

Extended Abstract: Structure Determination of Symmetric Protein Complexes by a Complete Search of Symmetry Configuration Space Using NMR Distance Restraints

Shobha Potluri¹ Anthony K. Yan¹ James J. Chou²
Bruce R. Donald^{1,3,4,5} Chris Bailey-Kellogg^{1,5}

¹ Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA

² Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

³ Department of Chemistry, Dartmouth College, Hanover, NH 03755, USA

⁴ Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA

⁵ Corresponding authors. 6211 Sudikoff Laboratory, Hanover, NH 03755. Email: {brd,cbk}@cs.dartmouth.edu

Symmetric homo-oligomers are protein complexes with similar subunits arranged symmetrically [10]. Figure 1 illustrates the structure of a symmetric homo-oligomer called phospholamban. Phospholamban is a membrane protein that helps regulate the calcium level inside the cell and hence aids in muscle contraction and relaxation [7]; ion conductance studies [5] also suggest that phospholamban might have a separate role as an ion channel. A detailed molecular-level understanding of homo-oligomeric structures provides insights into their functions and, in some cases, how to design appropriate drugs. Nuclear Magnetic Resonance (NMR) spectroscopy underlies many structural studies of homo-oligomers, but poses significant computational challenges in inferring three-dimensional structures from indirect (and often sparse) measurements of geometry.

We use two types of information in homo-oligomeric structure determination: distance restraints from nuclear Overhauser effect (NOE) data, and biophysical modeling terms evaluating packing quality. An inter-subunit distance restraint is of the form $\|\mathbf{p} - \mathbf{q}'\| \leq d$, where \mathbf{p} and \mathbf{q}' are atoms in different subunits of the complex, and d is the given distance for the restraint. We say that a structure is consistent with a distance restraint if \mathbf{p} and \mathbf{q}' are within d Å of each other. The experimental data are complemented by biophysical models of the (non-covalent) interactions that stabilize complexes. Figure 1(b,c) illustrates that the atoms of adjacent subunits of phospholamban are well-packed, interacting at just the right distance to hold the complex together. Packing interactions are typically evaluated with functions that model the van der Waals (vdW) energies between the atoms forming the complex [1, 4]. Our approach separately accounts for experimental data and biophysical modeling terms,

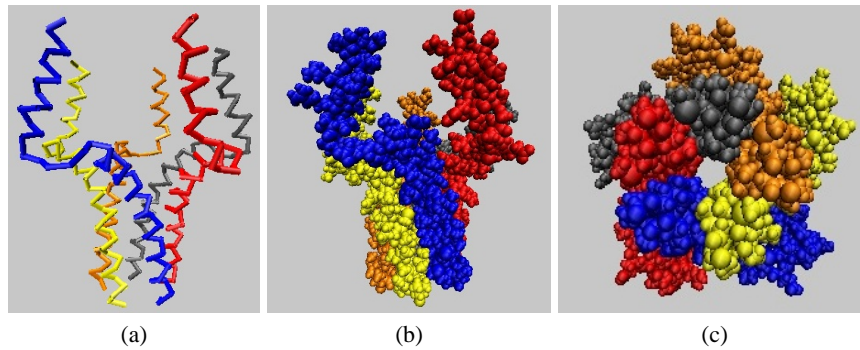


Fig. 1. Structure of Phospholamban. The five subunits are shown in different colors. (a) Wireframe (backbone trace) representation. (b) Van der Waals sphere representation of all the atoms. (c) Van der Waals sphere representation, viewed down the symmetry axis and illustrating the 5-fold symmetry.

and ultimately finds structures of symmetric homo-oligomers that are consistent with the inter-subunit distance restraints and that display high-quality inter-subunit packing interactions.

Traditional protocols [6] for structure determination of protein complexes from NMR data use simulated annealing and molecular dynamics to optimize a pseudo-potential combining both biophysical terms (including packing interactions) and terms evaluating consistency with experimental data. The goal is to find low-energy conformations, but these techniques may become trapped in local minima and miss structures consistent with the data. The precision in the determined structure is also strongly affected by the annealing temperature. Further, since these approaches combine data and packing, they cannot identify the contribution to the structure from the experimental data alone versus both data and packing. Alternative docking-based approaches [2, 3, 8] for structure determination typically involve a two-stage approach: generate a set of possible docked structures, and then score them. The possible structures are generated by a heuristic and/or grid-based sampling of the space of rotations and translations of one subunit with respect to another. The generated structures are scored by geometric/energetic functions, and can be filtered based on symmetry. However, the sampling in the generation step does not account for consistency with the data and thus may miss consistent structures. Wang et al. [11] developed a branch-and-bound algorithm to compute rigid body transformations satisfying potentially ambiguous inter-subunit distance restraints. In contrast to this approach, our algorithm exploits the kinematics of the ‘closed-ring’ constraint due to symmetry, and thereby derives an analytical bound for pruning, which is tighter and more accurate than the previous randomized numerical techniques.

