BIOINFORMATICS

Bioinformatics *Vol. 21 Suppl. 1 2005*
Pages. i292–i301
doi:10.1093/bioinformatics/bti1005

High-Throughput Inference of Protein-Protein Interfaces from Unassigned *NMR Data*

Ramgopal R. Mettu[a], **Ryan H. Lilien**[a,b] **Bruce Randall Donald**[a,c,d,*]

[a]*Dartmouth Computer Science Department, Hanover, NH 03755, USA*
[b]*Dartmouth Medical School, Hanover, NH 03755, USA*
[c]*Dartmouth Chemistry Department, Hanover, NH 03755, USA*
[d]*Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA*

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, where the arrangement is induced by a spherical quadratic form. 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.

1 INTRODUCTION[1]

Protein-protein interactions are well-studied in structural biology, and the structural basis for these interactions are useful in elucidating the biological role of the consituent proteins. As the Protein Structure Initiative [31] rapidly populates the "space of protein structures," an emerging goal of structural proteomics is to study not just individual proteins, but protein complexes and networks of protein interactions, as well as the molecular and structural basis for these interactions. High-throughput computational approaches for identifying the interface region between proteins in a complex can serve a useful role in studying these protein-protein interactions. Recent advances in solution NMR spectroscopy allow us to directly study the interaction between two proteins in solution; NMR is ideally suited to studying protein-ligand and protein-protein interactions [42]. In contrast to existing approaches that rely on *assigned* NMR data, in this paper we develop an efficient algorithm for identifying the interface between two proteins in a complex using *unassigned* NMR data.

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 [17, 2, 23, 1, 28, 30] 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,

*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] 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 complex form of a protein in protein complex; SVD, singular value decomposition; $SO(3)$, special orthogonal group in 3D; S^2, a 2-sphere in $\mathbb{R}^3$.

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 a pair 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, 26]. 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 [23, 22, 43, 1, 2] (see [17] 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 an order of magnitude less 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,[2] 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 [39]. 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, due to 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 *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, 26] or that multiple experiments utilizing selective isotopic labeling be performed [34]. We first consider a geometric version of the problem of identifying protein interfaces that asks us to cluster the cells of an *arrangement* on a 2-manifold. 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 Z_A. Then, in the second phase, we use a practical version of our geometric clustering algorithm that, given a size threshold, identifies the interaction core C_A. 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.

[2] Full details of the results in the paper, including additional figures, can be found in [27].

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 algorithm on different combinations of real and simulated NMR data from 7 proteins that shows it could be useful in high-throughput applications.

1.1 Previous Work

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., [42] 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, 26] or that multiple experiments utilizing selective labeling be performed [34]. The key difference between our work and much of the previous work is that we require only *unassigned* NMR data, and seek only to identify the residues involved in the interface region without predicting [11, 10, 26] 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 [24]. The goal of working with unassigned data is to minimize manual, wetlab, and computational time, as well as resources, needed, and to thus facilitate high-throughput examination of various structural properties of proteins [21, 23, 28, 29, 14, 43].

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 [26, 20, 11, 7, 8] compute the translation and rotation that brings the apo structures together to produce the complex structure. In general, the experimental NMR data must first be assigned; NOE data is particularly hard to assign due to chemical shift degeneracy [7, 8]. However, without experimental data, the accuracy of the predicted complex structure is determined solely by the energy function, and not by experimental observations of the complex in solution.

Another ubiquitous technique in the study of protein-protein interfaces is called *chemical shift mapping* [42, 38], which compares the change in HSQC spectra (see Section 1.2 below) for the free and complex spectra of the protein. To directly identify the interface region from chemical shift perturbations, the HSQC must be assigned. McCoy and Wyss [26] 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. With unassigned HSQC spectra, it is possible, through titration experiments, to identify which (unassigned) HSQC peaks have shifted [33].

In contrast to many 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. Kohlbacher *et al.* [20] use unassigned experimental ^{1}H spectra to score candidate dockings; however they do not use experimental data to directly identify the interface region. Compared to the work of [34] which uses selective labeling and unassigned NMR data, our approach is faster and cheaper since the amount of wetlab time is fixed for our technique 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.

