3-1-2007

Complete Configuration Space Analysis for Structure Determination of Symmetric Homo-oligomers by NMR

Shobha Potluri
Dartmouth College

Complete Configuration Space Analysis for Structure Determination of Symmetric Homo-oligomers by NMR

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in
Computer Science
by
Shobha Potluri
DARTMOUTH COLLEGE

Hanover, New Hampshire

March, 2007

Examining Committee:

(chair) Chris Bailey-Kellogg

Bruce R. Donald

Alan M. Friedman

Devin Balkcom

Charles K. Barlowe, PhD.
Dean of Graduate Studies

Abstract

Symmetric homo-oligomers (protein complexes with similar subunits arranged symmetrically) play pivotal roles in complex biological processes such as ion transport and cellular regulation. Structure determination of these complexes is necessary in order to gain valuable insights into their mechanisms. Nuclear Magnetic Resonance (NMR) spectroscopy is an experimental technique used for structural studies of such complexes. The data available for structure determination of symmetric homo-oligomers by NMR is often sparse and ambiguous in nature, raising concerns about existing heuristic approaches for structure determination. We have developed an approach that is *complete* in that it identifies all consistent conformations, *data-driven* in that it separately evaluates the consistency of structures to data and biophysical constraints and *efficient* in that it avoids explicit consideration of each of the possible structures separately. By being complete, we ensure that native conformations are not missed. By being data-driven, we are able to separately quantify the information content in the data alone versus data and biophysical modeling. We take a configuration space (degree-of-freedom) approach that provides a compact representation of the conformation space and enables us to efficiently explore the space of possible conformations. This thesis demonstrates that the configuration space-based method is robust to sparsity and ambiguity in the data and enables complete, data-driven and efficient structure determination of symmetric homo-oligomers.

Acknowledgements

My journey towards a doctorate degree has never been alone. Many people—mentors, family members, friends and acquaintances, have each contributed in their own special way towards this achievement.

My advisor and mentor Prof. Chris Bailey-Kellogg has made this difficult journey fun-filled and exciting. His enthusiasm and dedication for good high-quality research has driven me, many a time, to pursue problems that I thought were impossible to solve. Meetings and discussions with him have always been stimulating. The thought of pursuing a PhD would never have crossed my mind had it not been for him. Chris, I thank you for all your support and encouragement, both at a technical level and personal level. You have made a dream never dreamt, or even thought of as remotely possible, a reality.

I would also like to thank Prof. Bruce Donald for being a wonderful mentor and introducing me to my thesis topic. His prompt feedback and suggestions have driven this research to new heights. I have enjoyed all my research discussions with him and am grateful for his encouragement. Thanks to Prof. Alan Friedman who was always ready to give the biological perspective and point out things that, as a computer scientist, I would overlook. The weekly meetings with folks at Purdue have always been fun and are memorable

in more ways than one. I would also like to thank Prof. Devin Balkcom for being on my thesis committee and for providing valuable comments and suggestions on my research.

I am grateful and thankful to my husband, Narendra Chaparala, for his unwavering support through the last couple of years. His patience and willingness to always be there has helped me to get through several difficult times. His insistence on proper rest and relaxation has helped me retain my sanity. I am grateful to my mom, dad, pinni and babai for their encouragement over the years. My sister, Shilpa, has always been there when I needed a quick relaxation and a break from the regular routine. Her patience towards the "words of wisdom" (a side-effect of pursuing a PhD) I am ever ready to impart is always appreciated. I would also like to thank my in-laws for their confidence and the pride they take in my doctoral degree. I am also thankful to my cousins, Vicky, Deepthi and Krishna.

Thanks to the wonderful friends I have been fortunate enough to have. Tony has been a great collaborator and a good friend. His passion and excitement for research is very contagious and it was tremendous fun working with him. I have learnt a lot from him, both research-wise and otherwise. I would also like to thank the members of the CBK lab: Xiaoduan, John, Fei and Wei. The research (and several non-research) discussions with you guys have always been helpful and refreshing. Xiaoduan, you have been a great lab-mate all over the years, being patient towards all my talking and the never-ending phone calls.

Lacy George, Satarupa Basu and Virginia Anderson, you guys have been great room-mates and wonderful friends all through my stay at Dartmouth. I really appreciate your concern for the rest and nutrition that I always seemed to be in need of. You guys have al-

ways been there to talk to and drag me around to do something fun. Satarupa, your strength and attitude towards life is always inspiring. Virginia, you have been by my side all through my stay at Dartmouth. Thank you for being a wonderful friend. I would also like to thank Santhoji Katare, a friend and a mentor, who seems to be ever ready with a solution to my never-ending problems. I am eternally grateful to my friends at Purdue: Lakshmi Josyula (fondly known as Ani), Prasanna Karmakar, Bhagyalakshmi Bethala, Vidya Ganesh, Simran Banga, Brahmananada Reddy Vanga who have been with me, helping me get through the most difficult time of my life. I would also like to thank several other friends: Krishna Nandivada, Shrish Ranjan, Raavi Kalyani, Sirisha Ganta, Saroja Gadde, Sudha Gadicherla, Madhu, Hemanth Potluri, Aly Azeem, Tanmay Lele, Deepa Mukhtyar, Priyan Patkar, and Aparana Bhaskaran, who have each contributed in their own way. Last but not the least, I would like to thank Ratna aunty, for being a source of strength and courage through several difficult times.

I would also like to acknowledge Ryan Lilien and Ivelin Georgiev for providing me with support on using the software they had developed.

Contents

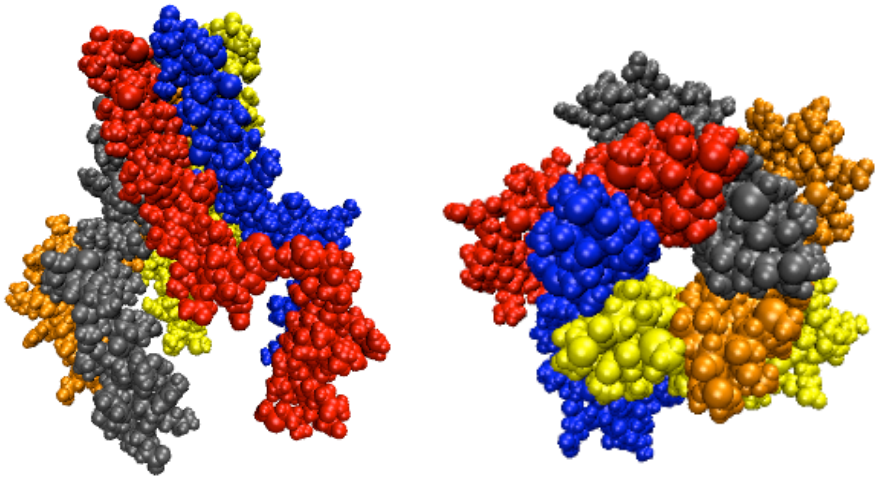
List of Tables

List of Figures

1. INTRODUCTION

Symmetric homo-oligomers are a class of protein complexes that have similar subunits arranged symmetrically. These complexes play pivotal roles in complex biological processes including ion transport and regulation, signal transduction, and transcriptional regulation. Structure determination of symmetric homo-oligomers is necessary in order to gain valuable insights into their mechanisms. For example, phospholamban is a symmetric homo-oligomer that is known to regulate the calcium levels between cytoplasm and sarcoplasmic recticulum and hence aids in cardiac muscle contraction and relaxation. Structure determination of phospholamban (Figure 1.1 [119]) enables us to understand how it regulates the intracellular calcium levels and how it could function as an ion channel. Malfunctioning of phospholamban could lead to congestive heart failure and cardiomyopathy. Knowledge of the structure not only enables understanding function but also aids in design of drugs to target such diseases.

The increasing availability of experimentally determined protein sequences and individual protein structures now allows research to focus on protein interactions and structure determination of protein complexes. Through a series of post-translational modifications and gene splicing mechanisms, the $\sim 30,000$ genes of the human genome are speculated to



(a) (b)

Fig. 1.1: Structure of phospholamban in two different views with different subunits shown in different colors.

give rise to 10^6 proteins [122]. The majority of these proteins interact with other proteins in processes that impact cellular structure and function driving the need for sophisticated approaches to understand these interactions.

The data available for determining structure of symmetric homo-oligomers from experiments such as Nuclear Magnetic Resonance (NMR) spectroscopy is often sparse and ambiguous in nature. This makes it necessary to develop algorithms which deal with the sparsity and ambiguity in a robust manner and which guarantee that native conformations are not missed. It is important to identify all structures consistent with the data and avoid false precision. Furthermore, there is need for approaches that accurately identify the structural constraint that the data provides and quantify the information content in the data.

1.1 Necessary Features

The features necessary in a structure determination method for protein complexes given data from experiments and structures of interacting proteins are as follows:

Complete: Despite the structural knowledge of the interacting proteins of a protein complex, the space of possible conformations of the complex is immense. To avoid missing native conformations a structure determination method must be *complete* in exploring the conformation space and evaluating each of the possible conformations. This becomes especially important when the experimental data available is sparse. Sparseness in the data could cause a structure determination method which is incomplete to be trapped in local minima. Requiring a structure determination method to be complete avoids any bias in the search and ensures that native conformations are not missed.

Efficient: The vast number of possible conformations makes explicit consideration of each of them computationally infeasible. This calls for approaches that are *efficient* in that they avoid this explicit enumeration and allow us to determine the complete set of structures.

Data-driven: Experimental data by itself is typically insufficient to determine a structure and needs to be complemented with information from biophysical constraints. Structures that satisfy both the data and and the biophysical constraints must be identified; this is typically done by subjecting the structures to energy minimization, a process we call "modeling". The lower the energy, the greater the satisfaction of the biophysical constraints. Such modeling involves choosing model parameters, and the final structure could

be biased by these choices. We need approaches that are *data-driven* in that conformations are first tested for consistency with data and only the consistent conformations are later evaluated for modeling. Being data-driven avoids over-reliance on subjective choices of parameters for modeling, and consequent false precision in determined structures. It also allows us to independently identify the amount of the structural constraint provided by data alone versus data and modeling. Independently identifying the structural constraint enables us to quantify the *information content* in the data which helps find redundant data as well as inconsistencies in the data which warrant further investigation.

Robust: The difficulty in dealing with complexes makes the experimental data available for structure determination of complexes sparse and ambiguous. Further, as with any experimental technique, the data is subject to experimental errors. Structure determination methods must be *robust* in that limited amount of inaccuracies and deficiencies in the data do not affect the correctness of the method.

1.2 Our Approach

1.2.1 Configuration Space

The focus of the current thesis is on structure determination of one class of protein complexes, homo-oligomers with cyclic-fold (C_n) symmetry, using data from NMR experiments. We have developed a structure determination method for symmetric homo-oligomers that has the necessary features of completeness, efficiency, data-drivenness and robustness.

We achieve this by developing a configuration space-based approach. The configuration space is formed from a minimal number of parameters that represent a complex, that is, its degrees of freedom. For example, a complex with two subunits of known structure has six degrees of freedom—three for translation and three for rotation of one subunit with respect to the other. Every possible conformation of the complex can be represented by a point in an n-dimensional space, where n is the number of degrees of freedom of the complex. Determining the complex structure then becomes a search in this n dimensional space.

The thesis statement is as follows:

> *Configuration space analysis enables complete, data-driven and robust structure determination of symmetric homo-oligomers.*

1.2.2 Architecture

Our method for structure determination of symmetric homo-oligomers takes as input a set of inter-subunit Nuclear Overhauser Effect (NOE) restraints, the subunit structure, and the oligomeric number of the complex. Here, inter-subunit NOE restraints refers to the data obtained from NMR experiments, subunit structure refers to the structure of one unit in the complex and oligomeric number refers to the number of subunits forming the complex. Each inter-subunit NOE restraint constrains the distance between atoms on the adjacent subunits. Throughout this thesis, we will assume that the backbone symmetry is exact and that the backbone of the subunit is known. The subunit structure could be available either from determining the "bound" structure of a monomer within the complex or from deter-

