allows the stable isotopic labeling of a given set of residue types, and thus allows us to constrain the amino acid type of an experimentally-recorded RDC if that type has been labeled. In our algorithm, this additional constraint can be used in the first phase to initially condition the probability distributions as they are constructed. This decreases the ambiguity in matching residues, and thus can improve both accuracy and sensitivity of the interaction zone. In practice, the most useful residue types for selective labeling can be determined from the primary sequence and apo structure, as well as from biophysical characterizations of which amino acid types are likely to be on the protein surface [9]. We show in Table 2 that our experimental results can be improved by using selective labeling; for each protein, we give a labeling that improves both the accuracy of the interaction zone and the sensitivity of the interaction core. By using selective labeling, we are able to improve the average accuracy of the interaction zone to 88% and the average sensitivity of the interaction core to 97%. Furthermore, we observed the same tradeoff between accuracy and sensitivity of the interaction core; however, the sensitivity of C_A was improved due to the constraint added by selective labeling in the first phase of our algorithm.

5 Conclusion

In this paper, we have formalized the problem of finding a protein interface from *unassigned* NMR data as a geometric clustering problem. We gave an optimal algorithm for the geometric clustering algorithm that runs in $O(n^3 \log n)$. Using this algorithm, we developed a practical algorithm for finding protein interfaces given unassigned chemical shifts, unassigned RDCs and a structural model of the apo protein that runs in $O(nk^3 + n^3)$ time. When run on NMR data for 7 proteins, we showed that our algorithm yielded results that were both accurate and had high sensitivity (i.e., a low false-positive rate), demonstrating that our algorithm is useful in practice. It would be interesting to see if our algorithm could be applied to proteins with multiple interface regions. In principle, our algorithm could be applied: in the second phase, instead of returning the cluster with highest score, we would return a set of clusters with high score as the interaction cores.

Protein	Accuracy	Sensitivity	Labeling
PEX13P	87%	94%	RDQKF
barnase	78%	85%	NGKT
barstar	91%	100%	RQKS
ubiquitin	74%	100%	RNDKT
CAD	85%	100%	QEHMS
HPr	88%	100%	EF
E1N	93%	100%	NV

Protein	Core diameter (Å)	Sensitivity
barstar	20	100%
	25	100%
	30	100%
	35	96%
HPr	20	100%
	25	91%
	30	93%
	35	84%

Table 2: **Results with selective labeling.** The table on the left shows the accuracy of the interaction zone and the sensitivity of the interaction core, as well as the labelings used in our experiments. We took the diameter of core (c_0) to be 20 Å. The table on the right shows the tradeoff between sensitivity and c_0; sensitivity is decreases as c_0 is increased.

Acknowledgements

The authors would like to thank Dr. Chris Bailey-Kellogg, Dr. Jack Kelley, Dr. Chris Langmead, Dr. Gerhard Wagner, Dr. Jeff Hoch, Dr. Lincong Wang, Anthony Yan and all members of Donald Lab for helpful discussions and comments.

References

[1] C. Bailey-Kellogg, S. Chainraj, and G. Pandurangan. A random graph approach to NMR sequential assignment. In *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology*, pages 58–67, 2004.

[2] C. Bailey-Kellogg, A. Widge, J. J. Kelley III, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. In *Proceedings of the Fourth Annual International Conference on Computational molecular biology*, pages 33–44, 2000.

[3] A. Y.-E. Ban, H. Edelsbrunner, and J. Rudolph. Interface surfaces for protein-protein complexes. In *Proceedings of the 8th International Conference on Research in Computational Biology*, pages 205–212, March 2004.

[4] A. M. Buckle, G. Schreiber, and A. R. Fersht. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0 Å resolution. *Biochemistry*, 33:8878–8889, 1994.

[5] J. Cavanagh, Fairbrother W. J., A. G. Palmer III, and N. J. Skelton. *Protein NMR Spectroscopy: Principles and Practice*. Academic Press, 1995.

[6] B. Chazelle, H. Edelsbrunner, L. Guibas, and M. Sharir. A singly-exponential stratification scheme for real semi-algebraic varieties and its applications. *Theoretical Computer Science*, 84:77–105, 1991.

[7] G. M. Clore. Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization. *Proceedings of the National Academy of Sciences*, 97:9021–9025, 2000.

[8] G. M. Clore and C. D. Schwieters. Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from ^{15}N-H chemical shifts mapping and backbone ^{15}N-H

residual dipolar couplings using conjoined rigid body/torsion angle dynamics. *Journal of the American Chemical Society*, 125:2902–2912, 2003.

