Simultaneous determination of subunit and complex structures of symmetric homo-oligomers from ambiguous NMR data

Himanshu Chandola  
Department of Computer Science  
Dartmouth College  
6211 Sudikoff Laboratory  
Hanover, NH 03755, USA  
himanshu@cs.dartmouth.edu

Bruce R. Donald  
Departments of Computer Science and Biochemistry  
Duke University  
Durham, NC 27708  
brd+acmbcb13@cs.duke.edu

Chris Bailey-Kellogg  
Department of Computer Science  
Dartmouth College  
6211 Sudikoff Laboratory  
Hanover, NH 03755, USA  
ckb@cs.dartmouth.edu

ABSTRACT

Determining the structures of symmetric homo-oligomers provides critical insights into their roles in numerous vital cellular processes. Structure determination by nuclear magnetic resonance spectroscopy typically pieces together a structure based primarily on interatomic distance restraints, but for symmetric homo-oligomers each restraint may involve atoms in the same subunit or in different subunits, as the different homo-oligomeric “copies” of each atom are indistinguishable without special experimental approaches. This paper presents a novel method that simultaneously determines the structure of the individual subunits and their arrangement into a complex structure, so as to best satisfy the distance restraints under a consistent (but partial) disambiguation. Recognizing that there are likely to be multiple good solutions to this complex problem, our method provides a guarantee of completeness to within a user-specified resolution, generating representative backbone structures for the secondary structure elements, such that any structure that satisfies sufficiently many experimental restraints is sufficiently close to a representative. Our method employs a branch-and-bound algorithm to search a configuration space representation of the subunit and complex structure, identifying regions containing the structures that are most consistent with the data. We apply our method to three test cases with experimental data and demonstrate that it can handle the difficult configuration space search problem and substantial ambiguity, effectively pruning the configuration spaces and characterizing the actual diversity of structures supported by the data.

Categories and Subject Descriptors

F.2.2 [Theory of computation]: Nonnumerical algorithms and problems; I.2.8 [Computing methodologies]: Problem solving, control methods, and search; J.2 [Computer applications]: Physical sciences and engineering; J.3 [Computer applications]: Life and medical sciences

General Terms

Algorithms

1. INTRODUCTION

Symmetric homo-oligomers are comprised of subunits that are identical in sequence and highly similar in structure and are arranged symmetrically; we study here cyclic symmetry, in which the subunits are placed like spokes on a wheel (Fig. 1). Symmetric homo-oligomers are thought to make up a majority of proteins; they play important roles in biological processes that include ion transport and regulation, signal transduction, and transcriptional regulation [7]. They are therefore a valuable target for structural studies, and nuclear magnetic resonance spectroscopy (NMR) provides the ability to analyze their structures and dynamics in solution.

Nuclear Overhauser Enhancement Spectroscopy (NOESY), which measures through-space interactions, provides the main source of structural information in standard NMR protocols. Nuclear Overhauser Effect (NOE) intensities are converted into distance restraints and assigned to pairs of protons, giving upper bounds on their interatomic distances. NOE distance restraints are typically used to frame NMR structure determination as an optimization problem combining biophysical modeling terms with pseudo-energy encodings of the constraints [8, 3, 9]. Statistical inference techniques have also been developed to combine modeling and experimental terms and characterize the resulting posterior distributions of structures [20, 4].

In a symmetric homo-oligomer, the high structural similarity of the subunits yields highly similar chemical environments for their atoms, rendering an atom in one subunit indistinguishable from the corresponding atom in another subunit under standard NMR experiments. Consequently, an NOE may involve two atoms in the same subunit or in two different subunits (with multiple pairs of subunits possible for symmetry higher than two) or both (in a mixture) [17], with no inherent means to resolve this ambiguity (Fig. 1). Some experimental strategies have been devised to separate intra- vs. inter-subunit restraints [11, 13, 27, 24], but they are difficult and have met with limited success.
The problem of determining the structure of a symmetric homo-oligomer from ambiguous NOEs was formulated by Nilges and coworkers following the standard NMR approach of optimizing a pseudo-energy function combining biophysical modeling terms and symmetry-enforcing restraints [17]. A simulated annealing protocol is employed to optimize the pseudo-energy, initially considering both intra- and inter-subunit interpretations for the restraints. After a round of optimization, the intra- vs. inter-subunit interpretations are reassessed; those that are inconsistent with the identified structures are eliminated. Another round of optimization is then initialized from the structures and reduced set of restraint disambiguations. The process is repeated in order to identify mutually consistent structures and disambiguations. This approach, called ARIA, has been used to determine a number of symmetric homo-oligomeric structures from ambiguous NOE data (see aria.pasteur.fr/aria-links/pdb-structures-calculated-using-aria), and has continued to expand in functionality, e.g., incorporating additional types of restraints such as RDC. The method, however, fails to give any assurances on the correctness of the obtained structures or to account for the possibility of missing structures. There are no guarantees that the heuristics to escape local minima will work, or that the greedy selections of disambiguations will lead to the native structures.