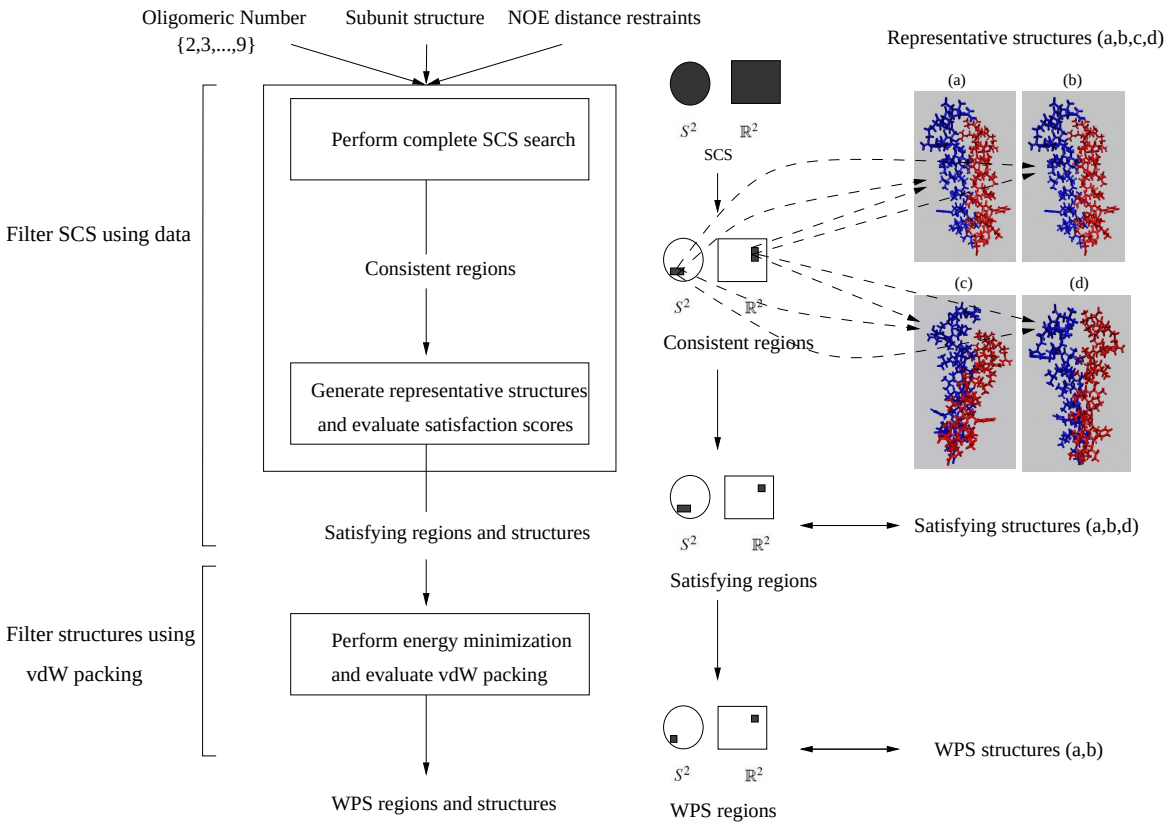


Fig. 1.2: The overall architecture for determining structures of symmetric homo-oligomers. The input to the protocol includes the subunit structure, the NOE restraints and the oligomeric number. The first phase of our two-phase approach involves a complete data-driven search using a branch-and-bound algorithm in the symmetry configuration space (SCS), the space of symmetry axis parameters. The 4-dimensional SCS is represented here as two 2D regions, a sphere representing the orientation space S^2 and a rectangle representing the translation space $\mathbb{R}^2$. The output from the branch-and-bound algorithm is a set of consistent regions in SCS that represent all structures satisfying the restraints. Representative structures are chosen from the consistent regions. These structures are evaluated for quality of restraint satisfaction, and those that are good enough are identified as satisfying structures. Regions corresponding to the satisfying structures in the consistent regions form the satisfying regions (represented by the dark regions). In the second phase, each of the satisfying structures is energy minimized and evaluated for quality of vdW packing. The output from this step is a set of well-packed satisfying (WPS) structures and WPS regions that represent conformations that are consistent with data and have high-quality vdW packing.

mining the structure of the monomer in its apo form. Experimentally it is possible to determine the subunit structure in complex prior to computing the oligomeric assembly [119]. The subunit structure in these cases is determined by using intra-molecular distance restraints from NOEs, backbone orientation restraints from residual dipolar couplings, and side-chain χ_1/χ_2 restraints from three-bond scalar couplings. In this dissertation, the modeling term we consider in evaluating structures is inter-subunit van der Waals (vdW) energy. Note that this can be replaced with other modeling terms such as electrostatics by considering the appropriate solvent models.

The key to our approach is the observation that, given the structure of a subunit, the structure of a C_n symmetric homo-oligomer is completely determined by identifying the position ($\mathbf{t} \in \mathbb{R}^2$) and orientation ($\mathbf{a} \in S^2$) of the symmetry axis. (We discuss other possible representations in Chapter 7.) Geometrically, the axis of symmetry is a line parallel to the unit vector $\mathbf{a}$ that intersects the xy-plane at the point $\mathbf{t}$. For a given axis, rotating the subunit n times by the angle of symmetry ($360^\circ/n$) around the symmetry axis yields the structure. Hence there are four degrees of freedom for a C_n symmetric homo-oligomer given the structure of the subunit (under the assumptions described in the previous paragraph), making its configuration space $S^2 \times \mathbb{R}^2$. We refer to this configuration space as the ***symmetry configuration space (SCS)***. Each possible conformation of the symmetric homo-oligomer is represented by a point in the 4-dimensional SCS. Hence, a complete search in the SCS allows us to identify all conformations consistent with the data.

Figure 1.2 shows the overall architecture. The approach consists of two phases. In the first phase, we perform a configuration space analysis by a complete data-driven search

in the 4D space of symmetry axis parameters, pruning out regions representing conformations that are inconsistent with the data. SCS is too large to search naïvely or exhaustively. Therefore, we have developed a novel branch-and-bound algorithm that is efficient and provably conservative in that it examines and conservatively eliminates non-satisfying regions. Without this algorithm, a complete, data-driven search would not be computationally feasible. At the end of the configuration space analysis, we return regions in SCS, the *consistent regions*, which contain all conformations that are consistent with the data. We choose *representative structures* from the consistent regions such that every conformation in the consistent regions is within an RMSD of τ_0 Å to at least one representative structure (τ_0 is a user-specified parameter). Due to the conservative bounds used in our search, the representative structures might contain conformations that are inconsistent with the data. The set of *satisfying structures* includes only those representative structures with restraint satisfaction scores below a chosen threshold. In the second phase of the overall approach, each of the satisfying structures is energy-minimized and scored based on van der Waals packing. The set of *WPS structures* includes those energy-minimized satisfying structures with van der Waals packing scores below a chosen threshold.

The following summarizes the contributions that the thesis makes

1. *Chapter 3*: We developed algorithms for complete, data-driven SCS search given the data and the subunit structure. We call this the "core algorithm" and developed it assuming that there is no noise, uncertainty or ambiguity in the NOE restraints and the subunit structure.

2. *Chapter 4*: In NMR experiments, ambiguity could exist in the identity of the atoms that correspond to a distance restraint. *NOE assignment* refers to the process that resolves the ambiguity as to which pairs of protons generated the observed data, and thus should be restrained in structure determination. We extended the core algorithm to handle ambiguous NOE data and perform NOE assignment.

3. *Chapter 5*: We extended the core algorithm to handle side-chain uncertainty in the subunit structure. We allow for side-chain uncertainty during the search through the SCS rather than only at the stage of energy minimization.

4. *Chapter 6*: We extended the core algorithm to be robust to noise and uncertainty in the NOE data. We provided algorithms to enable structural inference of symmetric homo-oligomers, that is, to probabilistically reason about the degree of agreement that possible conformations have to data and biophysical constraints.

In each of these chapters, the overall architecture is as shown in Figure 1.2. The chapters differ mainly in the search through the SCS. Chapters 4, 5 and 6 provide extensions of the core algorithm, enabling us to handle ambiguity in data, uncertainty in subunit structure and uncertainty in data respectively. Each of these chapters extends the core algorithm independent of the other extensions, demonstrating the applicability of our approach to handle noise, ambiguity and uncertainty in the input.

1.2.3 Validation of Our Approach

We validate our approach by testing it on a number of different symmetric homo-oligomers. The structures of each of these complexes have been determined either by X-ray crystallography or NMR. In structures determined by crystallography, the subunit structure was chosen as the structure of the first unit of the complex and the NOE data was simulated. Protons were added to the reference structure using the software CNS [23]. We simulated the NOE restraints by finding pairwise distances between protons on adjacent monomers. Every pair that had a distance less than 5 Å and almost exact symmetry (the mean of difference in distances across adjacent subunits is within 0.5 Å) was chosen as a restraint and an uncertainty of ± 1 Å was added. Choosing restraints with almost exact symmetry simulates the scenario of choosing inter-subunit restraints that have significant signal overlap. In structures determined by NMR, the subunit was chosen from the best representative conformers of the ensembles. Following is a brief description of each of the protein complexes we used to validate our approach. Table 1.1 summarizes the test cases.

In Chapter 3, we tested our approach on symmetric homo-oligomers with different oligomeric numbers. Phospholamban, is a symmetric homo-pentamer that satisfies the assumption that the subunit structure can be determined in complex. We chose as the reference structure the best representative conformer (as indicated in [119]) among the deposited 20 NMR structures (PDB id: 1ZLL). Nine inter-subunit NOE restraints were obtained [119] by using a mixture of labeled and unlabeled subunits and filtering NOE signals appropriately. The subunit structure was determined by a simulated annealing pro-

Tab. 1.1: Test Cases

Protein (symmetry)	PDB id	No. of restraints
Glycophorin A (2)	1AFO	6 (expt)
MinE (2)	1EV0	183 (expt)
Chicken cartilage matrix protein (CCMP) (3)	1AQ5	49 (expt)
Haemagglutinin (3)	1HTM	85 (simulated)
Kv1.2 potassium channel (4)	1QDV	32 (simulated)
Phospholamban (5)	1ZLL	9 (expt)
Gp31 co-chaperonin (7)	1G31	85 (simulated)

tocol using intra-molecular distance restraints from NOEs, backbone orientation restraints from residual dipolar couplings, and side-chain χ_1/χ_2 restraints from three-bond scalar couplings [119].

Human glycophorin A (PDB id: 1AFO) is a symmetric dimer. Six experimental NOE restraints were available. The authors reported twenty NMR structures and we chose as the subunit structure the first chain from the structure that best satisfies the restraints. Additional test cases in Chapter 3 include haemagglutinin (PDB id: 1HTM), a trimer; Kv1.2 potassium channel (PDB id: 1QDV), a tetramer; Gp31 co-chaperonin (PDB id: 1G31), a heptamer. These structures were determined by x-ray crystallography. Our simulations yielded 85 simulated restraints for haemagglutinin, 32 for the Kv1.2 potassium channel, and 85 for the Gp31 co-chaperonin.

We used two test cases to validate our ambiguity resolution and NOE assignment algorithms in Chapter 4. We could not use phospholamban or glycophorin A here, since we

could not simulate ambiguity for these complexes due to unavailability of required chemical shift data. We chose as test cases homo-dimeric MinE and homo-trimeric CCMP, for which inter-subunit NOE distance restraints and assigned chemical shift data were available in the BioMagResBank (BMRB) [131]. In both cases, isotopic labeling strategies were used to separate NOEs between atoms within a subunit from NOEs between subunits. We use the "bound" subunit structures and focus on determining the complex from the remaining ambiguous inter-subunit NOEs. We used as the subunit the first chain of the reference structures (stated by the authors to be the best representative conformers). The homo-dimeric topological specificity domain of *Escherichia coli* MinE [89] has a novel dimeric $\alpha\beta$-sandwich fold that has previously been determined (PDB id: 1EV0) by using both DYANA [64] and the NCS-symmetry potential of X-PLOR (with the ARIA method) [23]. Each subunit has 50 residues, and a total of 183 inter-subunit NOE restraints were deposited. The structure of the trimeric coiled-coil domain of chicken cartilage matrix protein (CCMP) [148] was originally determined using X-PLOR [23] (PDB id: 1AQ5). Each subunit has 47 residues, and a total of 49 inter-subunit NOE restraints were deposited.

We used the same two test cases, MinE and CCMP in Chapters 5 and 6.

The rest of the document is organized as follows. Chapter 2 reviews related work in structure determination of proteins, NOE assignment, side-chain uncertainty and structural inference. Chapter 3 describes our core algorithm in detail. Chapter 4 extends the core algorithm to handle ambiguity and perform NOE assignment. Chapter 5 discusses our approach to handle uncertainty in subunit structure. Chapter 6 presents our approach for structural inference and Chapter 7 presents conclusions and future work.

2. RELATED WORK

2.1 Protein Complex Structure Determination

2.1.1 Experimental Techniques

Traditionally, an atomic detail structure of a protein complex is determined using one of the two experimental techniques: X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. In X-ray crystallography [20], proteins are crystallized by slowly changing the conditions in a super-saturated protein solution. Electron density maps are then obtained from the diffraction patterns formed by beaming X-rays at the crystal. The structure of the protein complex is obtained from the diffraction patterns taken at several angles of rotation of the crystal. The main problems in structure determination by X-ray crystallography lie in production of sufficient quantities of the sample and in crystallization. Formation of crystals is a difficult and time-consuming process and is dependent on several parameters such as pH of the solution, purity of the protein, temperature, protein concentration, nature of the solvent and precipitant, and presence of added ions or ligands in the protein. Also, crystallization of membrane proteins is extremely difficult [25, 118].

Nuclear Magnetic Resonance (NMR) spectroscopy [20] is the other experimental tech-

nique that is widely used for protein complex structure determination. When placed in a strong magnetic field the hydrogen atoms in the protein complex spin in an equilibrium alignment with the field. With the application of radio frequency (RF) pulses, the nuclear spins become excited. When reverting to the equilibrium state, each atom emits a different frequency based on the electronic environment around its nucleus. By varying the RF pulses, multi-dimensional NMR spectra give information about angles and through-space and through-bond interactions between nuclei. These restraints help in identifying the 3D structure of the protein. The main problem with NMR is that it is restricted by the protein size. However, recent technical advances have allowed its application to systems as large as the 900 kDa GroEL-GroES complex [50].

Electron microscopy (EM) [130] is another experimental technique that provides low-resolution structural information for complexes. There are several variants of electron microscopy (EM), including single-particle EM [50], electron tomography [15] and electron crystallography of regular two-dimensional arrays of a sample [114]. In single-particle cryo-EM, the three-dimensional structure is reconstructed from two-dimensional projections taken from different angles. Small quantities of the sample are sufficient for cryo-EM. Cryo-EM can determine the electron density for complexes weighing greater than 200-500 kDa at low resolutions of approximately 5 Å [54, 72, 79, 155]. Although one cannot get atomic-level detail from cryo-EM density maps, the maps provide insights into the mechanism of large complexes. Electron tomography allows for the study of structures at a resolution of a few nanometers using multiple tiled views of the same object. It enables visualization of complexes in a cellular context [14, 61].

Several minimalist experiments, though not comprehensive enough to determine complex structures by themselves, provide low-resolution information about the relative positions of units in a complex. Proximity information between adjacent units in a complex can be obtained from experiments such as mutagenesis [2, 3, 19, 35, 43, 93, 110], hydrogen-deuterium exchange [9, 94], chemical cross-linking [12, 47, 68, 92, 135, 132, 137], Fluorescence Energy Transfer (FRET) [138, 152], Fourier transform infra-red spectroscopy (FTIR) [84]. Small angle X-ray scattering (SAXS) is another method that can provide low-resolution information about the shape of a complex [105, 134].