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 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. The chemical shift of a nucleus changes when its local electronic environment changes. Hence, the holo vs. apo spectrum indicates binding or conformational change, allowing us to identify residues in the interface region. Conversely, zero chemical shift change can indicate that binding has *not* occurred. We further assume that the holo structure does not undergo significant conformational change outside of the interface

region; similar assumptions are made by most docking protocols [10, 8, 26]. Once the identity of each peak's atoms (in the primary sequence) is known, chemical shift information can be useful in studying protein-ligand [38] and protein-protein [42] 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] 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.

Our algorithm also uses *residual dipolar coupling* (RDC) data [25, 36, 39]. 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:

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

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 [25]. 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 $\mathbf{S}$ [25]. There are a number of techniques to estimate the alignment tensor given *unassigned* RDCs [21, 23, 22, 14, 29, 43]. 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 [40, 41]. 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 [41]. 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 $\mathbf{S}$ 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 PROBLEM DEFINITION AND APPLICATION

In this section, we formally define a clustering problem in 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.

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 gives a commutative diagram of the mappings π, j, d, and d_*. Let the *cost* of $X \subseteq TP \setminus 0$ be defined as

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

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'

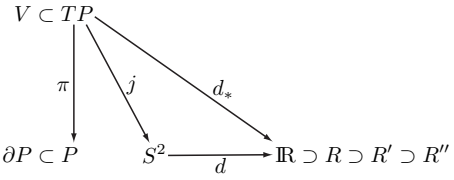

Fig. 1. Commutative diagram of the mappings used in our problem definition.

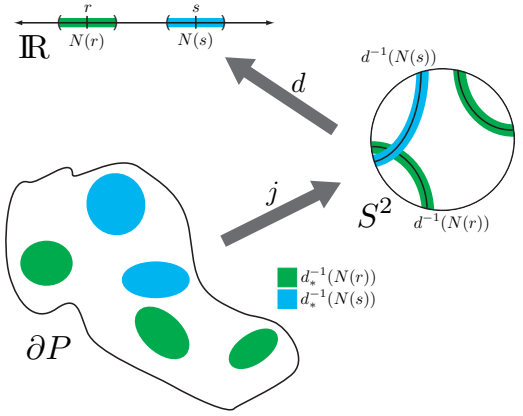

Fig. 2. Our clustering problem in an arrangement with the set $R' = \{r, s\}$. Starting with the neighborhood of R', i.e., the intervals $N(r)$ and $N(s)$ in $\mathbb{R}$, we consider the set of orientations (contained in S^2) that are associated with these intervals. These orientations are $d^{-1}(N(r))$ 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 diameter threshold c_0.

such that

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

Note that $d_*^{-1}(N(R'')) \cap \mathcal{B}(V) \cap \pi^{-1}(\partial P)$ represents possible 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 intersect only at 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 [21, 23, 22, 29, 14, 43] (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 1 Hz 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 2). 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's base point 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 intersect only at boundaries, and thus total number of 3-cells in $\mathcal{A}$ 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* [18], 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 [18] 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 $\Theta(n^3)$ [18]; 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 TO IDENTIFY PROTEIN INTERFACES

The algorithm in Section 2.3 is exact and combinatorially precise, but requires computation of algebraic surfaces. In this section we give a practical version of the the algorithm of Section 2.3. 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 algebraic 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., [33]).