An alternative approach is based on Rosetta [21]. It first computes monomer structures that satisfy chemical shifts and backbone NOEs; critically, it requires these NOEs already to have been assigned to specific atoms, with intra- vs. inter-subunit ambiguity resolved. It then constructs a complex structure by docking the monomers subject to a symmetric constraint, while also incorporating RDC-derived orientational constraints. Energy refinement is subsequently performed to generate final structures. This method is clearly susceptible to being trapped in local minima by the two-stage process, as well as by the use of stochastic search methods in each stage.

In order to correctly and completely characterize the diversity of structures consistent with the data, we present here a method that employs a branch-and-bound search over a configuration space representation of the subunit and complex structures. The approach builds on our earlier work in structure determination of symmetric homo-oligomers [5, 15, 18, 4], which handled special cases where special experimental techniques successfully resolved the intra- vs. inter-subunit ambiguity. The method presented here goes significantly further, simultaneously determining the subunit structure and the arrangement of the subunits into a complex structure, while partially disambiguating the intra- vs. inter-subunit interpretations of the input NOE distance restraints. Since multiple structures may be consistent with the data, we provide a guarantee of completeness to within a user-specified resolution: our method generates a set of representative structures such that any other structure that is sufficiently consistent with the data has a sufficiently similar representative. As we discuss further below, we determine just the backbone structure within secondary structure elements (SSEs), though the resulting SSE-based structures could readily be used as inputs for loop closure algorithms using additional experimental data [22], with sidechain packing and energy minimization algorithms then employed to obtain the best full conformations.

2. METHODS

Our overall goal is to take as input a protein sequence and experimental NMR data, and produce as output the complex structure. We leverage earlier work on determination of SSE backbone structures from residual dipolar coupling (RDC) data: RDC-Panda [26] employs a tree-based search algorithm to find sequences of backbone torsion angles in the SSEs that best explain the experimental RDC data. The RDC data also allows determination of the orientation (but not position) of the symmetry axis (uniquely for 3-fold and higher, one of the three eigenvectors of the Saupe Matrix for 2-fold) [2].

Thus we focus here on the problem in which the input includes the number of subunits (a scan over possibilities could be performed [18]), the SSE backbone structures, the symmetry axis orientation, and a set of ambiguous NOE distance restraints, and the output is a structure placing the SSE backbone structures relative to each other and to the symmetry axis (Fig. 2), thereby generating the subunit and complex structure. As discussed in the introduction, there is not likely to be a unique such structure best satisfying the restraints, so the output is actually a set of representatives, such that any other structure satisfying a sufficient number of restraints is sufficiently similar. We define “sufficiently similar” in terms of root mean squared distance (RMSD), and a “sufficient number of restraints” relative to the best solution. We note that RMSD is not a metric in the mathematical sense, and simply use it to decide when to stop branching on small regions in the SCS. Thus our method can be seen as complete to within a user-specified resolution, and does not suffer from problems of myopia, local minima, and so forth.

We first detail the representation of the structure and restraints, including the various types of ambiguity. Then we develop methods to assess entire cells within the configuration space for consistency with the restraints under the ambiguous interpretations and to assess the uniformity of structures within the cells. Finally we develop a branch-and-bound algorithm and postprocessing analysis to identify the representative structures.

2.1 Representation

As summarized in Fig. 2, we place the global origin at the center of mass of the backbone atoms of one of the SSEs (the
Figure 2: Configuration space representation for simultaneous determination of subunit and complex structure of symmetric homo-oligomers. The backbone SSE structures and symmetry axis orientation are determined from RDC data, so our problem is to place the SSEs relative to each other and to the symmetry axis based on ambiguous NOE distance restraints. The configuration space is parameterized by placing the center of one “fixed” SSE at the origin, and specifying the 3D translations $s_i$ of the centers of the other SSEs, as well as the position $t$ of the intersection of the symmetry axis with the $x$-$y$ plane (its orientation $a$ is precomputed). Specifying the $s_i$ then generates a subunit structure (here cyan), while rotating it around the axis $(a,t)$ generates a complex structure (here a dimer, with the second subunit in red). In assessing restraints, we also use the distances $k_{ij}$ of atom $j$ in SSE $i$ to the center of the SSE. For side-chain atoms, this requires determination of the side-chain conformation, which we choose from a set of rotamers.