2.1.2 Biophysical Modeling

Any structure determination method must ensure satisfaction of certain biophysical constraints. The energy of the structure gives a measure of how well the biophysical constraints are satisfied and is typically obtained as [85],

$$\begin{aligned} E &= E_{\text{covalent}} + E_{\text{noncovalent}} \\ E_{\text{covalent}} &= E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}} \\ E_{\text{noncovalent}} &= E_{\text{electrostatic}} + E_{\text{vanderWaals}} \end{aligned} \tag{2.1}$$

There can be several implementations for each of these terms. Typically, the bond and angle terms are modeled as harmonic potentials centered around the "ideal" bond length derived either from experiment or theory. The dihedral or torsional terms typically have multiple minima and their actual form varies with the implementation. The class of covalent terms could also include an "improper" dihedral term to account for the planarity of

aromatic rings.

The non-covalent or non-bonded terms are much more computationally costly because they include many additional pairwise interactions per atom. The van der Waals term is typically modeled using a "6-12 Lennard-Jones potential", where the attractive forces fall off with distance as r^{-6} and repulsive forces as r^{-12}, where r represents the distance between two atoms. Generally a cutoff radius is used so that atom pairs whose distances are greater than the cutoff have a van der Waals interaction energy of zero. For the electrostatic terms, the basic functional form is the Coulomb potential, which depends on the effective dielectric function for the medium and falls off as r^{-1}.

The energy terms include parameters for force constants, van der Waals multipliers, and other constant terms. These parameters, together with the equilibrium bond, angle, and dihedral values, partial charge values, atomic masses and radii, and definitions of the energy terms, are collectively known as a *force field* [102]. The parameters are typically obtained from agreement with experimental values or theoretical calculations results. Two widely used force fields are AMBER [26, 33] and CHARMM [21].

2.1.3 Structure Determination by NMR

The standard protocol for structure determination by NMR [150], given spectra, involves the successive steps of peak picking, chemical-shift assignment, NOE assignment and collection of other conformational constraints, and finally structure calculation and refinement. *Peak picking* refers to the identification of peaks in each collected NMR spectrum. Each

peak corresponds to an interaction between a pair of atoms. The challenge of peak picking is that peaks often overlap. Stronger spectrometer magnets and multidimensional NMR experiments could make peak-picking easier, but are expensive. The variation in the resonance frequency of a nucleus due to its chemical environment is called the *chemical shift*. The coordinates of a peak in an NMR spectrum correspond to the chemical shifts of the interacting nuclei that generated that peak. *Chemical shift assignment* is the process of assigning observed chemical shift values to specific atoms. NOESY spectra contain peaks corresponding to magnetic interactions between nuclei through space. *NOE assignment* is the process of identifying the atoms involved in generating NOE peaks in a NOESY spectrum. Generally, the backbone and side-chain chemical shifts are assigned from a variety of 3D NMR experiments. This information is then used to assign NOEs and derive distance *restraints* for structure calculation. Each distance restraint exists between pairs of atoms and constrains the distance between the corresponding pairs of atoms. Also, torsion angle restraints and residual dipolar couplings are obtained from NMR experiments. The structure calculation and refinement process identifies structures that satisfy all the restraints and are also consistent with physical modeling terms.

Given data from NMR experiments in the form of restraints between atoms in the complex, structure determination is typically done by simulated annealing and molecular dynamics protocols. The protocols could either be performed in the cartesian space [22, 23, 24] or in torsion angle space [63, 64]. The energy function that is minimized includes both physical energy terms such as those discussed in the previous section and pseudo-energy terms to check for consistency with experimental data. The advantage of torsion angle

dynamics over cartesian space dynamics is that the number of degrees of freedom is decreased since the covalent structure parameters (such as bond lengths and bond angles) are kept fixed during the minimization process. The rugged landscape of the energy function could cause the simulated annealing / molecular dynamics mechanisms to get trapped in local minima. Furthermore, the precision in the determined structure is strongly affected by the "temperature" used in the simulated annealing protocols.

In the context of symmetric homo-oligomers, the simulated annealing / molecular dynamics protocols have been extended for determining structure by incorporating additional constraints to ensure symmetry, e.g., by penalizing differences between the subunits [126] or by fixing multiple copies of the backbone and allowing only the side chains to move to fit the restraints [119]. DYANA [64] and CYANA [71, 62] have been extended for homo-dimer structure determination, while the symmetry-ADR method [112, 116, 117] is applicable to higher-order oligomers. When the data available is sparse and ambiguous, as is typically the case for symmetric homo-oligomers, the problems of local minima are further exacerbated.

2.1.4 Computational Techniques

Computational techniques [130] for predicting the structures of complexed proteins mainly fall into two categories: homology modeling and docking. In homology modeling for complexes [5, 7, 106], a structural model for a complex is constructed based upon a template protein complex, another protein complex with good sequence similarity and whose struc-

ture is known. The question that arises is how similar the sequences must be. It was found that proteins of the same fold interact in a similar way if the sequence identity is above approximately 30% [6]. The applicability of homology modeling is limited by the availability of a structure with high sequence similarity.

Docking-based approaches model the structures of the complex given the structures of the subunits involved in the complex. Docking strategies usually involve a two-stage approach: generate a set of possible docked structures, and then score them. The possible structures are typically generated by sampling the space of rotations and translations of the docked subunit with respect to the fixed subunit [28, 52, 59, 104, 48, 128, 140]. The fast Fourier transform (FFT)-based method developed by Katchalski-Katzir et al. [86] is used by most of the docking algorithms to efficiently sample the space of possible docked configurations. Alternatively, geometric hashing [115] samples conformations consistent with shape complementarity.

Scoring a docking prediction is usually based on some kind of a force field (Section 2.1.2), which assigns an energy to all atom or residue pairs after subjecting the predicted model to energy minimization [60]. Due to the cost involved in energy minimization, faster scoring functions [66, 75, 133] are also used to discriminate among predicted models which include elements such as geometric complementarity [40], contact and overlap checks [13, 55] and counts of hydrogen bonds [10]. When experimental structures for the subunits of a complex are not known, docking is applied on homology models of the individual subunit [136].

Docking-based methods are immediately applicable to symmetric homo-oligomers, by

docking two subunits of the homo-oligomer. Pierce et al. [121], Duhovny et al. [41], and Comeau et al. [32] have all used docking-based approaches followed by filtering based on symmetry and scoring by minimizing energy functions to predict structures of homo-oligomers. In each of the above-mentioned approaches, it is desirable to use use data from experiments to validate the structures predicted.

Several docking-based approaches have been developed that use experimental data. These approaches can be divided into two classes: techniques that use the data during the scoring stage and techniques that use the data during the search stage. Techniques that use the data during the scoring stage generate a set of possible candidate structures and then filter them based on the consistency with the data. Adams et al. generated a set of possible candidates by sampling and then scored them using mutagenesis data for structure determination of two transmembrane proteins, glycophorin A and phospholamban [2, 3]. Mutagenesis data [11, 53] was used to evaluate the structures obtained from the docking programs HEX [128] and GRAMM [140]. Similarly, mass spectrometry hydrogen/deuterium exchange data [9] was used with the program DOT [104]. Also, NMR chemical shift perturbation data and residual dipolar couplings [38, 97, 109] were used with the programs FTDOCK [52], AUTODOCK [111] and BIGGER [120]. In approaches that use the experimental data during the search stage, the most popular technique is to incorporate the data in an energy term which is low if the restraints that the data provides are satisfied [31, 34, 39, 51, 58, 107, 112, 129, 139]. TREEDOCK [45] uses anchor points based on experimental data. Another approach to incorporate the data during the search stage is by assigning residues that are restrained by the data more weight in fast Fourier

transform-based rigid body docking approaches (weighted geometric docking) [16].

Wang et al. [144] developed another approach for structure determination of protein complexes using restraints from NMR experiments. They propose a branch-and-bound algorithm [4], AMBIPACK, to determine structures of protein complexes using potentially ambiguous inter-subunit distance restraints obtained from NMR experiments. AMBIPACK assumes that the subunit structures are rigid. It finds a set of transformations of the rigid substructures that satisfy the restraints. It uses a hierarchical subdivision of the space of possible conformations and a branch-and-bound algorithm to eliminate infeasible regions of the space of possible transformations. Local search methods then focus on the remaining space.

None of the existing approaches for structure determination come with guarantees of completeness, and they all fail to take advantage of the 'closed-ring' constraint of a symmetric homo-oligomer. The closed-ring constraint arises from the fact that in C_n symmetric homo-oligomers, the subunits are arranged in a ring. This implies that a restraint existing between subunits 1 and 2 also exists between subunit n and 1, where n is the number of subunits. As we will see, our approach uses the oligomeric number to enforce an *a priori* symmetry constraint. In this sense it is analogous to the manner in which non-crystallographic symmetry is handled in molecular replacement for x-ray crystallography [98]. Our approach directly exploits the kinematics of the 'closed-ring' constraint, and thereby derive an *analytical* bound for pruning inconsistent conformations, which is tighter and more accurate than previous randomized numerical techniques [144].

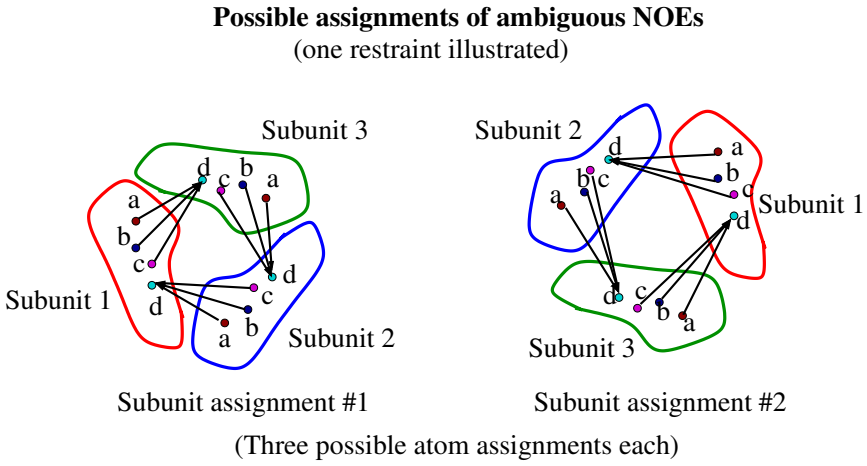


Fig. 2.1: Different possible assignments (arising from subunit and atom ambiguity) for one ambiguous NOE are illustrated for a homo-trimer. The NOE could be ordered between subunits as 1-2-3 or 1-3-2, and has chemical shift degeneracy between atoms a, b, and c.

2.2 NOE Assignment

Automated NOE assignment is a key bottleneck in structure determination by NMR. Because of the limited accuracy of chemical shift values, many NOE peaks cannot be attributed to a single pair of atoms. In symmetric homo-oligomers, this problem is further complicated by the additional ambiguity that arises in the identity of the subunit and whether the peak is between atoms within a subunit or between subunits.

The sources of ambiguity in inter-subunit NOE assignment for symmetric homo-oligomers are (Figure 2.1): *intra- vs. inter-subunit ambiguity*, whether the restrained atoms are within the same subunit in the complex or in different ones; *subunit ambiguity*, the subunits to which the restrained atoms of an inter-subunit NOE belong; *atom ambiguity*, the identities

of the restrained atoms among those with chemical shifts similar to the NOE peak. Intra- vs. inter-subunit ambiguity can be resolved experimentally, by a combined analysis of 3D ^{15}N-edited, double ^{13}C-filtered NOESY of a mixed sample and standard 3D ^{13}C-edited NOESY of a homogeneous ^{15}N, ^{13}C-labeled sample [156, 142]. Atom ambiguity can (in principle) be resolved experimentally by using 4D or higher dimensional NOESY spectra [87, 49]. However, in general, subunit ambiguity, and atom ambiguity for 3D NOESY spectra, require computational solutions.

Subunit ambiguity is a fundamental problem in dealing with homo-oligomers. While in some special cases it might be possible to manually resolve the ambiguity by identifying a self-consistent direction (e.g., in coiled coils [148]), in general an NOE could involve a subunit i and any one of subunits $i+1$, $i+2$, ... We call a determination of the relative positions of the subunits the *subunit assignment* of the NOE. Most previous attempts at resolving subunit ambiguity have done so simultaneously with resolving atom ambiguity, and have employed heuristic techniques (see below). As we discuss below, these approaches risk being trapped in local minima and missing correct assignments.

One previous approach that rigorously deals with subunit ambiguity is AMBIPACK [144] (see Section 2.1.4). AMBIPACK explicitly enumerates combinations of subunit assignments, but only for a limited number of the least ambiguous NOEs. Furthermore, AMBIPACK handles only (counter)clockwise subunit ambiguity. By (counter)clockwise subunit ambiguity we mean that an NOE could either go clockwise or counterclockwise in a ring of monomers, resulting in 2 possible assignments. Arbitrary subunit ambiguity means that an NOE could go to any of the other $n-1$ monomers, resulting in $n-1$ possible assignments.

Also, the presented AMBIPACK algorithm and results do not consider atom-ambiguous NOEs.

Atom ambiguity in NOEs is caused by chemical shift degeneracy. We call a determination of the atoms the *atom assignment* of the NOE. A number of techniques, such as ARIA [113], NOAH/DIAMOND [151], AUTOSTRUCTURE [73], and PASD [83], have been developed to address atom ambiguity in the context of monomer structure determination. These techniques follow an iterative assignment strategy: an initial set of unambiguous NOEs is used to generate some number of structures, which are then used to evaluate the consistency of atom assignments of the remaining ambiguous NOEs. The NOEs that are inconsistent with the ensemble are pruned. The remaining NOEs are incorporated in subsequent rounds of structure ensemble determination, and the process is repeated until no further improvements in atom assignments or structures can be obtained. An alternative approach to assign NOE restraints in monomer structure determination using a rotamer ensemble-based algorithm and residual dipolar couplings was developed by Wang and Donald [145].