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. 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 first compute an estimated alignment tensor using the algorithm of [23], and fix the RDC map d. Our algorithm then partitions the set R into two sets, M, RDCs that are associated with perturbed chemical shifts, and $M' = R \setminus M$. We then probabilistically match RDCs in M' with NH vectors V_A by eliminating the highest joint-probability match, and successively conditioning match probabilities on previous eliminations (cf. [22]). After all RDCs in M' have been matched, we output the remaining NH vectors as the interaction zone Z_A. In the second phase, we filter Z_A further by using the algorithm of Section 2.3 as follows. First, we compute an approximation to ∂P by taking a uniform sample (at a fixed resolution) of ∂P. We make use of the MSMS [35] 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 proceed as in Section 2.3 and output the residues associated with the highest-scoring cluster as the interaction core C_A. During the first phase of the algorithm, tensor estimation requires $O(nk^3)$ time, and the set $V_A - V'$ requires $O(n^2)$ time to construct. In the second phase of the algorithm, the set $\mathcal{S}_A$ requires $O(m \log m)$ time, where m is the number of atoms in A, and the clustering step requires $O(n^3)$ time. The overall running time of our algorithm is then $O(nk^3 + m \log m + n^3) = O(nk^3 + n^3)$.

4 RESULTS AND DISCUSSION

We implemented and tested our algorithm 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 [32], ubiquitin-CUE [19], barnase-barstar [4], E1N-HPr [15] protein-protein complexes and the Pex13P-Pex14P [12] protein-peptide complex. We assume that the manual (and generally time-consuming) experimental studies for these complexes have produced the true interface regions, and compare the results of our algorithm against them. We report the *accuracy* (the fraction of the interface region identified 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) of the interaction core (Figure 3).

For our experiments, we used experimental RDC data for a single aligning medium for E1N, HPr, and ubiquitin available from the BioMagResBank (BMRB) [37]. 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 that can be incorporated as follows. For 2 aligning media, each RDC r is given one probability distribution per medium; we match experimental RDCs to NH vectors by taking the maximum joint probability that the RDCs in both media match 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 1 Hz. 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 Figure 3 give the average accuracy and sensitivity over 10 trials for each protein. For our test cases, each execution of our algorithm required about 2

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

(a)

Protein	c_0	Sensitivity
barstar	20	100%
	25	100%
	30	99%
	35	81%
HPr	20	100%
	25	91%
	30	88%
	35	73%

(b)

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

(c)

Protein	c_0	Sensitivity
barstar	20	100%
	25	100%
	30	100%
	35	96%
HPr	20	100%
	25	91%
	30	93%
	35	84%

(d)

Fig. 3. Results. (a) Accuracy of interaction zone and sensitivity of interaction core; (b) Tradeoff between sensitivity and c_0; (c) Accuracy of interaction zone and sensitivity with selective labeling; (d) Tradeoff between sensitivity and c_0 with selective labeling. For (a) and (c), the diameter of the interaction core, c_0, was set to 20 Å.

or 3 minutes of CPU time on average. 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 (for a visualization of the output of our algorithm on the proteins in the E1N-HPr complex see [27]). A key feature of our algorithm is the ability to choose the diameter threshold c_0 for the interaction core. 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. Figure 3(b) 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, 26]. 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 improve the accuracy of matching experimental RDCs to NH vectors. This can, in turn, 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]. Figure 3(c) shows 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 observe the same tradeoff between accuracy and sensitivity of the interaction core (see Figure 3(d)); 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. 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 generalized: in the second phase, we would return a set of clusters with high score, rather than a single cluster, as the interaction cores.

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. In *Proc. RECOMB 2004*, pp. 58–67.

[2] C. Bailey-Kellogg, A. Widge, J. J. Kelley III, M. J. Berardi, J. H. Bushweller, and B. R. Donald. *J. Comp. Bio.*, 7(3–4):537–558.

[3] A. Y.-E. Ban, H. Edelsbrunner, and J. Rudolph. In *Proc. RECOMB 2004*, pp. 205–212.

[4] A. M. Buckle, G. Schreiber, and A. R. Fersht. *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. *Theoretical Computer Science*, 84:77–105, 1991.

[7] G. M. Clore. *Proc. Natl. Acad. Sci. USA*, 97:9021–9025, 2000.

[8] G. M. Clore and C. D. Schwieters. *J. Am. Chem. Soc.*, 125:2902–2912, 2003.

[9] B. I. Dahiyat and S. L. Mayo. *Science*, 278(5335):82–87, 1997.