“fixed” SSE, $\#0$ in one of the subunits (the “fixed” subunit, $\#0$). These choices are arbitrary in terms of experimental information. The center of mass of the $i$th SSE is then expressed as a 3D vector $s_i$, with $s_{00} = 0$. The position of the symmetry axis is specified by the translation $t$ of its intersection with the $x$-$y$ plane; its orientation $a$ is known.

The coordinates of the atom $j$ within SSE $i$ are specified by translation $k_{ij}$ from the SSE center. For backbone atoms, the translation is fixed, but for side-chain atoms, they depend on the side-chain conformation, which in turn is restrained by the data and guided to avoid steric clash. We adopt a rotamer-based representation of side-chain conformations, allowing them to vary over low-energy representations mined from experimental structures. We use discrete rotamers that are members of the “penultimate rotamer library” [14], but the method can use any such rotamer library. The side-chain translation vector for atom $j$ in SSE $i$ is thus a member of a set $K_{ij}$ precomputed from the rotamer library. For simplicity of notation, we also use such a set, containing a single member, for backbone atoms.

Putting together these parameters, we have a set $Q_{rij}$ of possibilities for the coordinates $q_{rij}$ of atom $q_{rij}$, atom $j$ in SSE $i$ and subunit $r$:

$$Q_{0ij} = \{ s_i + k_{ij} | k_{ij} \in K_{ij} \}$$

$$Q_{rij} = \{ R_a(\alpha)(q_{0ij} - t) + t | q_{0ij} \in Q_{0ij} \} \quad r > 0$$

where $R_a(\alpha)$ is a three dimensional rotation by angle $\alpha$ around axis $a$, where $\alpha = 2\pi/c$ for $c$ subunits.

The parameters $t$ and $s_i \ (i \in \{1, \ldots, m - 1\})$ define the backbone structure of a subunit with $m$ SSEs; determining them is our goal. We call their possible values, in $\mathbb{R}^2 \times (\mathbb{R}^3)^{m-1}$, the Symmetry Configuration Space (SCS). The side-chain rotamers are useful in assessing restraint satisfaction and avoiding steric clash, but need not be (and indeed likely are not) completely determined.

The NOE restraints are expressed in terms of norm inequalities on interatomic distances. In our representation, the restraint $(p, q, \delta)$, indicating that atoms $p$ and $q$ must be within distance $\delta$, becomes:

$$\exists \ p \in P, \ q \in Q \ s.t. \ ||p - q|| \leq \delta$$

where $P$ and $Q$ are sets of atom positions for atoms $p$ and $q$ respectively and $|| \cdot ||$ is the Euclidean distance. With respect to our representation, we know the atom and SSE indices of $p$ and $q$; if we also knew the subunit indices, then $P$ and $Q$ would be determined as one of the $Q_{rij}$. The fact that we do not know which subunits are involved in each restraint is the key confounding factor in simultaneous determination of subunit and complex structure.

Let us consider how to represent this ambiguity regarding the interpretation of a restraint: is it within a single subunit (“intra”) or between two subunits (“inter”), and if between two, which two. We note that experimentally the restraints are “mirrored”—an intra restraint is satisfied in all subunits, and an inter restraint is satisfied in all pairs of subunits the same spacing apart around the cycle. (This is why the choice of fixed subunit doesn’t matter.) In our approach, we consider all such interpretations. Only if the interatomic distance is too large in all interpretations (Eq. 3) do we consider the restraint to be violated. Thus we essentially express ambiguity as a logical OR over the interpretations.

There is likewise ambiguity in the side-chain conformation— we do not know which rotameric conformation is correct. For this ambiguity, we likewise use a logical OR to express the fact that, as long as some pair of rotamers places the atoms within $\delta$, the restraint is not considered violated.

One final source of ambiguity in our representation actually comes with the input SSE backbone structures, each of which is subject to a 180 degree rotation around each of the axes in a manner that yields 4 images of the structure, called “orientations” [26] (Fig. 3). We do not directly represent this ambiguity, but instead simply solve for each combination independently.

### 2.2 Cell-based restraint analysis

The configuration space gives a compact representation for all possible structures; our goal is to find within it re-
resentatives for all structures that are sufficiently consistent with the data. To consider the feasible region in SCS, let us for a moment ignore the various sources of ambiguity. The SCS-to-Euclidean conversion (Eq. 1 and 2) is a linear transformation, and each distance restraint (Eq. 3) specifies a ball in Euclidean space. The pseudoinverse of the transformation (which is indeed of rank 3) transforms the feasible ball to an ellipsoid in SCS. A feasible configuration is the ellipsoid crossed with the the null space, which makes it an infinite-length cylinder with an ellipsoidal cross-section. Thus we need to compute the intersections of these cylinders. Once we incorporate the ambiguity, however, we would have to compute an exponential number of such intersections.