Previous approaches for dealing with both subunit and atom ambiguity in symmetric homo-oligomers perform a tight loop of NOE assignment and structure determination as described in the previous paragraph for monomers. Atom ambiguity is handled using approaches described in the previous paragraph and subunit ambiguity is handled by incorporating additional constraints to ensure symmetry [112].

The iterative techniques discussed above all follow a best-first strategy without backtracking, and thus must contend with the problem of multiple local minima. The large

rearrangements potentially required to escape local minima are difficult to identify by local search techniques. The order in which assignments are chosen may affect the outcome, because the determined ensemble attempts to satisfy the chosen assignments, and the ensemble is then used to evaluate subsequent assignments. For example CYANA uses the heuristic of choosing the most consistent NOEs at each iteration, but of course this does not guarantee correctness. To ensure settling into the proper minimum, these techniques typically require a significant amount of over-restraint (e.g., ARIA requires > 5 NOEs per residue). However, with higher-order oligomers, available inter-subunit data can be quite sparse, due to the size of the proteins and the decrease in sensitivity of peaks, along with the decrease in peak intensity due to double filtering and editing strategies. Finally, as the number of ambiguous NOEs increases, the ruggedness of the potential landscape increases, making it much more difficult and time-consuming to find valid solutions. Consequently, the level of atom ambiguity that current techniques can handle is also limited (e.g., CYANA requires a ^{1}H chemical shift match tolerance ≤ 0.03 ppm). Furthermore, the presence of subunit ambiguity in higher-order oligomers makes all NOEs inherently ambiguous and exacerbates the landscape ruggedness.

2.3 Side-chain Uncertainty

In structure determination of protein complexes given the structures of the subunits, the uncertainty in subunit structure is a key issue. The uncertainty could arise either from flexibility or from deficiencies in the experimental data. When subunits form a complex, they

could undergo conformational change. The conformational change could be either small-scale, fast motions in their side-chains, or large-scale, slow motions in their backbones. In this thesis we assume that the backbone undergoes no conformational change and no backbone uncertainty is considered. We consider only side-chain uncertainty.

Rigid docking is an area of research that determines structures of complexes assuming that the structures of the subunits in complex are exactly known and no conformational changes occur. Flexible docking [66] on the other hand deals with conformational changes in the subunits. There are several approaches that handle flexibility in the side-chains. One approach first performs rigid docking to get a set of possible candidate structures. Flexibility is introduced only at the later stage when each of the candidates is subjected to energy minimization [1, 32]. Jackson and co-workers [74] used a self-consistent mean field approach to iteratively refine the side-chain conformations generated by their rigid-body docking program, FTDOCK. In order to avoid discarding valid solutions during rigid docking, the rigid docking employs a "soft" potential, considering only a portion of each residue (such as just the backbone) [141, 153]. HADDOCK [39] introduces flexibility increasingly in three stages. In the first stage, the subunits are assumed to be rigid; in the second stage, side-chains are allowed to move; and in the third stage, both side-chains and backbones are made flexible. The classical approaches [112] of structure determination by NMR handle flexibility by iteratively refining the side-chain conformations based on energy functions.

Another approach to handle flexibility is to consider ensembles of the subunits and dock the ensemble rather than single conformers [30]. The ensembles could be obtained from experiments by collecting either crystal structures (apo or bound to different ligands) or

the set of structures obtained from NMR experiments [90]. They could also be obtained by subjecting existing conformations to molecular dynamics simulations [18, 29]. Each of these approaches fails to be complete in exploring the space of backbone conformations while handling side-chain flexibility.

One could imagine applying techniques developed for handling side-chain uncertainty in protein-ligand docking to docking protein complexes. In protein-ligand docking, only ligand flexibility and/or flexibility only at the binding site is considered and all possible conformations can be explicitly considered [57, 80, 91, 95, 108, 125, 146]. SLIDE [154] is an approach where the authors follow a minimal rotation hypothesis (side-chains move minimally from existing conformations). Such protein-ligand docking approaches would not work as well for protein-protein docking because of the combinatorial explosion in the number of possible conformations.

Another approach to handle side-chain flexibility is to restrict side-chain conformations to a limited set of possibilities. Certain side-chain conformations are higher in energy than others due to possible steric clashes [66]. This causes the distributions of side-chain angles to be non-uniform, with preference towards low-energy conformations. This property of side-chains being non-uniformly distributed is exploited by modeling side-chains with a discrete set of possible conformations called *rotamers*. Rotamers provide a reduced representation of the side-chain conformational space. In a rotamer library [100, 123, 82] each amino acid is represented by a set of common low energy conformations. A rotamer library could be backbone-independent, that is, not depending on the backbone conformation, or backbone-dependent, that is, depending on the backbone conformation. The libraries are

typically obtained by statistical analysis on high resolution protein structures in the protein databank [17] by clustering observed conformations, or by dividing dihedral angle space into bins, and determining an average conformation in each bin [81].

Rotamer libraries have been used in determination of side-chain conformations mainly in the context of single protein structure determination. In this case, the backbone is assumed to be known and rotamers are used in determining side-chains. Leach and Lemon [96] have developed one such approach that explores the side-chain conformational space given the backbone. Each residue has a set of possible conformations, arising from the different rotamers considered. A rotamer is pruned, that is removed from a residue's set, if the contribution to the total energy is reduced by using an alternative rotamer of that residue. The so-called dead-end elimination criterion developed by Desmet et al. [37] (in the context of protein design) was used to provably prune such rotamers. Among the set of remaining rotamers, an A*-search [69] is done to identify side-chain conformations that lead to low-energy structures. Althaus et al. [8] developed an approach where side-chain optimization (pruning rotamers) is done by either (a) fast heuristic approach or (b) an exact but slower integer linear programming based approach. None of the mentioned rotamer-based approaches have been applied in the context of protein complex structure determination.

One approach that has used rotamers in determining protein complexes is ROSETTA-DOCK [143]. The rotamer libraries are supplemented with side-chain conformations taken from the unbound subunit structures. This approach performs side-chain optimization by torsion-angle space minimization. The energy-minimization techniques used by this approach could fail to explore the space of all possible conformations and hence could miss

native conformations.

2.4 Structural Inference

The classical approaches for structure determination by NMR [22, 23, 64, 62] formulate structure determination as an optimization problem. The value that is optimized is the energy term that includes the physical energy of the molecule (Section 2.1.2) and pseudo-energy that represents agreement with the data. Formulating structure determination as an optimization problem implies that there exists a unique solution, that is a unique conformation satisfying the data and biophysical constraints. When data is of high quality, this assumption leads to valid solutions. But, the uncertainty and noise typically present in the data makes this assumption questionable. To overcome this, an ensemble of structures are calculated by repeating the optimization procedure several times. This, however, does not take into account the noise and uncertainty in the data and makes it difficult to evaluate the quality of the ensemble of structures. Furthermore, it is not obvious how to choose parameters such as the relative weight between the physical energy and the pseudo-energy.

The Inferential Structure Determination (ISD) approach developed by Wolfgang et al. [149] addresses these concerns by formulating NMR structure determination as an inference problem. Their approach deals with noise and uncertainty in data in a robust manner using probabilistic inference techniques, where each structure is assigned a probability measure based on its agreement with the data and prior biophysical information. This approach was developed in the context of protein structure determination by NMR. Using

Bayes theorem [77], the posterior probability in a structure is given as the product of the likelihood of the data and the prior support for the structure. A structure that is in better agreement with the data is given a higher probability. In this approach no assumptions about the existence of a unique solution are made. Although the framework is robust in handling uncertainty and noise in the data, the approach fails to come with the guarantee of being complete. Possible conformations are obtained by sampling the posterior distribution using a stochastic Markov Chain Monte Carlo [27] method. The sampling could miss conformations with high posterior density. The conclusions on precision of structures, though better than the classical approaches, still fail to be completely accurate due to the stochastic sampling.

3. CONFIGURATION SPACE ANALYSIS OF SYMMETRIC HOMO-OLIGOMERS

This chapter describes the core of our complete, data-driven algorithm for structure determination of symmetric homo-oligomers using distance restraints from NMR. Section 1.2.2 gave a high-level description of our algorithm and here we describe each of the steps in detail.

3.1 The Core Algorithm

3.1.1 Complete Search of SCS

Given the subunit structure, a set of inter-subunit distance restraints and the oligomeric number as the input, we must identify all possible symmetry axis parameters, $(\mathbf{a}, \mathbf{t}) \in (S^2 \times \mathbb{R}^2)$, such that corresponding structures satisfy all the restraints. Here, $\mathbf{a}$ represents the orientation of the axis and $\mathbf{t}$ represents the relative position between the axis and one of the subunits. An exact algebraic formulation for identifying all possible values of $(\mathbf{a}, \mathbf{t})$ is possible, but it would yield high-degree polynomials that are expensive to solve exactly. Hence, we develop here a branch-and-bound algorithm that searches the SCS

and conservatively eliminates regions that provably cannot satisfy all the restraints. The branch-and-bound approach facilitates a complete search over all possible conformations and ultimately returns consistent regions in SCS.

As Figure 3.1 shows, the branch-and-bound search follows a tree structure and performs a recursive search through regions in SCS. Each node in the tree is an SCS *cell*—a 4-dimensional hyper-cuboid defined by values representing extrema along each of the four dimensions. At each node of the branch-and-bound search, we test whether any point in the cell represents a consistent conformation. If such a point possibly exists, we *branch* and partition the cell into smaller sub-cells. We continue branching until we can either *eliminate* or *accept* each cell. We *eliminate* a cell by computing *bounds* on the conformations it represents and determining that they all violate at least one restraint or contain significant steric clashes. 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 of τ_0 Å of each other and each restraint is satisfied by at least one conformation represented by the cell.

Bounding

We evaluate a cell for potential steric clash only when the cell is "small" enough ($\leq 5^\circ$ in S^2 and ≤ 0.5 Å in $\mathbb{R}^2$). In testing for a steric clash, we separately consider each of the 16 structures represented by the corners of the 4D cell. A cell is pruned only when each of the 16 structures has severe steric clashes (see Section 3.2) between its atoms. This

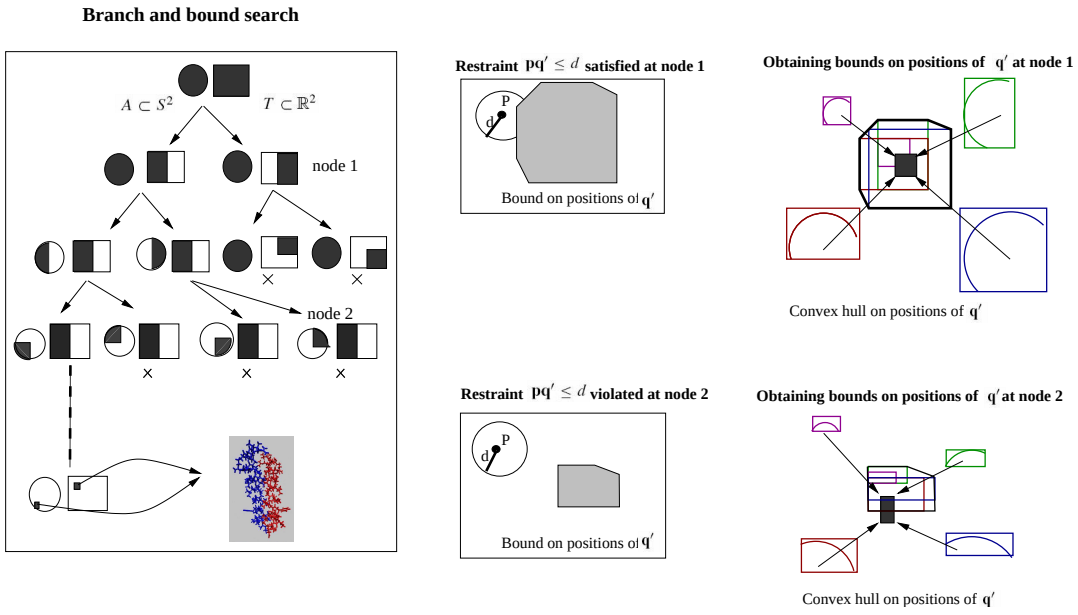


Fig. 3.1: The branch-and-bound algorithm proceeds as a tree search in SCS, the space of orientations and translations of the symmetry axes ($S^2 \times \mathbb{R}^2$). (left) The 4-dimensional SCS is represented as two 2D regions, a sphere representing the orientation space S^2 and a rectangle representing the translation space $\mathbb{R}^2$. The dark shaded regions at each node of the tree represent the region in SCS being explored ($A \times T \subset S^2 \times \mathbb{R}^2$). Ultimately (bottom left of the tree) the branch-and-bound algorithm returns regions in 4D space representative of structures that possibly satisfy all the restraints. (middle) At each node, we test satisfaction of each restraint of form $\|\mathbf{p} - \mathbf{q}'\| \leq d$ by testing intersection between the ball of radius d centered at $\mathbf{p}$ and the convex hull bounding possible positions of $\mathbf{q}'$. If there exists an intersection between the ball and the convex hull for each restraint, further branching is done (as for node 1); otherwise, the entire node and its subtree are pruned (as for node 2). (right) The orientations and translations for a node restrict the possible positions of $\mathbf{q}'$. For each node, the four colored curves represent all possible positions of $\mathbf{q}'$ when considering the node's orientation space at each of the four corners of the node's translation space. The colored boxes represent the corresponding axis-aligned bounding boxes (AABBs). The convex hull of the four AABBs (the black box) bounds the positions of $\mathbf{q}'$ for the node.

soft pruning allows for conformations with a few steric clashes in side chains that can be overcome through energy minimization performed later. (A provably correct test for steric clashes is presented in Chapter 5.)