[10] A. Dobrodumov and A. M. Gronenborn. *Proteins: Structure, Function, and Genetics*, 52:18–32, 2003.

[11] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin. *J. Am. Chem. Soc.*, 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. *Mol. Cell*, 10:1007–1017, 2002.

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

[14] M. A. Erdmann and G. S. Rule. 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. *Nat. Struct. Biol.*, 6(2):166–173, 1999.

[16] S. Grzesiek and A. Bax. *J. Biomol. NMR*, 3(6):627–638, 1993.

[17] P. Güntert. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43:105–125, 2003.

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

[19] R. S. Kang, C. M. Daniels, S. A. Francis, S. C. Shih, W. J. Salerno, L. Hicke, and I. Radhakrishnan. *Cell*, 113:621–630, 2003.

[20] O. Kolhbacher, A. Burchardt, A. Moll, A. Hildebrandt, P. Bayer, and H.-P. Lenhof. *J. Biomol. NMR*, 20(1):15–21.

[21] C. J. Langmead and B. R. Donald. In *Proc. IEEE CSB 2003*, pp. 209–217.

[22] C. J. Langmead and B. R. Donald. *J. Biomol. NMR*, 29(2):111–138, 2004.

[23] C. J. Langmead, A. K. Yan, L. Wang, R. H. Lilien, and B. R. Donald. *J. Comp. Bio.*, 11(2–3):277–298, 2004.

[24] R. H. Lilien, M. Sridharan, and B. R. Donald. Technical Report 492, Dartmouth College Computer Science Department, March 2004. `http://www.cs.dartmouth.edu/reports/`.

[25] J. A. Losonczi, M. Andrec, W.F. Fischer, and Prestegard J.H. *J. Magn. Reson.*, 138(2):334–42, 1999.

[26] M. A. McCoy and D. F. Wyss. *J. Am. Chem. Soc.*, 124:2104–2105, 2002.

[27] R. R. Mettu, R. H. Lilien, and B. R. Donald. Technical Report 530, Dartmouth College Computer Science Department, January 2005. `http://www.cs.dartmouth.edu/reports/`.

[28] G.T. Montelione, D. Zheng, Y. J. Huang, K.C. Gunsalus, and T. Szyperski. *Nat. Struct. Biol.*, 7(11):982–985, 2000.

[29] L. C. Morris, H. Valafar, and J. H. Prestegard. *J. Biomol. NMR*, 29:1–9, 2004.

[30] H.N. Moseley and G.T. Montelione. *Curr. Opin. Struct. Biol.*, 9(5):635–642, 1999.

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

[32] T. Otomo, H. Sakahira, K. Uegaki, S. Nagata, and T. Yamazaki. *Nat. Struct. Biol.*, 7:658–662, 2000.

[33] C. Peng, S. W. Unger, F. V. Filipp, M. Sattler, and S. Szalma. *J. Biomol. NMR*, 29(4):491–504, 2004.

[34] M. L Reese and V. Dötsch. *J. Am. Chem. Soc.*, 125:14250–14251, 2003.

[35] M. F. Sanner, A. J. Olson, and J.-C. Spehner. In *Proc. ACM-SoCG 1995*, pp. C6–C7.

[36] A. Saupe. *Angew. Chem.*, 7:97–112, 1968.

[37] B.R. Seavey, E.A. Farr, W.M. Westler, and J.L. Markley. *J. Biomol. NMR*, 1(3):217–236, 1991.

[38] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik. *Science*, 274:1531–1534, 1996.

[39] N. Tjandra and A. Bax. *Science*, 278:1111–1114, 1997.

[40] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. *Proc. Natl. Acad. Sci. USA*, 92:9279–9283, 1995.

[41] L. Wang and B. R. Donald. *J. Biomol. NMR*, 29:223–242, 2004.

[42] E. R. P. Zuiderweg. *Biochemistry*, 41(1):1–7, 2002.

[43] M. Zweckstetter and A. Bax. *J. Am. Chem. Soc.*, 123(38):9490–1, 2001.