Thus instead of trying to exactly compute the feasible portion of the configuration space, we subdivide the space into regions containing relatively similar structures and evaluate the discretized regions. In particular, we use a cell-based representation of configuration space regions, where a “cell” is an axis-aligned box \( T \times S_1 \times \ldots \times S_{m-1} \) with \( T \) an axis-aligned rectangle in \( \mathbb{R}^2 \) containing the symmetry axis translations and \( S_i \) an axis-aligned cuboid in \( \mathbb{R}^3 \) containing the translations of SSE \( i \). In evaluating a cell, we want to know how many restraints its various structures satisfy. The algorithm described below starts from a single cell covering a large SCS volume, and adaptively subdivides it to find cells representing structures satisfying the restraints.

First let us bound the positions of the atoms. Recall that Eq. 1 and 2 define the possible positions based on specific choices for \( t \) and \( s_i \); \( k_{ij} \) is constant when assessing a particular rotamer. Our cell representation allows \( t \) and \( s_i \) to range across axis-aligned boxes \( T \) and \( S_i \). Thus for an atom in the fixed subunit, extending Eq. 1 over \( s_i \in S_i \) simply displaces the box \( S_i \) by the vector \( k_{ij} \). For an atom in another subunit, Eq. 2, when expanded out, becomes:

\[
Q_{ij} = \{ R_a(\alpha)(s_i + k_{ij} - t) + t \mid t \in T, s_i \in S_i \} = \{ R_a(\alpha)(s_i - t) + t + R_a k_{ij} \mid t \in T, s_i \in S_i \} \tag{4}
\]

Again, recognizing that \( T \) and \( S_i \) are boxes, we can see that this is a linear transformation of a Minkowski difference of two convex polyhedra, another convex polyhedron. We compute the extreme points of this polyhedron by ranging \( t \) and \( s_i \) over only the corners of \( T \) and \( S_i \).

So for any backbone atom, we bound the possible positions, over the whole cell, with a convex polyhedron. For side-chain atoms, we have sets of polyhedra over the different rotamer-defined positions (translating by \( k_{ij} \)), and we employ an OR as described above.

The bound of atomic coordinates over a cell then enables us to assess steric clash, along with satisfaction of a restraint over all conformations defined by a cell.

**Steric clash.** The square of the distance between two atoms is a convex function, and thus its maximum is achieved at a pair of extremal points of the atoms’ bounding polyhedra. We test each such pair \((2^3 \times 2^3 \text{ for intra and } 2^3 \times 2^3 \times 2^2 \text{ for inter})\). If the maximum distance is less than 1.5 Å, we can infer that all structures in the cell exhibit steric clash for that atom pair. For efficiency, we only test pairs of atoms involved in NOEs.

**Completely satisfied.** If the maximum distance (as described for steric clash) between a pair of atoms in an NOE restraint is less than the NOE distance, the restraint is satisfied for every structure in the cell.

**Completely violated.** A restraint cannot be satisfied unless some pair of points, one for each atom’s bounding polyhedron, is within the NOE distance. Thus we simply compute the shortest inter-polyhedral distance, and consider the restraint to be violated for every structure in the cell if that distance exceeds the threshold.

While these tests allow us to evaluate the two extreme cases for each restraint, the overall quality of a structure rests on satisfaction of multiple restraints simultaneously. In contrast, with these tests the point used to evaluate one restraint may be different from that used for another restraint. Unfortunately an ability to exactly assess simultaneous satisfaction of an arbitrary set of restraints would also give us the ability to perform side-chain packing, an NP-hard problem [1]. Thus we develop an algorithm for limited simultaneous restraint satisfaction, correctly bounding the true evaluation that would be produced by a full assessment.

The algorithm is described formally in Alg. 1. We form position-specific sets such that \( R_i \) contains all restraints in which some atom from residue \( i \) participates. (Since each restraint has two atoms, it can appear in two such sets.) We work from N terminus to C terminus. When considering residue \( i \), we examine each of its possible rotamers, identifying the one (\( a \)) that supports the satisfaction of the most restraints. (We don’t allow a restraint to be considered satisfied for both its different residues.) To do so, for each remaining restraint involving an atom from residue \( i \), we consider all possible rotamers \( b \) for the other atom, and evaluate the resulting interatomic distance over the given cell. Here function \( d \) computes Euclidean distance after applying the configuration space transformation for parameters \( c \) and using the intra-SSE distance vectors \( k \) for rotamers \( a \) and \( b \), as in Eq. 1 and 2. After identifying the best such rotamer, we add to our list all the restraints it satisfies, and continue.