To test whether we can eliminate a cell G due to restraint violation, we independently consider each restraint, $\|\mathbf{p} - \mathbf{q}'\| \leq d$, where $\mathbf{p}$ and $\mathbf{q}'$ are positions of atoms in adjacent subunits in clockwise order. Let position $\mathbf{q}$ correspond to $\mathbf{q}'$ in the same subunit to which $\mathbf{p}$ belongs. We then compute $G\mathbf{q}$, the set of all possible positions of $\mathbf{q}'$ under the symmetries defined by G. If there is an empty intersection between $G\mathbf{q}$ and the ball of radius d centered at $\mathbf{p}$, then none of the structures represented by G satisfy the restraint. Formally, let $B(\mathbf{p}, d)$ be the solid ball in $\mathbb{R}^3$ which has radius d and is centered at the point $\mathbf{p}$. If $G\mathbf{q} \cap B(\mathbf{p}, d) = \emptyset$ then G is eliminated. Under the assumption of exact symmetry, each restraint $\|\mathbf{p} - \mathbf{q}'\| \leq d$ implies another restraint $\|\mathbf{p}' - \mathbf{q}\| \leq d$, where $\mathbf{q}$ and $\mathbf{p}'$ are atoms on the adjacent subunits in counterclockwise order. The satisfaction of this restraint is tested in a similar manner.

The region $G\mathbf{q}$ is characterized by high-degree polynomials and it is computationally expensive to test for intersections with $G\mathbf{q}$. Hence we approximate $G\mathbf{q}$ by a conservative *bounding region*, $W(G, \mathbf{q})$, that completely contains $G\mathbf{q}$ (i.e., $G\mathbf{q} \subset W(G, \mathbf{q})$), but is simpler to compute than $G\mathbf{q}$. The conservative nature ensures that intersection tests between $W(G, \mathbf{q})$ and $B(\mathbf{p}, d)$ provably prune out only structures inconsistent with data. If $W(G, \mathbf{q}) \cap B(\mathbf{p}, d) = \emptyset$, then we know $G\mathbf{q} \cap B(\mathbf{p}, d) = \emptyset$. We compute $W(G, \mathbf{q})$ by first deriving a bounding region from the orientation space (S^2) and then finding the bound from SCS ($S^2 \times \mathbb{R}^2$). The details are as follows.

Bound from orientation space: Let $A \subset S^2$ be a part of the unit sphere representing the

set of orientations in G. Let $A\mathbf{k}$ represent the new positions of $\mathbf{k} \in \mathbb{R}^3$ after rotation around each axis in A by the angle of symmetry, $\alpha = 360^\circ/n$. Since rotations preserve distances, $A\mathbf{k}$ must lie on a sphere of radius $\|\mathbf{k}\|$ centered at the origin. We bound $A\mathbf{k}$ by a spherical cap (region of a sphere which lies above or below a given plane) formed by the intersection of the sphere and a ball. The center and radius of the ball are obtained as follows. Let $\mathbf{a}$ be an axis in A passing through the origin such that all other axes in A lie within a ball of radius r_a centered at $\mathbf{a}$. Hence, $A \subset S^2 \cap B(\mathbf{a}, r_a)$. Let $\mathbf{k}'$ be the position of $\mathbf{k}$ rotated by α radians around axis $\mathbf{a}$. The position $\mathbf{k}'$ is the center of the ball approximating $A\mathbf{k}$ and is obtained as follows:

$$\mathbf{k}' = (\mathbf{k} \cdot \mathbf{a})\mathbf{a} + (\sin\alpha)(\mathbf{a} \times \mathbf{k}) + (\cos\alpha)\,(\mathbf{k} - (\mathbf{k} \cdot \mathbf{a})\mathbf{a})\,. \tag{3.1}$$

The radius, r_k, of the ball approximating $A\mathbf{k}$ is computed as

$$r_k = r_a\Big(\sqrt{(\sin\alpha)^2\|\mathbf{k}\|^2 + (1-\cos\alpha)^2(\mathbf{k}\cdot\mathbf{a})^2} + |1-\cos\alpha|\,\|\mathbf{k}\|\Big)\,. \tag{3.2}$$

Finally, we define our bounding region to be the spherical cap $C = B(\mathbf{k}', r_k) \cap S(\mathbf{0}, \|\mathbf{k}\|)$ where $S(\mathbf{0}, \|\mathbf{k}\|)$ denotes the sphere of radius $\|\mathbf{k}\|$ which is centered at the point $\mathbf{0}$, the origin. By construction, $A\mathbf{k} \subset C$. We approximate C with a bounding box that is aligned along the x-, y- and z-axes—an axis-aligned bounding box (AABB), which we refer to as $V(A, \mathbf{k})$. We use $V(A, \mathbf{k})$ to help us perform a quick test for intersection. If this quick test is passed, we then perform a second, more careful and expensive test for intersection using a tighter bounding region for $A\mathbf{k}$. This tighter bound is the smallest axis-aligned bounding box (AABB) that contains $A\mathbf{k}$. The dimensions of the box are found by performing a numerical global minimization (and maximization) on the x-, y- and z-coordinates of $\mathbf{k}'$.

The global optimization is done by gridding A and starting a gradient descent from each of the points. The details of both our intersection tests and the Equations (3.1) and (3.2) are provided in Supplementary materials of Potluri et al. [124].

Bound from SCS: Let $T \subset \mathbb{R}^2$ and $A \subset S^2$ denote the sets of translations and orientations in G. The region $G\mathbf{q}$ represents the positions of $\mathbf{q}$ on rotation around each axis in $A \times T$ by the angle of symmetry α. To bound $G\mathbf{q}$ and determine $W(G, \mathbf{q})$, we need to find the orientation-based bound (as above) for each translation $\mathbf{t} \in T$. We choose our bounding region $W(G, \mathbf{q})$ as an approximation of the convex hull of $G\mathbf{q}$. Using the properties of convex hulls and the fact that T is convex, we are able to prove that the convex hull of $G\mathbf{q}$ is determined by just the corners of T (see Supplementary materials [124] for proof). Let $H(U)$ be the convex hull of $U \subset \mathbb{R}^3$. It can be shown that $H(G\mathbf{q}) = H(\bigcup_{i=1}^{4}\{A\mathbf{q}_{t_i} + \mathbf{t}_i\})$ where $\mathbf{t}_i$ are the four corners of T and $\mathbf{q}_{t_i}$ denotes the position $(\mathbf{q} - \mathbf{t}_i)$. We use our bounds on regions of $A\mathbf{k}$, the AABB $V(A, \mathbf{k})$, to bound $A\mathbf{q}_{t_i}$. We then bound $G\mathbf{q}$ by finding the convex hull of the AABBs at the corners of T. This convex hull is our bounding region $W(G, \mathbf{q})$. It can be proved that $W(G, \mathbf{q})$ is a conservative bounding region for $G\mathbf{q}$ (see Supplementary materials [124]). That is, $G\mathbf{q} \subset W(G, \mathbf{q})$. Hence, testing satisfaction of a restraint $\|\mathbf{p} - \mathbf{q}'\| \leq d$ requires testing for the intersection of the convex hull of the AABBs at the corners of T (which is a bounding region for $G\mathbf{q}$), with a solid ball centered at $\mathbf{p}$ with radius d. We test the intersection between a ball and a convex polyhedron using the method described in [44].

Branching

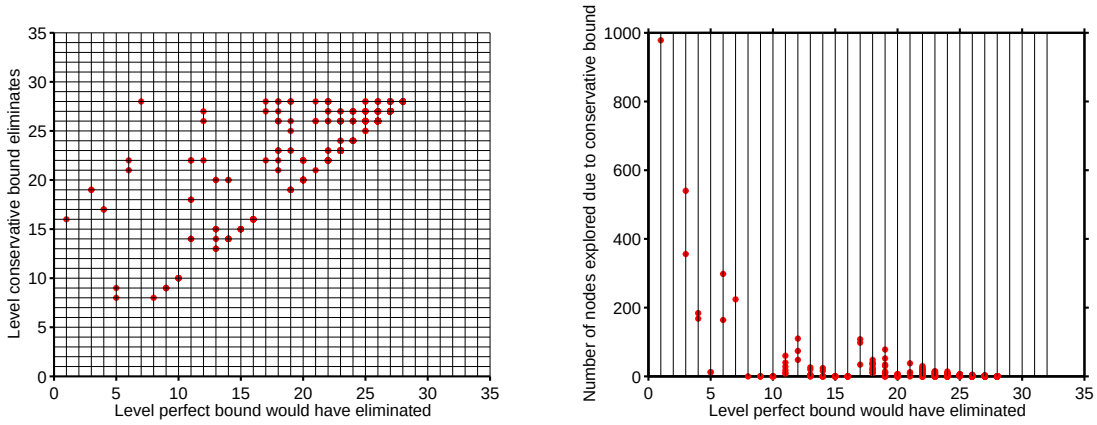
In partitioning a cell into sub-cells, we considered two approaches to choose the dimension to branch along. In the first approach, we seek to divide the cell into two regions along the dimension that will cause one of the restraints to be violated by all the conformations represented by one of the sub-cells. This kind of a division will allow us to efficiently eliminate the sub-cell. We use the following heuristic to achieve this. For each restraint $\|\mathbf{p} - \mathbf{q}'\| \leq d$, we compute $\mathbf{q}'$ for each of the corners of the cell. We then identify the dimension that has the largest difference in $\|\mathbf{p} - \mathbf{q}'\|$ distances for its pair of corners. We partition along that dimension.

In the second approach, we seek to partition a cell, $G = A \times T$, into sub-cells along the dimension that will make our conservative bound as tight as possible. In finding the conservative bound, we approximate the orientation space, A, by a spherical cap. Hence, the bound is tightest when the orientation space is as "square" as possible. Let V_c denote the volume occupied by $G\mathbf{q}$, the bound from G. Let V_r denote the sum of the volumes of the bounds from A at the four corners of T. We choose to partition along the translation space instead of the orientation space when the ratio of V_c to V_r is larger than 1. To partition the translation space, we divide along the x or y dimension, whichever is larger. Similarly, we choose the dimension of maximum length when the orientation space is chosen. In practice, we have found the second heuristic to work better than the first.

Time Complexity of the Search

The time complexity of the hierarchical subdivision depends on the number of nodes explored and the time spent at each node. First let us get an estimate of the number of nodes explored. Let C denote the set of satisfying cells (leaf cells of the hierarchical subdivision) and d be the depth of the search tree. First let us estimate the number of nodes explored when the geometric bound, $G\mathbf{q}$, is perfect, that is we can exactly compute the region of $G\mathbf{q}$. In this case, a worst-case bound on the number of nodes explored is $O(Cd)$. This comes from the fact that at each level of the search tree, we would in the worst case accept $|C|$ cells. The cells that cannot be part of C at each level are immediately eliminated, leading to a constant time to explore the eliminated nodes.

In the case that the bound is conservative, additional nodes are explored. This is because some invalid nodes, nodes that cannot be part of C, are not immediately eliminated. The question then is for each invalid node, k, how many subdivisions must occur for all the sub-cells of k to be eliminated. Given that k is at level l, we need to identify the level at which k is eliminated. Figure 3.2(a) shows the correlation between the level at which the perfect bound and the conservative bound eliminate invalid nodes for one of the test cases (glycophorin A). (Similar results were obtained on other test cases.) As can be seen, the plot is mostly diagonal. The farther away from the diagonal a point is, the greater the number of additional levels. The number of additional levels explored (say m) does not imply that the number of additional nodes explored is exponential ($2^{m+1} - 1$). Figure 3.2(b) indicates the number of additional nodes explored due to the conservative nature of the bound. As



(a) (b)

Fig. 3.2: Empirical analysis of the conservative nature of the bound on glycophorin A. Relationship between the level at which the perfect bound would have eliminated an invalid node to (a) the level at which the conservative bound eliminates it and (b) the number of additional nodes explored due to the conservative bound.

seen, for all invalid nodes beyond level 10, less than 100 additional nodes are explored.

Now, let us look at the time complexity for testing our geometric bound in a cell of the SCS.

Theorem 3.1 *The time complexity for checking satisfaction of a restraint for a pair of atoms in a cell of the 4D SCS is* $O(1)$.

Proof: Consider checking satisfaction of a restraint of the form $\|\mathbf{p}-\mathbf{q}'\| \leq d$ ($\mathbf{p}$ and $\mathbf{q}'$ are atoms on different subunits and d is the distance of the restraint) in a cell of the 4D SCS, $G = A \times T \subseteq S^2 \times \mathbb{R}^2$. This requires determining if there is a non-empty intersection between a ball (centered at $\mathbf{p}$ and of radius d) and $G\mathbf{q}$, the region of possible positions of $\mathbf{q}'$ when the symmetry axis is in G. Since $G\mathbf{q}$ is hard to compute exactly, we approximate it by computing a bound on $A\mathbf{q}$ at the four corners of T and then taking the convex hull of the four bounds (Section 3.1.1). The resulting convex hull is a superset of $G\mathbf{q}$, which we use in our conservative tests for restraint satisfaction. The bound on $A\mathbf{q}$ is obtained as an axis-aligned box (AABB). The time complexity, t, for checking satisfaction of a restraint in a cell, would then be the sum of (a) the time complexity, t_b, for computing the bound on $A\mathbf{q}$ at each of the four corners of T, (b) the time complexity, t_c, for computing the convex hull of the bounds at the corners, and (c) the time complexity, t_i, for computing the intersection between the convex hull and a ball.

(a) The time complexity, t_b, for computing a bound on $A\mathbf{q}$ includes evaluating an analytical expression to compute the center and radius of a bounding ball (Equations 3.1 and 3.2) and computing an axis-aligned box (AABB) of the intersection of the bounding

ball and a sphere. This requires constant time, $t_b = O(1)$. Since the bound on $A\mathbf{q}$ must be obtained at each of the four corners of T, $t_b = 4 \cdot O(1) = O(1)$.