Our approach, described in detail in [9], is *complete* in that it tests *all* possible structures, and it is *data-driven* in that our algorithm has two separate stages where the first stage only tests structures for consistency with the data, and the second stage

evaluates the consistent structures for vdW packing. Completeness ensures that our algorithm does not miss any solutions because it returns a superset of all structures which are consistent with the data. This avoids bias in the search, as well as any potential for becoming trapped in local minima. The data-driven nature of our method allows us to independently quantify the amount of structural constraint provided by data alone, versus both data and packing. This avoids over-reliance on subjective choices of parameters for energy minimization [1], and consequent false precision in determined structures.

Given a set of inter-subunit NOE restraints, the subunit structure and oligomeric number (number of subunits forming the complex) as input, our approach determines the 3D structure of a symmetric homo-oligomer. (We note that it is possible to experimentally determine the subunit structure prior to computing the complex [7].) Given a single (fixed) sub-unit structure, the entire structure of the homo-oligomer is determined by the position and orientation of the symmetry axis. We take a configuration space-based approach and represent each possible structure of the symmetric homo-oligomer by a point in the four-dimensional space of symmetry-axis parameters, which we call the *symmetry configuration space* (SCS), $S^2 \times \mathbb{R}^2$. Geometrically, a point in \mathbb{R}^2 represents the position of the symmetry axis, and a unit vector in S^2 gives the orientation of the symmetry axis. We must identify all points in SCS representing symmetry axes that lead to structures consistent with the given set of inter-subunit distance restraints. Let $R_{\mathbf{a}}(\theta) \in SO(3)$ be a rotation around the unit vector \mathbf{a} by $\theta = 2\pi/n$ radians, where n is the oligomeric number. Let $\mathbf{t} \in \mathbb{R}^2$ be the point where the axis of rotation pierces the xy -plane, specifying the location of the symmetry axis. For an atom \mathbf{q} in the fixed subunit, the corresponding atom in the adjacent subunit, \mathbf{q}' , when the symmetry axis is at (\mathbf{a}, \mathbf{t}) , is obtained as $\mathbf{q}' = T_{\mathbf{at}}(\mathbf{q}) = R_{\mathbf{a}}(\theta)(\mathbf{q} - \mathbf{t}) + \mathbf{t}$. We wish to find the set

$$M = \{(\mathbf{a}, \mathbf{t}) \mid \mathbf{a} \in S^2, \mathbf{t} \in \mathbb{R}^2, \|\mathbf{p} - T_{\mathbf{at}}(\mathbf{q})\| \leq d \forall \text{ ordered triples } (\mathbf{p}, \mathbf{q}, d) \in D\}, \quad (1)$$

where D is the set of inter-subunit distance restraints, each specifying atoms \mathbf{p} and \mathbf{q} in the fixed subunit and distance d . A restraint constrains the maximum distance between \mathbf{p} and $T_{\mathbf{at}}(\mathbf{q})$, the atom corresponding to \mathbf{q} in the adjacent subunit when the symmetry axis is at (\mathbf{a}, \mathbf{t}) . The set M corresponds to all points in SCS that satisfy all the restraints.

In order to compute the set M , we perform a search over the SCS. The SCS is too large to search naïvely or exhaustively. Therefore, we have developed a novel branch-and-bound algorithm to search the SCS that is efficient and provably conservative in that it examines and conservatively eliminates regions in SCS inconsistent with the data. Without this algorithm, a complete, data-driven search would not be computationally feasible. The branch-and-bound search performs a search of the SCS by hierarchically subdividing it. Each node in the tree is a SCS *cell*—a 4-dimensional hypercuboid defined by values representing extrema along each of the four dimensions. At each node of the hierarchical subdivision, we test whether any point in the cell represents a consistent structure. If such a point possibly exists, we *branch* and partition the cell into smaller sub-cells. We continue branching until we