The resulting estimate of satisfied restraints is loose since when considering a residue position, only the rotamers for that position are necessarily used consistently over the restraints involving that position, while the rotamers for the other atoms in the restraints are unconstrained. We can show by induction that it is a correct overestimate.

**Theorem 2.1.** The size of set \( S \) at iteration \( i \) of Alg. 1 is an overestimate of the size of the optimal set of restraints from \( R_0 \cup \ldots \cup R_i \) that can be satisfied when choosing for each position \( i \) through \( i \) a unique rotamer.

**Proof.** Let the set computed by our algorithm be denoted by \( X_i \) and the optimal set (with a unique rotamer per

---

**Algorithm 1** Simultaneous restraint satisfaction bound

**Input:** Sets \( R_i \) (1 \( \leq i \leq n \)) of restraints with at least one atom from residue \( i \)

**Input:** Sets \( A_i \) (1 \( \leq i \leq n \)) of rotamers for residue \( i \)

**Input:** Cell \( C \).

**Output:** \(|S_i|\): bound on number of satisfied restraints

\[
S_i \gets \emptyset
\]

for \( i = 1 \rightarrow n \) do

\[
U \leftarrow R_i \setminus S_i
\]

\[
a \leftarrow \arg \max_{a \in A_i} \{ (p, q, \delta) \in U \mid \exists b \in A_{\text{real}(q)}, \ c \in C \text{ s.t. } d(p, q, a, b, c) \leq \delta \}
\]

\[
S_i \leftarrow S_i \cup \{ (p, q, \delta) \in U \mid \exists b \in A_{\text{real}(q)}, \ c \in C \text{ s.t. } d(p, q, a, b, c) \leq \delta \}
\]

end for
position) by $O_i$. The proof is by induction. For the base case $i = 1$, our algorithm finds the maximum number of satisfied restraints using any rotamer at position 1, an overestimate of the actual number which would restrict the other side of the restraint; i.e., $|X_i| \geq |O_i|$.

For the inductive step we will prove that if our hypothesis holds for $i$, i.e., $|X_i| \geq |O_i|$, then it also holds for $i + 1$. Assume for contradiction $|X_{i+1}| < |O_{i+1}|$. Let $O_i' = O_{i+1} \cap (R_1 \cup \ldots \cup R_i)$; note that it might be completely different from $O_i$. Let $\Delta_{i+1} = R_{i+1} \setminus (R_1 \cup \ldots \cup R_i)$ be the new restraints involving only positions $i+1$ and higher. We have two possibilities:

1. $|O_i'| > |O_i|$. This immediately contradicts the optimality of $O_i$.

2. $|O_i'| \leq |O_i| \leq |X_i|$. Then for $|O_{i+1}| > |X_{i+1}|$, it must be that $|O_{i+1} \cap \Delta_{i+1}| > |X_{i+1} \cap \Delta_{i+1}|$. But since $\Delta_{i+1}$ includes only restraints between residues $i + 1$ and higher and the rotamer consistency requirement of our algorithm only applies to $i + 1$, any restraint added to $O_i'$ for $O_{i+1}$ can also be added to $X_i$ for $X_{i+1}$. Thus $|O_{i+1} \cap \Delta_{i+1}| \leq |X_{i+1} \cap \Delta_{i+1}|$, a contradiction.

In either case we derive a contradiction, so it must be that $|X_{i+1}| = |O_{i+1}|$, and the induction carries through. □

As a corollary, when it terminates at $i = n$, Alg. 1 produces an overestimate of the size of the optimal set of consistently satisfied restraints.

2.3 Cell structural uniformity assessment

In searching the configuration space, we need to be able to determine whether a cell represents a more-or-less uniform set of structures. We adopt the criterion here that all structures must be within a user-specified RMSD to each other. We now develop an estimate for the maximum RMSD between structures within a cell, without having to convert the continuous set in configuration space to conformation space.

First let us consider the distance between a particular structure and any other structure in the cell, as follows. Given a point $x$, at particular $t$ and $s_i$ (for $i$ ranging over the SSEs), define function $g_x(y)$ as the square of the RMSD to the fixed $x$ from another point $y$ in the cell, at $t + \Delta t$ and $s_i + \Delta s_i$. For an atom in the fixed subunit, its contribution to the squared distance is $||x - y||^2 = ||\Delta s_i||^2$. For an atom in another subunit (at defined rotation angle $\alpha$), the squared distance is similarly

$$SD = \left(\sum_{ij} (R_{\alpha}(\sigma_i)(s_{ij} + k_{ij}) - t + \Delta t)^2\right)$$

$$= \left(\sum_{ij} (R_{\alpha}(\sigma_i)(s_{ij} + \Delta s_i + k_{ij}) - t + \Delta t)^2\right)$$

By inspection we can determine that the lower corner (with smallest value for each element) and upper corner (with largest for each) yield the largest $u$, as they provide the maximal $||\Delta s_i||$.