[9] B. I. Dahiyat and S. L. Mayo. *De novo* protein design: fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[10] A. Dobrodumov and A. M. Gronenborn. Filtering and selection of structural models: Combining docking and NMR. *Proteins: Structure, Function, and Genetics*, 52:18–32, 2003.

[11] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125:1731–1737, 2003.

[12] A. Douangamath, F. V. Filipp, A. T. J. Klein, P. Barnett, P. Zou, T. Voorn-Brouwer, M.C. Vega, O. M. Mayans, M. Sattler, B. Distel, and M. Wilmanns. Topography of independent binding of alpha-helical and PPII-helical ligands to a peroxisomal SH3 domain. *Mol. Cell*, 10:1007–1017, 2002.

[13] S.W. Englander, T.R. Sosnick, J.J. Englander, and L. Mayne. Mechanisms and uses of hydrogen exchange. *Curr. Opin. Struct. Biol.*, 6:18–23, 1996.

[14] M. A. Erdmann and G. S. Rule. Rapid protein structure detection and assignment using residual dipolar couplings. Technical Report 195, Department of Computer Science, Carnegie-Mellon University, 2002.

[15] D. S. Garrett, Y.-J. Seok, A. Peterkofsky, A. M. Gronenborn, and G. M. Clore. Solution structure of the 40,000 M_r phosphoryl transfer complex between the N-terminal domain of Enzyme I and HPr. *Nature Structural Biology*, 6(2):166–173, 1999.

[16] G. Gemmecker, W. Jahnke, and H. Kessler. Measurement of fast proton exchange rates in istopically labeled compounds. *Journal of the American Chemical Society*, 115(24):11620–11621, 1993.

[17] A. Grishaev and M. Llinas. Protein structure elucidation from NMR proton densities. *Proceedings of the National Academy of Sciences*, 99(10):6713–6718, 2002.

[18] S. Grzesiek and A. Bax. Measurement of amide proton exchange rates and NOEs with water in ^{13}C/^{15}N-enriched calcineurin B. *Journal of Biomolecular NMR*, 3(6):627–638, 1993.

[19] P. Güntert. Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43:105–125, 2003.

[20] D. Halperin. Arrangements. In J. E. Goodman and J. O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, pages 389–412. CRC Press, New York, NY, 1997.

[21] R. S. Kang, C. M. Daniels, S. A. Francis, S. C. Shih, W. J. Salerno, L. Hicke, and I. Radhakrishnan. Solution structure of a CUE-Ubiquitin complex reveals a conserved mode of Ubiquitin binding. *Cell*, 113:621–630, 2003.

[22] O. Kohlbacher, A. Burchardt, A. Moll, A. Hildebrandt, P. Bayer, and H.-P. Lenhof. Structure prediction of protein complexes by a nmr-based protein docking algorithm. *Journal of Biomolecular NMR*, 20:15–21, 2001.

[23] C. J. Langmead and B. R. Donald. 3D-Structural Homology Detection via Unassigned Residual Dipolar Couplings. *Proc. IEEE Computer Society Bioinformatics Conference (CSB), Stanford University, Palo Alto, CA (August 11-14)*, pages 209–217, 2003.

[24] C. J. Langmead and B. R. Donald. An Expectation/Maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *Journal of the Biomolecular NMR*, 29(2):111–138, 2004.

[25] C. J. Langmead and B. R. Donald. High-Throughput 3D Structural Homology Detection via NMR Resonance Assignment. *Proc. IEEE Computer Society Bioinformatics Conference (CSB), Stanford University, Palo Alto, CA*, 2004. 278–289.

[26] C. J. Langmead, A. K. Yan, L. Wang, R. H. Lilien, and B. R. Donald. A polynomial time nuclear vector replacement algorithm for automated NMR resonance assignment. In *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology*, pages 176–187, April 2003.

[27] C. J. Langmead, A. K. Yan, L. Wang, R. H. Lilien, and B. R. Donald. A polynomial time nuclear vector replacement algorithm for automated NMR resonance assignment. *Journal of Computational Biology*, 11(2–3):277–298, 2004.

[28] R. H. Lilien, M. Sridharan, and B. R. Donald. Identification of novel small molecule inhibitors of core-binding factor dimerization by computational screening against NMR molecular ensembles. Technical Report 492, Dartmouth College, March 2004.

[29] J.A. Losonczi, M. Andrec, W.F. Fischer, and Prestegard J.H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson*, 138(2):334–42, 1999.