(b) The time complexity, t_c, for computing the convex hull of the AABBs obtained at the four corners of T would be $O(s \log s)$ in the worst case where s is the number of input points to compute the hull. Since we are finding the convex hull of four AABBs, each with eight vertices, s in our case would be 32. Hence, $t_c = O(1)$.

(c) The time complexity, t_i, for computing the intersection between the convex hull and a ball is proportional to the number of triangles of the convex hull. For each triangle, we test if it intersects the ball. Since the complexity for testing each triangle is constant, $t_i = O(f)$, where f is the number of triangles of the convex hull. Using Euler's formula, we can prove that the number of triangles for the convex hull in 3D is at most $2p - 4$ [36]. Here p is the number of vertices of the convex hull which is at most the number of input points, 32 in our case. Hence, $f \leq 2 \cdot 32 - 1$, obtaining $t_i = O(1)$.

Hence the time complexity for checking satisfaction of a possible NOE assignment in a cell of the 4D SCS, $t = t_b + t_c + t_i = O(1)$. □

3.1.2 Determining Satisfying Structures

Our algorithm guarantees that the consistent regions it returns represent every conformation of the symmetric homo-oligomer that is consistent with the data. Because we prune conservatively, they might also represent additional structures inconsistent with the data. To identify the most consistent conformations, we generate representative structures from

the consistent regions. Choosing good representatives ensures that we do not miss native structures (as can occur with sampling-based docking approaches; see Section 2.1.4). We choose the set of representative structures as follows. A cell is accepted as part of the consistent regions only when all structures it represents are within τ_0 Å of each other. We first obtain the structures from the centroids of the cells in the consistent regions. We then cluster these structures using an agglomerative complete linkage hierarchical clustering [67] that allows two structures to be within a cluster only if their backbone RMSD is within τ_0 Å. The centroids of the clusters then form the set of representative structures. This procedure ensures that every structure of the consistent region is within τ_0 Å of at least one representative structure.

Some representative structures might be inconsistent with the data, due to the conservative bounds used when pruning regions. We define the *satisfaction score* for each structure as the sum over the violated NOE restraints of the difference in the expected and the observed distance.

The set of *satisfying structures* are those representative structures with satisfaction scores below a threshold. Note that each satisfying structure represents a set of cells in the consistent regions. The union of all such cells form the *satisfying regions*.

3.1.3 Determining WPS Structures

Having obtained the set of satisfying structures, we now evaluate each of them for packing quality. We first energy-minimize them with the LBFGS conjugate-gradient minimization

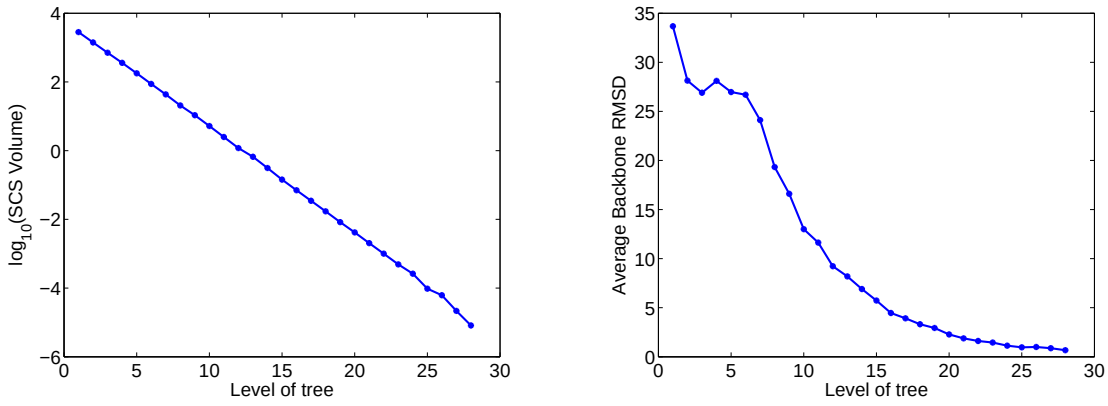
algorithm (10,000 minimization steps) in CNS [23]. The energy function being minimized includes the NOE restraint energy terms as well as the modeling energy terms of vdW (6-12 Lennard-Jones potential), bond length, bond angle, dihedral, and improper energies [23]. We harmonically restrain the backbone and the NOE restraints to ensure that we maintain the symmetry and satisfy the restraints. The minimization accounts for flexibility, uncertainty and asymmetry in the side chains and should help obtain conformations of the side chains that provide good van der Waals packing.

We define the *packing score* of an energy-minimized structure as the difference between the vdW energy of a subunit in the structure when it is in the complex, and the vdW energy of the subunit when it is not in the complex. The difference ensures that only the vdW energy of the inter-subunit surfaces is taken into consideration and indicates the preference of the subunit to be in its apo form versus holo form. A positive value for the difference implies that the subunit prefers its apo form and that the complex is not well-packed.

We then define the set of *WPS structures* as those satisfying structures with packing scores below a threshold, chosen here as 0 kcal/mol since well-modeled structures should have negative packing scores. We refer to the union of cells in the consistent regions corresponding to the WPS structures as the *WPS regions*.

Evaluating Structural Constraint

Our data-driven approach allows us to evaluate the constraint in structure provided by data alone versus data and modeling. We use two metrics. The first metric evaluates the variance



(a) (b)

Fig. 3.3: Correlation of the level of the search tree with (a) log SCS volume and (b) average backbone RMSD of structures represented by the cells.

in the position of each atom across the set of satisfying (or WPS) structures. The second metric finds the 4D volume of the satisfying (or WPS) regions. By comparing these measures for satisfying versus WPS structures and regions, we evaluate the structural constraint from data alone versus both data and packing together. In order to get an intuition of how the volume corresponds to uncertainty in the conformations, we obtained the correlation between volume and RMSD to the level of the search tree. Figure 3.3(a) indicates the correlation between the 4D volume and the level of the search tree and Figure 3.3(b) indicates the correlation between the average backbone RMSD among conformations represented by cells at a level and the level of the search tree. (These figures were obtained by running the core approach on each of our test cases and taking the average over all test cases.) The log (volume) decreases linearly as the level of the tree increases. Similarly, the average

backbone RMSD decreases as the level of the tree increases. From these two figures we can deduce that the volume is roughly proportional to the RMSD of the structures.

Determining Oligomeric Number

Our method also allows computational determination of the oligomeric number of the complex. The confidence in the determined oligomeric number depends on the information content in the available data. We can determine whether the available data suffices to determine the oligomeric number with high confidence. For each possible oligomeric number, we determine a set of WPS structures. We place higher confidence in the oligomeric number that has WPS structures with better vdW packing. Thus we determine the oligomeric number using the NMR data and vdW packing. Our approach provides for an independent verification of the oligomeric state, which is typically determined using experiments such as chemical cross-linking followed by SDS-PAGE, or by equilibrium sedimentation.

We expect structures obtained from the correct oligomeric number to satisfy the data and have vdW packing better than structures from other possible oligomeric numbers. Hence, restraint satisfaction and vdW packing should help discriminate among putative oligomeric numbers. We test this possibility by searching in 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.

3.2 Results

We validated our core SCS algorithm by testing it on a number of proteins of different oligomeric numbers (Section 1.2.3) describes these test cases). The subunit structure was placed with its center of mass at the origin and its principal axis along the z-axis. The translation parameters, x and y, indicate the position of the symmetry axis relative to the origin. Similarly the orientation parameters theta and phi indicate the orientation of the symmetry axis on a unit sphere. Theta denotes the latitude and phi denotes the longitude on the unit sphere. We declared a structure to have a steric clash when there were at least five pairs of atoms such that each pair was separated by less than 1.5 Å. Energy minimization of structures, in the second phase of our approach, allows side chain flexibility such that one or two steric clashes can be eliminated. The value of τ_0, the user-defined similarity level, was chosen as 1 Å^2. The threshold for the satisfaction score was chosen as 1 Å. We will first discuss our results on phospholamban, the homo-pentamer that motivated the research in this thesis, and then discuss the results on several other proteins: glycophorin A, haemagglutinin, potassium channel and co-chaperonin.

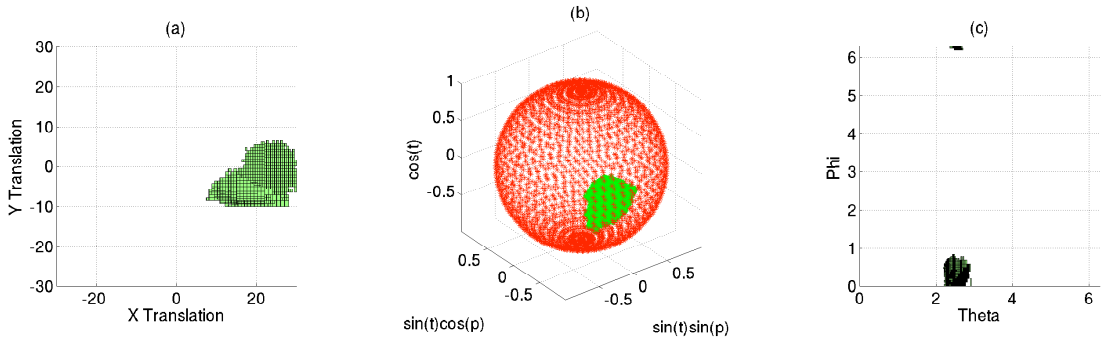
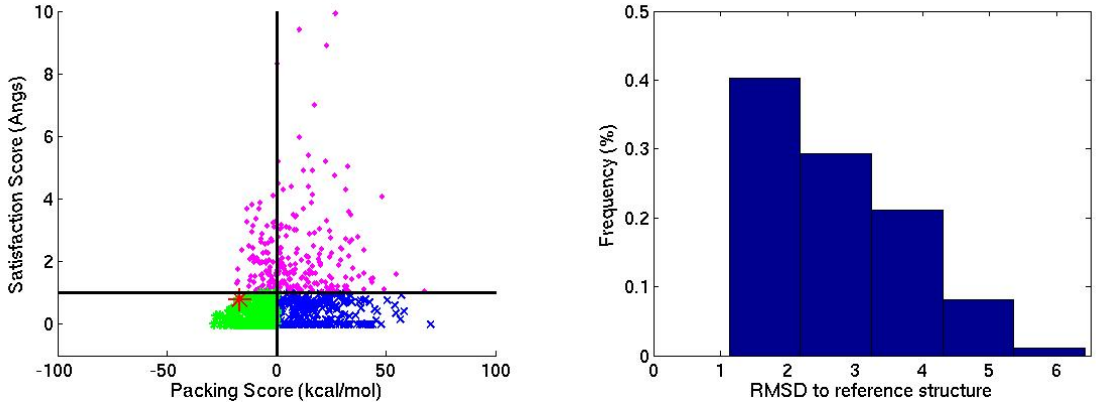


Fig. 3.4: Phospholamban Results: Consistent regions using the nine experimental restraints and an oligomeric number of 5. Subplot (a) indicates the translation parameters of the consistent regions. Subplots (b) and (c) indicate orientation parameters of consistent regions both on a sphere and when projected onto a plane of theta and phi angles.

3.2.1 Results on phospholamban

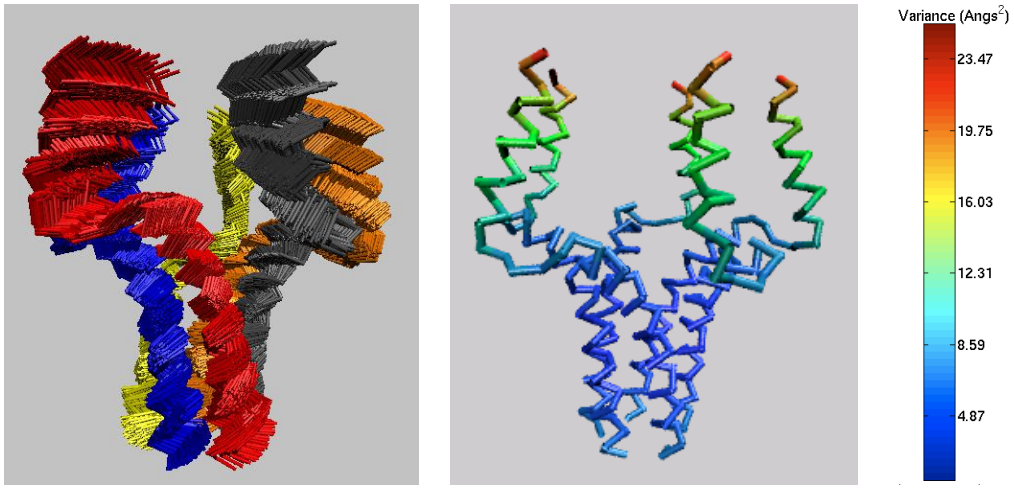
The branch-and-bound algorithm was run on phospholamban using the nine available experimental restraints. The algorithm returned a set of 5195 cells corresponding to the leaves remaining after pruning infeasible regions (restraint violation or steric clash) of the 4D configuration space. The solution cells are shown in Figure 3.4. A huge chunk of the 4D space (approximately 45238 $Å^2$-rad^2) was pruned leaving consistent regions as shown in the figure. The remaining volume in SCS is 1.24 $Å^2$-rad^2.

Figure 3.5(a) plots the packing scores versus the satisfaction scores for phospholamban. The set of satisfying structures has an overall range of 1.07 Å to 8.85 Å backbone RMSD to the reference structure. This range indicates the diversity in structures possible using just



(a) (b)

Fig. 3.5: Phospholamban results: (a) Restraint satisfaction score vs. packing score for all representative structures. The vertical and horizontal lines indicate the chosen cutoffs for WPS structures: 1 Å for the satisfaction score and 0 kcal/mol for the packing score. The green stars and the blue crosses indicate the set of satisfying structures. The magenta points indicate the set of representative structures that are eliminated due to high satisfaction scores. The green stars indicate the set of WPS structures and the red star indicates the reference structure. (b) Histogram of backbone RMSD to the reference structure for the WPS structures.