2.4 Search algorithm

We now develop a branch-and-bound search algorithm that hierarchically subdivides the SCS, using the cell-based evaluations to assess restraint satisfaction within the cells and to identify terminal cells that need not be further divided. Ultimately the search identifies cells that satisfy the most restraints and that have sufficiently uniform structures. Some of the cells might be similar to each other, so a clustering process yields the final set, from which are generated representatives for all structure sufficiently consistent with the data.

The search is initialized with a cell that is a cross product of $S_i$ and $T$ that are sufficiently large to contain all SCS points that could satisfy the restraints. The SCS with the most restraints to others is established as the origin, and the symmetry axis is taken as the $z$ axis to restrict how far away the subunits can be situated. Initially each restraint could be interpreted as either intra or inter. However, we can eliminate some of the intra possibilities prior to beginning the search, by identifying pairs of atoms in the same SSE (and thus independent of SCS choices) that cannot be within the NOE distance under any choice of rotamers. Constraints for such pairs must be considered as inter only. This preprocessing does not change the results, but reduces the size of the search space that must be explicitly considered and the number of tests that must be performed during the search.

The search maintains a priority queue of cells and associated viable restraint interpretations. Priority is determined by the number of violated restraints in a cell. The search also maintains a cutoff value of the fewest violated restraints by any terminal cell.

When a cell is removed from the priority queue, it is subdivided along its longest dimension. The child cells are as-
sessed for structural uniformity and for restraint satisfaction, as follows:

- Cells that are sufficiently uniform (we use a threshold of 1 Å RMSD for our results) are considered terminal and tested by Alg. 1 for restraint satisfaction, updating \( \tau \) when appropriate.

- Cells that are sufficiently small (we use a threshold of 2 Å in each dimension) are tested for complete satisfaction (the expense of the test is not justified for larger cells). If the number of restraints that aren’t completely satisfied is at most \( \tau \), then the cell appears good but not sufficiently uniform. Thus we repeatedly subdivide it until the subcells are sufficiently uniform. We consider them terminal and continue as above.

- Other cells are tested for complete violation. If more than \( \tau \) many restraints are completely violated, the cell is pruned. Otherwise it is added to the priority queue.

Upon termination, we reevaluate the terminal cells, ordered by the number of violations, against the final \( \tau \). Those having fewer violations than the final \( \tau \) are tested for steric clash within the structure represented by the cell center. Those that pass are considered accepted cells. The accepted cells are clustered to reduce redundancy while still ensuring that all satisfying structures are represented by a sufficiently close solution. The clustering is performed using the Euclidean distance metric on a KD-Tree constructed from the cell centers. The RMSD upper bound used to assess structural uniformity in the cell (Sec. 2.3) also enables us to use Euclidean distances on SCS points after scaling the coordinates appropriately.

Representative structures are generated from the centers of the final clustered cells. For each cell center, an individual subunit is generated from the SSE translations; the complex is then generating by rotating around the translated symmetry axis \( m - 1 \) times.

3. RESULTS

We applied our approach to three test cases with experimental NOE restraints: (1) MinE [12] (PDB id 1EV0), a dimer with one \( \alpha \)-helix and two \( \beta \)-strands restrained by 1109 NOEs (926 intra + 183 inter); (2) B. subtilis Anti-TRAP (PDB id 2KO8), a trimer with one \( \alpha \)-helix and two \( \beta \)-strands (along with a third that is unrestrained and thus not considered here) restrained by 863 NOEs (378 intra + 485 inter); (3) the cytoplasmic domain structure of BM2 proton channel from influenza B virus [25] (PDB id 2KJ1), a tetramer with two \( \alpha \)-helices restrained by 400 NOEs (340 intra + 60 inter). NOEs were obtained from the BioMagResBank (BMRB) [23] and the intra vs. inter resolution was ignored. We used the deposited SSE backbone structures and axis orientation, since these proteins lacked the RDC data necessary to determine them by RDC-Panda. The RMSD cut-off for cell uniformity was set to 1 Å and the initial maximum NOE restraint violation \( \tau \) was set to ten percent of the total number of restraints.