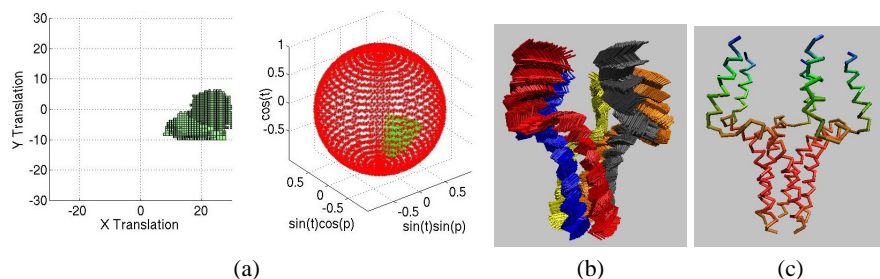


Fig. 2. Phospholamban Results: (a) Region of 4D space output by our branch and bound approach using the nine experimental restraints and knowledge of C_5 symmetry. The solution space of translation parameters and rotation parameters (theta angle denoted by t and phi angle denoted by p) on a sphere is shown. (b) The set of WPS structures after alignment to the structure with lowest packing score. Different subunits are in different colors. (c) Variance of the atoms illustrated by a color scale with blue indicating maximum variance and red minimum variance.

can either *eliminate* or *accept* each cell. We *eliminate* a cell when all the structures represented by the cell violate at least one restraint (see below) or contain several atoms that significantly clash with each other. We conservatively *accept* a cell as part of the consistent regions when all the structures it represents either provably satisfy all the restraints or are within an RMSD (root mean square deviation) of τ_0 Å (a user-defined similarity level) of each other and each restraint is satisfied by at least one structure represented by the cell. At the end of the branch-and-bound search, we return regions in SCS, the *consistent regions*, which provably contain all structures that are consistent with the data.

To test whether we can eliminate a cell G due to restraint violation, we independently consider each restraint, $\|\mathbf{p} - \mathbf{q}'\| \leq d$. We would like to compute $G\mathbf{q}$ (recall that \mathbf{q} corresponds to \mathbf{q}' in the fixed subunit), the set of all possible positions of \mathbf{q}' under the symmetries defined by G . Since the region $G\mathbf{q}$ is characterized by high-degree polynomials and it is computationally expensive to test for intersections with $G\mathbf{q}$, we approximate $G\mathbf{q}$ by a *conservative bounding region* that completely contains $G\mathbf{q}$. If there is an empty intersection between the conservative bounding region and the ball of radius d centered at \mathbf{p} , then all the structures represented by G violate the restraint and we eliminate G .

Figure 2(a) shows the consistent regions in SCS for phospholamban based on the nine experimentally-determined distance restraints. For the sake of illustration, we show the consistent regions as separate 2- d projections into S^2 and \mathbb{R}^2 . The volume of the consistent regions in the SCS is $1.24 \text{ \AA}^2\text{-radian}^2$. This volume indicates the constraint on structure provided by data alone. The larger the volume, the lesser the constraint.

Once the consistent regions have been identified, we choose *representative structures* from them such that every structure in the consistent regions is within an RMSD of τ_0 Å to at least one representative structure. Note that this sampling is different

from the sampling in docking-based approaches in that the native structures are always within τ_0 Å to at least one of the representative structures. Due to the conservative bounds used in our search, the representative structures might contain structures that are inconsistent with the data. The set of *satisfying structures* includes only those representative structures with restraint satisfaction scores below a chosen threshold. We then evaluate each of the satisfying structures by energy-minimizing and scoring them based on van der Waals packing. The set of *well-packed satisfying (WPS) structures* includes those energy-minimized satisfying structures with van der Waals packing scores below a chosen threshold. Thus, we ultimately return a set of structures consistent with data and packing representing any consistent, well-packed structure to within an RMSD of τ_0 Å.

The structural uncertainty in a set of structures can be quantified by the average variance in the positions of the atoms. The satisfying structures of phospholamban have a variance of 12.32 Å; the incorporation of vdW packing reduces this to 6.80 Å for the well-packed satisfying structures. Figure 2(b) illustrates the set of WPS structures for phospholamban. Figure 2(c) illustrates the variance of the atoms in the set of WPS structures. There is less uncertainty in the lower half of each subunit than in the upper half, since there are more experimental restraints in the lower half. Our complete approach hence allows us to identify the atoms of the complex that have high structural uncertainty. Further, it allows us to separately quantify the amount of structural constraint provided by data alone (satisfying structures), versus data and packing (WPS structures).