(a) (b)

Fig. 3.6: Phospholamban structures: (a) The set of WPS structures after alignment to the structure with lowest packing score. Different chains are in different colors. (b) Variance of the backbone atoms illustrated by the color scale shown with blue indicating maximum variance and red minimum variance.

the nine experimental inter-subunit NOE restraints. The average variance in the positions of the atoms in the set of satisfying structures is 12.32 Å^2. The satisfying region volume (1.24 Å^2-rad^2) and the average variance indicate the constraint on structure provided by the data alone.

Figure 3.5(a) also shows the scores of the WPS structures. The reference structure has a satisfaction score of around 0.8 Å and packing score of -17 kcal/mol, and it lies in the WPS region. Figure 3.5(b) shows the backbone RMSD of the reference structure to each of the WPS structures. Incorporating packing quality reduces the maximum RMSD to the reference structure from 8.85 Å to 6.24 Å. The area of the translation space is reduced from

290 Å^2 to approximately 135 Å^2 and that of the orientation space from 0.40 radian^2 to approximately 0.23 radian^2. The volume of the 4D space is reduced from 1.24 Å^2-radian^2 in the set of satisfying structures to approximately 0.51 Å^2-radian^2. All these values indicate the additional constraint that packing quality imposes on the structure of phospholamban. The average variance in the positions of the atoms is reduced from 12.32 Å^2 to 6.80 Å. Figure 3.6(a) illustrates all the WPS structures and Figure 3.6(b) illustrates the variance of each of the backbone atoms. The figures show that there is more uncertainty in the amphipathic helices than the transmembrane helices. The average variance in the amphipathic helices is 10.75 Å^2 whereas that in the transmembrane helices is 2.96 Å^2. This is because the experimental restraints are between residues in the transmembrane helices. This agrees with the observation in [119] that the amphipathic helices are less well determined. From this we conclude that we need more restraints in the amphipathic helices to determine the structure with greater precision.

We further tested whether the experimental data available are sufficient to choose one oligomeric number over others with reasonable confidence. When we applied our approach to determine the oligomeric number on phospholamban, WPS structures were present only for oligomeric numbers of tetramer, pentamer, hexamer, and heptamer. The lowest packing scores obtained were $E_l(4) = -21.80$ kcal/mol, $E_l(5) = -28.44$ kcal/mol, $E_l(6) = -19.28$ kcal/mol, and $E_l(7) = -15.52$ kcal/mol. $E_l(5)$ has the lowest packing score, correctly suggesting that the pentamer is the most feasible oligomeric number. We expect that with the availability of more experimental data, we can determine the oligomeric number with greater confidence.

3.2.2 Results on other proteins

The branch-and-bound algorithm was run on each of the four proteins: glycophorin A, haemagglutinin, potassium channel and co-chaperonin (Section 1.2.3). The results are shown in Figures 3.7, 3.8, 3.9, 3.10. For the dimer case (glycophorin), the search is performed on half of the orientation space. This is because symmetry axes oriented exactly reverse of each other lead to the same dimer structure. Note the small region in configuration space (translation space 1.10 Å^2 and rotation space 0.001 rad^2) for haemagglutinin. The reason for this is the presence of 85 restraints and the coiled-coil fold of the complex. The tetramer and the heptamer, having more pancake-like folds, have considerable spreads in translation and rotation spaces even with the availability of 32 and 85 restraints respectively.

Tables 3.1 and 3.2 summarize the results. Figure 3.11 shows for each test case a plot of packing scores versus satisfaction scores. Table 3.1 shows the results on the satisfying regions and satisfying structures and Table 3.2 shows the results on WPS regions and WPS structures. The tables and figure clearly show that except for human glycophorin A, the remaining test cases have identical sets of satisfying structures and WPS structures. Further, the spread in the satisfying region and the WPS region is almost the same. The reason for this similarity might be that we have used all possible restraints (85, 32, and 85) for these test cases. This use of all restraints causes almost all the satisfying structures returned by the branch-and-bound algorithm to have high-quality vdW packing and hence to belong to the set of WPS structures.

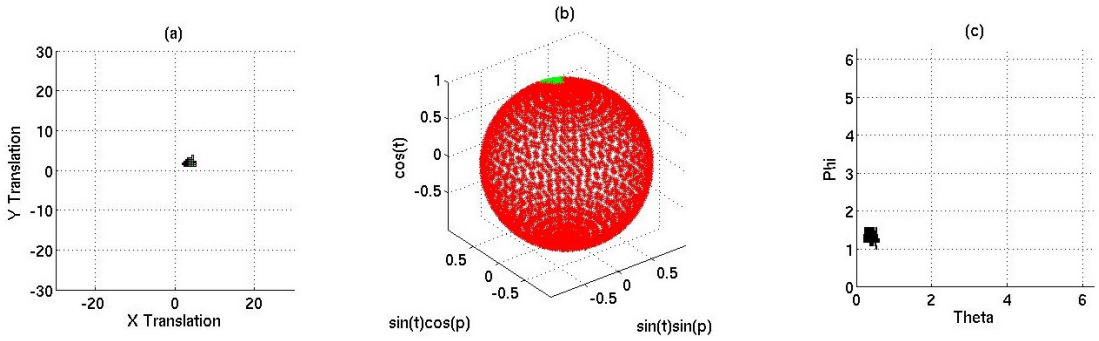


Fig. 3.7: Glycophorin A Results: Consistent regions using the six experimental restraints and an oligomeric number of 2. Subplot (a) indicates the translation parameters of the consistent regions. Subplots (b) and (c) indicate orientation parameters of consistent regions both on a sphere and when projected onto a plane of theta and phi angles.

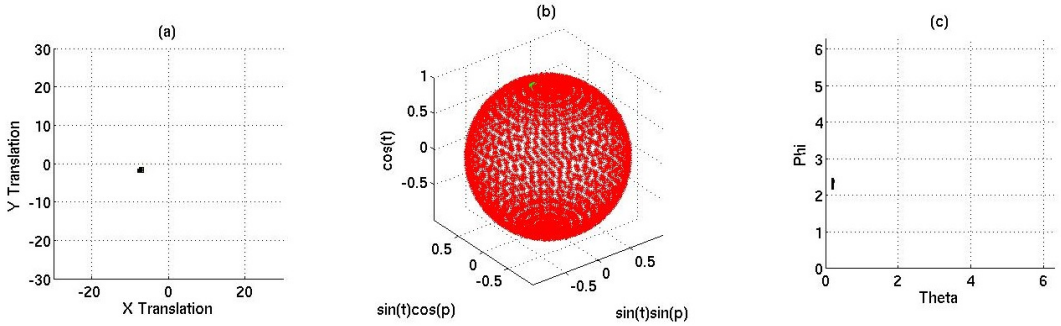


Fig. 3.8: Haemagglutinin Results: Consistent regions using the 85 simulated restraints and an oligomeric number of 3. Subplot (a) indicates the translation parameters of the consistent regions. Subplots (b) and (c) indicate orientation parameters of consistent regions both on a sphere and when projected onto a plane of theta and phi angles.

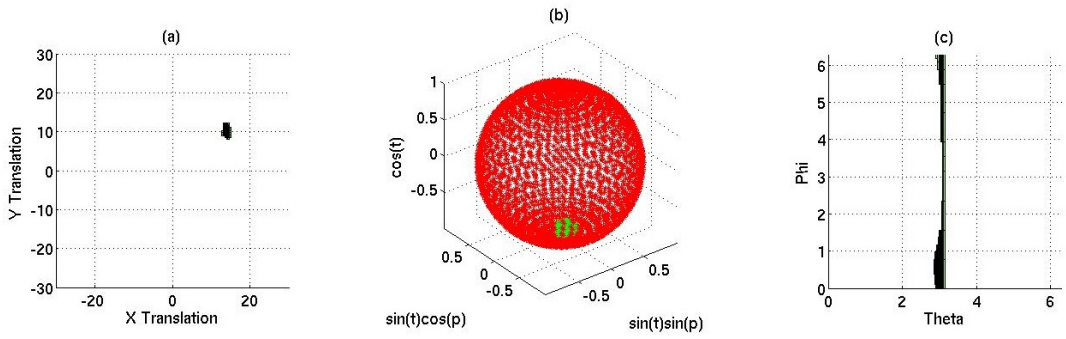


Fig. 3.9: Kv1.2 Potassium Channel Results: Consistent regions using the 32 simulated restraints and an oligomeric number of 4. Subplot (a) indicates the translation parameters of the consistent regions. Subplots (b) and (c) indicate orientation parameters of consistent regions both on a sphere and when projected onto a plane of theta and phi angles.

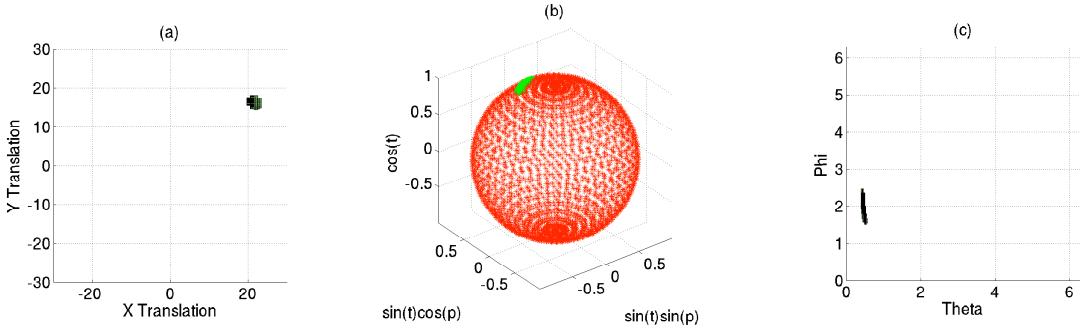


Fig. 3.10: Gp31 co-chaperonin Results: Consistent regions using the 85 simulated restraints and an oligomeric number of 7. Subplot (a) indicates the translation parameters of the consistent regions. Subplots (b) and (c) indicate orientation parameters of consistent regions both on a sphere and when projected onto a plane of theta and phi angles.

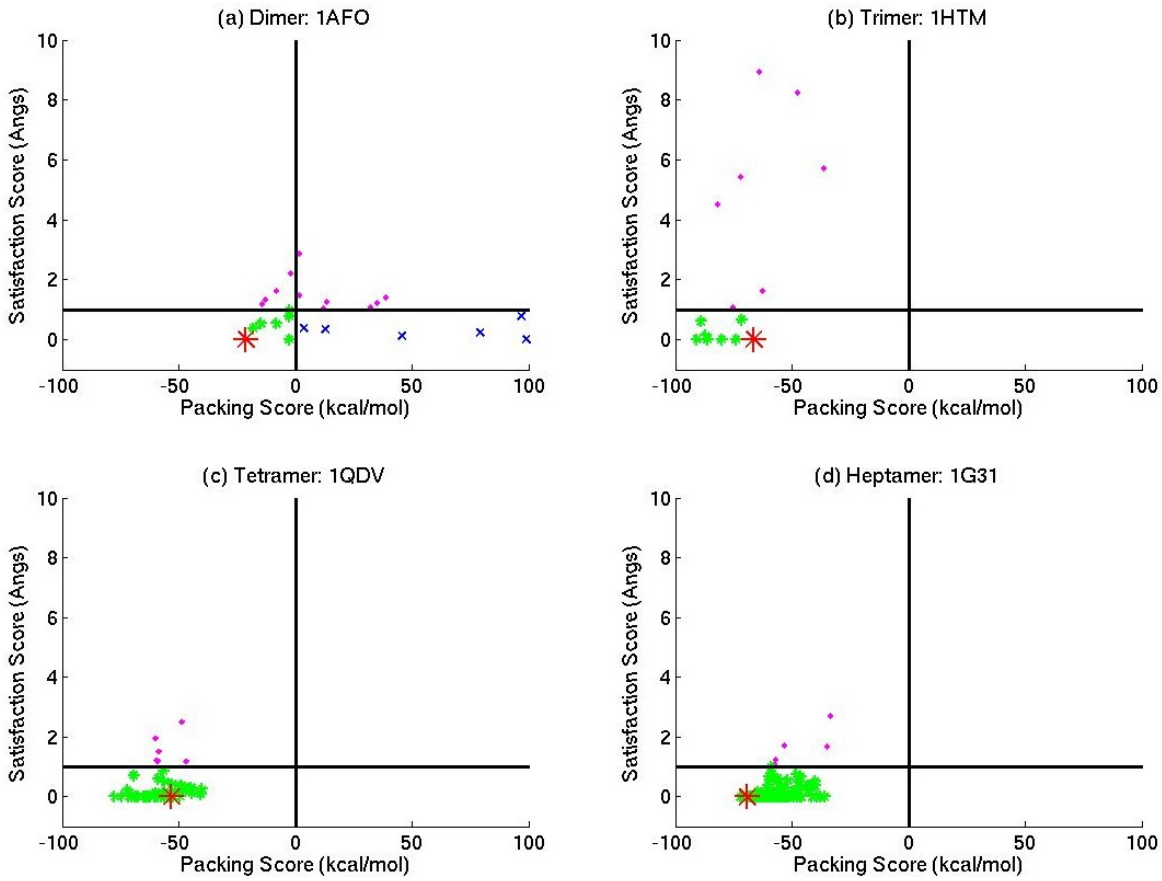


Fig. 3.11: Restraint satisfaction score vs. packing score for all satisfying structures of (a) human glycophorin A (dimer:1AFO), (b) influenza haemagglutinin (trimer:1HTM), (c) Kv1.2 potassium channel (tetramer:1QDV), and (d) Gp31 co-chaperonin (heptamer:1G31). The vertical and horizontal lines indicate the chosen cutoffs for WPS structures: 1 Å for the satisfaction score and 0 kcal/mol for the packing score. The green stars and the blue crosses (when present) indicate the set of satisfying structures. The magenta points indicate the set of representative structures that are eliminated due to high satisfaction scores. The green stars indicate the set of WPS structures and the red star indicates the reference structure.