We first determined which restraints supported only an inter interpretation, as the involved atoms were in the same SSE but no rotamer choice could place them close enough. For MinE, 104 restraints were classified as inter, all consistent with the deposited interpretation. For Anti-TRAP, 173 restraints were classified as inter, but 60 of them were actually intra according to the deposited interpretation. For BM2, 8 restraints were classified as inter, 4 of which were actually intra. The misclassifications were due to the use of discrete rotamers, which did not come sufficiently close; possible fixes include relaxing the distance threshold or using rotamer “voxels” [6]. Note that while the preprocessing forced some incorrect interpretations, they are due to the geometric model and the same interpretations would ultimately have resulted from the search algorithm; the preprocessing is simply a time-saving measure.

We now characterize our results in terms of the identified feasible region of the configuration space and the structures contained within it. We show that, even with significant intra vs. inter ambiguity and a large, complex configuration space, the algorithm is able to identify compact feasible regions most consistent with the data. We also show that the resulting structures identified by our method capture the variability in the deposited ones. We further show that ours are substantially more diverse than those in the deposited ensemble, though we recognize that by focusing just on the SSEs, our results overestimate the structures consistent with the data (as NOEs and packing with loops could further constrain the allowable conformations).

3.1 Configuration space search

The configuration space search yields a set of accepted cells representing the feasible regions; each cell specifies the 2D translation of the symmetry axis relative to the fixed subunit (\( T \)) and the 3D translation of each SSE relative to the fixed SSE (\( S_i \)). Recall that we do the search independently for each set of SSE orientations. Fig. 4 illustrates the accepted cells for the most populated orientation set (i.e., the one with the largest volume) for each of our three test cases. Note that while the different components of the cells are displayed separately, not all combinations of these components are accepted.

For MinE, eight of the sixteen SSE orientation combinations led to accepted cells. The two most populated combinations were nearly equal in number and different from the deposited structure. The first combination (34% of the remaining volume) had the sheet rotated around \( z \) axis and the second (33% of the remaining volume) had both beta strand and alpha helix rotated around the \( z \) axis. The combination of the deposited structure contained 13% of the remaining volume, as did another combination with the alpha helix rotated around the \( z \) axis. We later discuss the resulting conformations, but wanted to point out here that the differences in orientations in the configuration space indeed lead to differences in conformations; e.g., the average RMSD for samples in the two most populated orientation sets was 4.6 Å with a maximum of 12.8 Å. The remaining volume of the translational component of the symmetry axis was \( 2 \times 10^{-5} \) that of the initial volume. The symmetry axis was highly constrained by inter-subunit restraints.

For the SSEs, the relative volumes of the translational components were \( 1 \times 10^{-5} \) for SSE 1 and \( 2 \times 10^{-4} \) for SSE 2. The \( \beta \)-strand was restrained to the fixed SSE by 237 NOEs, yielding a relatively restricted remaining translational component (10 Å\(^2\)). The \( \alpha \)-helix, in contrast, was relatively unrestrained, with only 26 NOEs to the fixed SSE, resulting
in much more translational uncertainty (32 Å³). There are no intra restraints between the two non-fixed SSEs and the inter restraints are therefore valuable in pinning down the structure.

There were accepted cells for five of the sixteen possible SSE orientation combinations for Anti-TRAP; the most populated combination (72% of the remaining volume) was the same as in the deposited structure while the other combinations produced relatively few cells. The second most populated combination (14%) was the combination in which the beta strand (residues 9-11) is rotated around the z axis. The translation component volume for the largest orientation combination was $2 \times 10^{-3}$ that of the total volume, while the SSE translation volumes were $2 \times 10^{-4}$ and $1 \times 10^{-4}$ those of the originals. There was much more uncertainty in the position of the symmetry axis here, compared to MinE, due to significantly fewer unambiguous inter-subunit restraints characterized during the preprocessing. There was also substantial uncertainty in the translations of the SSEs, 89 and 104 Å³, as various combinations of ambiguous assignments of different restraints allowed the cells to escape pruning. Interestingly, SSE 2's translation cells fell into two distinct groups, with the second SSE much further away from the fixed one in one than in the other.

For BM2 only one SSE orientation combination, that in the deposited structure, produced accepted cells. $7 \times 10^{-4}$ of the symmetry axis translation volume was accepted, while $1 \times 10^{-5}$ of the SSE translation volume remained. Though there are only four inter-subunit restraints between the fixed SSE in each subunit, the axis translation was tightly characterized (7 Å³), and likewise the two restraints from SSE 1 to the fixed SSE sufficed to reduce its translational uncertainty to 47 Å³. The NOEs acted in concert with backbone steric clash to drastically prune the configuration space.

3.2 Example structures

Clustering the accepted cells to reduce redundancy (while still representing all solutions) resulted in 1229 representatives for MinE, 377 for Anti-TRAP, and 195 for BM2. Representative structures were generated from the center of each such cell. To illustrate the diversity of structures, we performed further agglomerative clustering and selected an example from each of the most distinct groups. See Fig. 5.