Our approach also provides for an independent verification of the oligomeric number, which is typically determined using experiments such as chemical cross-linking followed by SDS-PAGE, or by equilibrium sedimentation. We determine the oligomeric number by extending our search space to include a search over possible oligomeric numbers. We refer to this extended space as the *extended symmetry configuration space* (ESCS), $\mathbb{Z}_9 \times S^2 \times \mathbb{R}^2$, where \mathbb{Z}_9 is the set of possible oligomeric numbers of 2 to 9. We first obtain the set of WPS structures for each oligomeric number. We immediately prune out those oligomeric numbers that have no WPS structures. This allows us to determine the oligomeric number with high certainty when only a single oligomeric number has WPS structures. When several oligomeric numbers have WPS structures, we determine the oligomeric number as follows. Let $E_l(m)$ and $E_l(n)$ represent the lowest packing scores of the WPS structures from oligomeric numbers of m and n respectively. If $E_l(m) < E_l(n)$, the difference $E_l(n) - E_l(m)$ indicates the confidence we have in preferring m versus n as the oligomeric number. On applying this approach to determine the oligomeric number of phospholamban, the pentamer has the lowest packing score causing us to correctly conclude that the pentamer is the most feasible oligomeric number.

In summary, we have developed a novel approach that performs a complete, data-driven search to identify all structures of a homo-oligomeric complex that are consistent with NOE restraints and display high-quality vdW packing. Our tests on phospholamban and four other proteins demonstrate the power of our method in determining and evaluating homo-oligomeric complex structures. Our approach is particularly important in sparse-data cases, where relying on an incomplete, biased search

may result in missing well-packed, satisfying conformations. Examination of the entire solution space further enables objective evaluation of the amount of structural uncertainty. Finally, we show that it is possible to determine the oligomeric number directly from NMR data. The details of our methods and results are available in our paper [9].

Acknowledgments

We would like to acknowledge members of the CBK lab and Ivelin Georgiev from the BRD lab for helpful discussions. This work was supported in part by the following grants, to BRD: National Institutes of Health (R01 GM 65982) and National Science Foundation (EIA-9802068 and EIA-0305444); CBK: National Science Foundation (IIS-0444544 and IIS-0502801); JJC: Smith Family Award for Young Investigators. JJC is a Pew scholar.

References

1. A. T. Brunger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren. Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Cryst.*, D54:905–921, 1998.
2. S. R. Comeau and C. J. Camacho. Predicting oligomeric assemblies: N-mers a primer. *Journal of Structural Biology*, 150(3):233–44, 2005.
3. D. Duhovny, R. Nussinov, and H. J. Wolfson. Efficient unbound docking of rigid molecules. In Gusfield et al., editor, *Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI)*, Lecture Notes in Computer Science 2452, pages 185–200. Springer Verlag, 2002.
4. L. G. Dunfield, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides. 8. empirical potential energy algorithm for the conformational analysis of large molecules. *J. Phys. Chem.*, 82:2609–2616, 1978.
5. R. J. Kovacs, M. T. Nelson, H. K. Simmerman, and L. R. Jones. Phospholamban forms Ca²⁺-selective channels in lipid bilayers. *J. Biol. Chem.*, 263:18364–18368, 1988.
6. M. Nilges. A calculation strategy for the structure determination of symmetric dimers by ¹H NMR. *Proteins*, 17(3):297–309, 1993.
7. K. Oxenoid and J. J. Chou. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *PNAS*, 102:10870–10875, 2005.
8. B. Pierce and Z. Weng. M-ZDOCK: A grid-based approach for C_n symmetric multimer docking. *Bioinformatics*, 21(8):1472–1476, 2005.
9. S. Potluri, A. K. Yan, J. J. Chou, B. R. Donald, and C. Bailey-Kellogg. Structure determination of symmetric protein complexes by a complete search of symmetry configuration space using NMR distance restraints. *Proteins*, in press, 2006.
10. Goodsell D. S. and Olson A. J. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.*, 29:105–53, 2000.
11. C. E. Wang, T. L. Pérez, and B. Tidor. AMBIPACK: A systematic algorithm for packing of macromolecular structures with ambiguous distance constraints. *Proteins*, 32:26–42, 1998.