[30] T. Matsuda, T. Ikegami, N. Nakajima, T. Yamakazi, and H. Nakamura. Model building of a protein-protein complexed structure using saturation transfer and residual dipolar coupling without paired intermolecular NOE. *Journal of Biomolecular NMR*, 29:325–338, 2004.

[31] M. A. McCoy and D. F. Wyss. Structures of protein-protein complexes are docked using only NMR restraints from residual dipolar couplings and chemical shift perturbations. *Journal of the American Chemical Society*, 124:2104–2105, 2002.

[32] J. Meiler and D. Baker. Rapid fold determination using unassigned NMR data. *Proceedings of the National Academy of Sciences*, 100(26):15404–15409, 2003.

[33] G.T. Montelione, D. Zheng, Y. J. Huang, K.C. Gunsalus, and T. Szyperski. Protein NMR spectroscopy in structural genomics. *Nature Structural Biology*, 7(11):982–985, 2000.

[34] L. C. Morris, H. Valafar, and J. H. Prestegard. Assignment of Backbone Resonances from Minimal NMR Data Using Connectivity, Torsion Angle Constraints, and Chemical Shifts. *Journal of Biomolecular NMR*, 29:1–9, 2004.

[35] H.N. Moseley and G.T. Montelione. Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol.*, 9(5):635–642, 1999.

[36] National Institute of General Medical Sciences, National Institutes of Health. The Protein Structure Initiative. `http://www.nigms.nih.gov/psi/`.

[37] T. Otomo, H. Sakahira, K. Uegaki, S. Nagata, and T. Yamazaki. Structure of the heterodimeric complex between CAD domains of CAD and ICAD. *Nature Structural Biology*, 7:658–662, 2000.

[38] C. Peng, S. W. Unger, F. V. Filipp, M. Sattler, and S. Szalma. Automated evaluation of chemical shift perturbation spectra: New approaches to quantitative analysis of receptor-ligand interaction NMR spectra. *Journal of Biomolecular NMR*, 29(4):491–504, 2004.

[39] M. L Reese and V. Dötsch. Fast mapping of protein-protein interfaces by NMR spectroscopy. *Journal of the American Chemical Society*, 125:14250–14251, 2003.

[40] M. F. Sanner, A. J. Olson, and J.-C. Spehner. Fast and robust computation of molecular surfaces. In *Proceedings of the 11th Annual ACM Symposium on Computational Geometry*, pages C6–C7, June 1995.

[41] A. Saupe. Recent Results in the field of liquid crystals. *Angew. Chem.*, 7:97–112, 1968.

[42] B.R. Seavey, E.A. Farr, W.M. Westler, and J.L. Markley. A relational database for sequence-specific protein NMR data. *Journal of Biolmolecular NMR*, 1(3):217–236, 1991.

[43] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik. Discovering high affinity ligands for proteins: SAR by NMR. *Science*, 274:1531–1534, 1996.

[44] F. Tian, H. Valafar, and J. H. Prestegard. A Dipolar Coupling Based Strategy for Simultaneous Resonance Assignment and Structure Determination of Protein Backbones. *J. Am. Chem. Soc.*, 123:11791–11796, 2001.

[45] N. Tjandra and A. Bax. Direct Measurement of Distances and Angles in Biomolecules by NMR in a Dilute Liquid Crystalline Medium . *Science*, 278:1111–1114, 1997.

[46] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution. *Proc. Natl. Acad. Sci. USA*, 92:9279–9283, 1995.

[47] L. Wang and B. R. Donald. Exact solutions for internuclear vectors and backbone dihedral angles from nh residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *Journal of Biomolecular NMR*, 29:223–242, 2004.

[48] Y. Xu, D. Xu, O. H. Crawford, J. R. Einstein, and E. Serpersu. Protein structure determination using protein threading and sparse NMR data. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology*, pages 299–307, 2000.

[49] D. Zheng, Y.J. Huang, H.N. Moseley, R. Xiao, J. Aramini, G.V. Swapna, and G.T. Montelione. Automated protein fold determination using a minimal NMR constraint strategy. *Protein Science*, 12(6):1232–1246, 2003.

[50] D.E. Zimmerman, C.A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G.T. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology*, 269(4):592–610, 1997.

[51] D.E. Zimmerman and G.T. Montelione. Automated analysis of nuclear magnetic resonance assignments for proteins. *Curr Opin Struct Biol.*, 5(5):664–673, 1995.

[52] E. R. P. Zuiderweg. Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry*, 41(1):1–7, 2002.

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

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