MinE. The top level of the dendrogram represents an RMSD of 10 Å, indicating substantial diversity in the structures. However, chopping the tree into 8 clusters yields compact groups, each with no more than 1.5 Å RMSD among its members. Fig. 5(top) illustrates one sample from each cluster. Between these samples, the non-fixed alpha helix for the first subunit had an RMSD up to 24 Å, while the beta sheet had an RMSD up to 16 Å.

Anti-TRAP. The accepted cells yielded much more similar structures, with a maximum RMSD at the top of the dendrogram of 4 Å. There are only three clusters that have a maximum RMSD of 3.5 Å in structures within them. The example structures in Fig. 5(middle) illustrate this relative uniformity of identified representatives. The SSE 2a had the most variance between these samples, as much as 23 Å.

BM2. This structure was the best determined, with an RMSD of only 0.7 Å at the top of the dendrogram, resulting from the relatively compact set of accepted cells. The six example structures from the top-most clusters (Fig. 5, bottom) emphasize this point.

3.3 Comparison to deposited structures

The structures identified by our method represent the deposited structures well. Fig. 6 shows that for MinE, the minimum-energy deposited structure is 0.94 Å away from the closest member in our ensemble. Similarly, the minimum-energy deposited structure for Anti-TRAP is represented by

![Figure 4: Accepted SCS cells for the test cases. (top) The translation of the x-y intersection of the symmetry axis. (middle, bottom) The translation of the non-fixed SSEs. For MinE, one β-strand is fixed while the other β-strand (middle) and the α-helix (bottom) are translated. For Anti-TRAP, the α-helix is fixed and the β-strands (middle and bottom) translated. For BM2, one α-helix is fixed and the other (middle) translated.](image-url)
Figure 5: Diverse example structures from satisfying SCS cells. The subunits are colored differently with the subunit containing the fixed SSE colored in black.

Figure 6: Superpositions of lowest-energy deposited structure (red) and closest representative from our search (blue). Parenthesized numbers are RMSDs.
4. CONCLUSION

We have developed an approach to fully account for intra- vs. inter-subunit ambiguity in NOE data for symmetric homooligomers, simultaneously determining the subunit and complex structures most consistent with the data. This new approach builds on and brings together work we have done on subunit structures alone and on assembly of complex structures given subunit structures. In contrast to other approaches that are heuristic in nature and can get trapped in local minima, our approach partitions a configuration space that represents all possible structures, using a set of restraint satisfaction tests to identify the regions that best satisfy the data. Our search procedure enables us to provide guarantees on the results, namely that any structure sufficiently consistent with the data is sufficiently close to one of the identified representatives.

We employ a powerful yet limited configuration space representation of complex structures. We focus on rigid SSEs and rigid rotamers, allowing for some flexibility or uncertainty in those by relaxing NOE distances. A more rigorous approach (further expanding the scope of the completeness guarantee) would be to further expand the configuration space with “voxels” around the SSEs and rotamers (analogous to [10]). Our approach does not currently consider the loops connecting the SSEs, though NOEs involving loops may also be useful in pruning the space. It remains future work to integrate loop closure [22] as part of an “end-to-end” complete configuration space-based structure determination software. An additional component of an end-to-end approach involves other ambiguities, including NOE assignment (similar to [19]), as well as multiple possible symmetry axes [16]. We note that the configuration space approach is sufficiently general to extend to other types of complex structures, beyond symmetric homo-oligomers.

We demonstrated with three test cases that our approach effectively prunes the search space and identifies diverse structures consistent with the data. As a branch-and-bound style algorithm, theoretical characterizations of expected efficiency and output size cannot readily be determined. In practice we found the wall-clock time to range from less than an hour to few hours with a multithreaded implementation on an eight-core machine, and the number of representative structures to range from about two hundred to over thousand. These measures depended on the number of subunits and the number of ambiguous restraints remaining after initial classification as intra vs. inter. As more test cases with the necessary data become available, it will be interesting to characterize the scaleability of the algorithm with the extent and quality of the data, desired resolution, and complexity of the resulting structure. In the mean-time, additional simulation studies can help characterize the impacts of these parameters on the run-time and output size.

Software is available from the authors by request.

Acknowledgement

We thank Jeff Martin, Kyle Roberts, and other members of the Donald lab for helpful comments. This work was supported by National Institutes of Health grants NS-79929 (to B.R.D.), GM-65982 (to B.R.D.), and GM-78031 (to B.R.D.), along with NSF grant CCF-0915388 to C.B.K. We also gratefully acknowledge computational resources provided by NSF grant CNS-1205521.

ACM-BCB 2013
REFERENCES