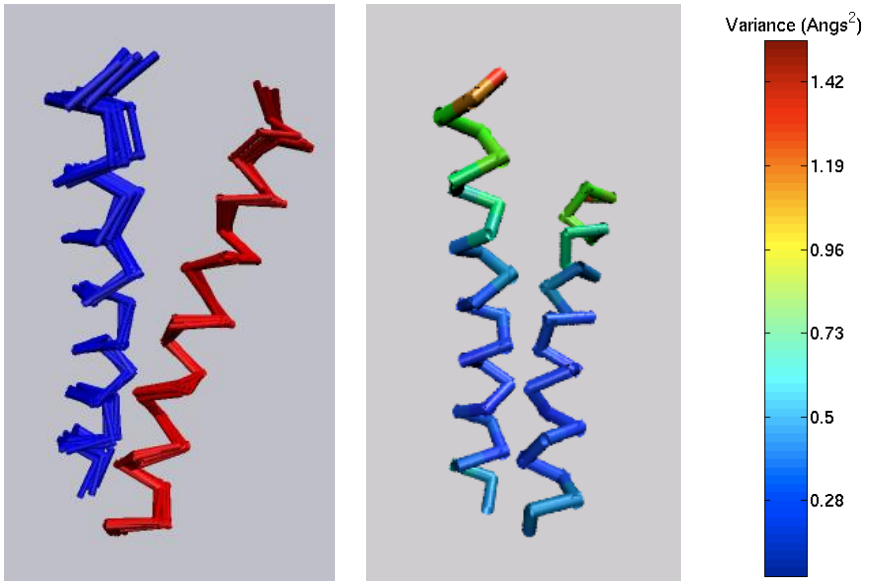
Tab. 3.1: Satisfying structure results: backbone RMSD of the set of satisfying structures to the reference structure, the uncertainty in SCS represented by the area of the translation (T) and orientation (A) space for the satisfying region and the 4D volume of the satisfying region, and finally the variance in the position of atoms for the set of satisfying structures.

Protein (symmetry)	PDB id	No. of restraints	RMSD (Å)		Uncertainty in SCS			Variance in atoms (Å^2)		
			min	max	T (Å^2)	A (rad^2)	Volume (Å^{2}rad^2)	min	max	mean
Glycophorin A (2)	1AFO	6 (expt)	0.61	1.77	4.72	0.06	0.06	0.34	2.62	0.97
Haemagglutinin (3)	1HTM	85 (simulated)	0.86	1.08	1.10	0.001	4e-4	0.07	0.64	0.22
Kv1.2 potassium channel (4)	1QDV	32 (simulated)	0.92	2.85	7.47	0.07	0.13	0.24	4.91	1.24
Phospholamban (5)	1ZLL	9 (expt)	1.07	8.85	289.53	0.40	1.24	2.87	43.97	12.32
Gp31 co-chaperonin (7)	1G31	85 (simulated)	0.40	2.72	21.20	0.07	0.15	0.36	7.73	1.66

Tab. 3.2: WPS structure results: backbone RMSD of the set of WPS structures to the reference structure, the uncertainty in SCS represented by the area of the translation (T) and orientation (A) space for the WPS region and the 4D volume of the WPS region, and finally the variance in the position of atoms for the set of WPS structures.

Protein (symmetry)	PDB id	No. of restraints	RMSD (Å)		Uncertainty in SCS			Variance in atoms (Å^2)		
			min	max	T (Å^2)	A (rad^2)	Volume (Å^{2}rad^2)	min	max	mean
Glycophorin A (2)	1AFO	6 (expt)	0.61	1.77	2.97	0.09/2	0.04	0.07	1.51	0.47
Haemagglutinin (3)	1HTM	85 (simulated)	0.86	1.08	1.10	0.001	4e-4	0.07	0.64	0.22
Kv1.2 potassium channel (4)	1QDV	32 (simulated)	0.92	2.85	7.47	0.07	0.12	0.24	4.91	1.24
Phospholamban (5)	1ZLL	9 (expt)	1.07	6.24	135.50	0.23	0.51	1.52	24.96	6.80
Gp31 co-chaperonin (7)	1G31	85 (simulated)	0.40	2.72	21.20	0.07	0.15	0.36	7.73	1.66

Figures 3.12, 3.13, 3.14, and 3.15 illustrate the uncertainty in the set of WPS structures for each of the test cases. Despite using 32 and 85 restraints, the potassium channel and co-chaperonin have considerable spread in the translation space. The variance in the position of the atoms is also high, with a maximum as high as 4.9 Å^2 and 7.7 Å^2, respectively. The higher uncertainty is because the chosen restraints (restricted to those with exact symmetry) are not distributed all along the inter-subunit surface, but are concentrated toward one end of the surface (Figure 3.16(a)). On the other hand, haemagglutinin is a long helical trimer and the 85 restraints are distributed across the entire inter-subunit surface, thereby yielding less uncertainty (Figure 3.16(b)). This indicates the effect of *independence* of restraints on the uncertainty.

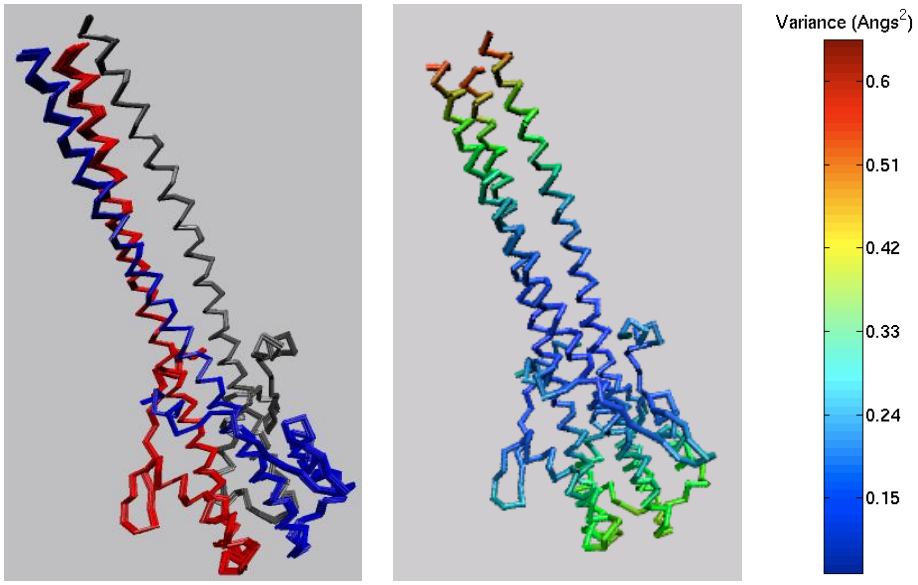


(a) (b)

Fig. 3.12: Glycophorin A (1AFO) structures: (a) The set of WPS structures after alignment to the structure with lowest packing score. Different chains are in different colors. (b) Variance of the backbone atoms illustrated by the color scale shown with blue indicating maximum variance and red minimum variance.

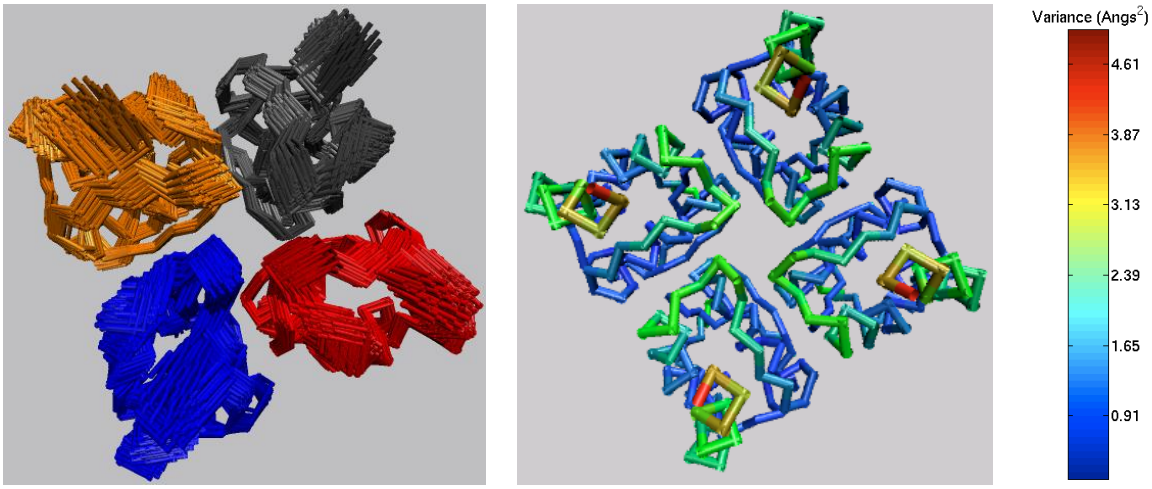
The reference structure lies in the WPS region for all cases, and Table 3.2 indicates the range of RMSDs to the reference structure. Figure 3.17 plots histograms of the backbone RMSD of the set of WPS structures to the reference structure. The histograms peak at 1 Å RMSD, which indicates that the structures obtained are close to the reference structure. The potassium channel (1QDV) and co-chaperonin (1G31) have larger ranges. It is interesting to note that the dimer, with as few as six experimental restraints, provides for comparatively much less uncertainty. This smaller uncertainty might be because the restraints are spread out across the subunit.

To test the hypothesis that the position of the restraints affects the uncertainty in the structure, we performed the following test. We tested the change in structural uncertainty



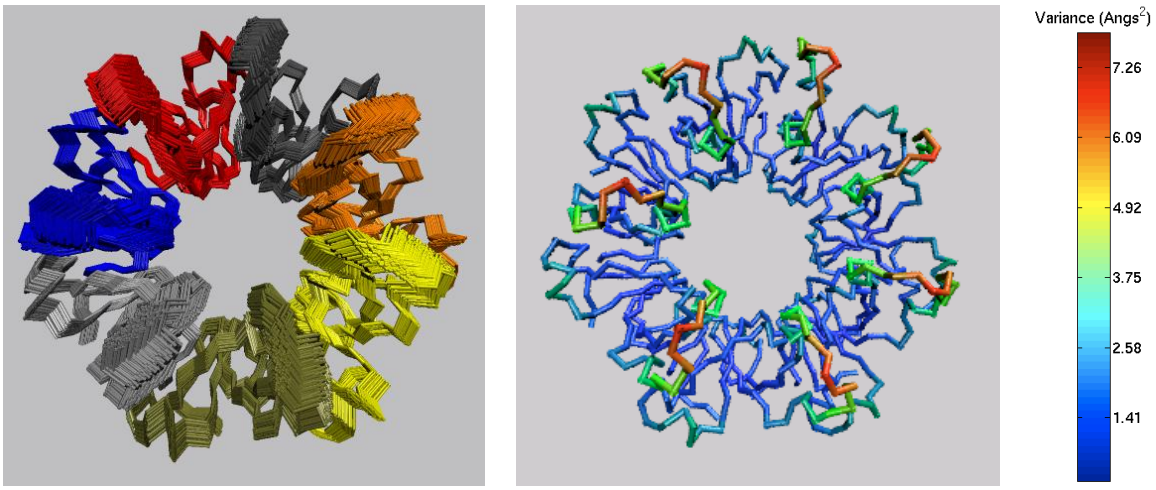
(a) (b)

Fig. 3.13: Haemagglutinin (1HTM) structures: (a) The set of WPS structures after alignment to the structure with lowest packing score. Different chains are in different colors. (b) Variance of the backbone atoms illustrated by the color scale shown with blue indicating maximum variance and red minimum variance.



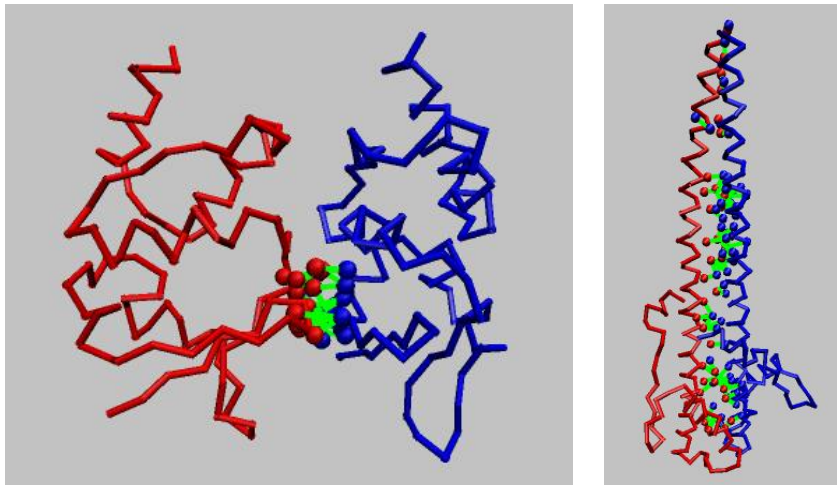
(a) (b)

Fig. 3.14: Kv1.2 potassium channel (1QDV) structures: (a) The set of WPS structures after alignment to the structure with lowest packing score. Different chains are in different colors. (b) Variance of the backbone atoms illustrated by the color scale shown with blue indicating maximum variance and red minimum variance.



(a) (b)

Fig. 3.15: Gp31 co-chaperonin (1G31) structures: (a) The set of WPS structures after alignment to the structure with lowest packing score. Different chains are in different colors. (b) Variance of the backbone atoms illustrated by the color scale shown with blue indicating maximum variance and red minimum variance.



(a) (b)

Fig. 3.16: The spread of simulated restraints across the first two chains in the reference structure of (a) Kv1.2 potassium channel (1QDV) (b) haemagglutinin (1HTM). The red and blue segments indicate the first two chains with the green lines indicating the restraints. The red and blue balls are the atoms on the chains between which the restraints exist.

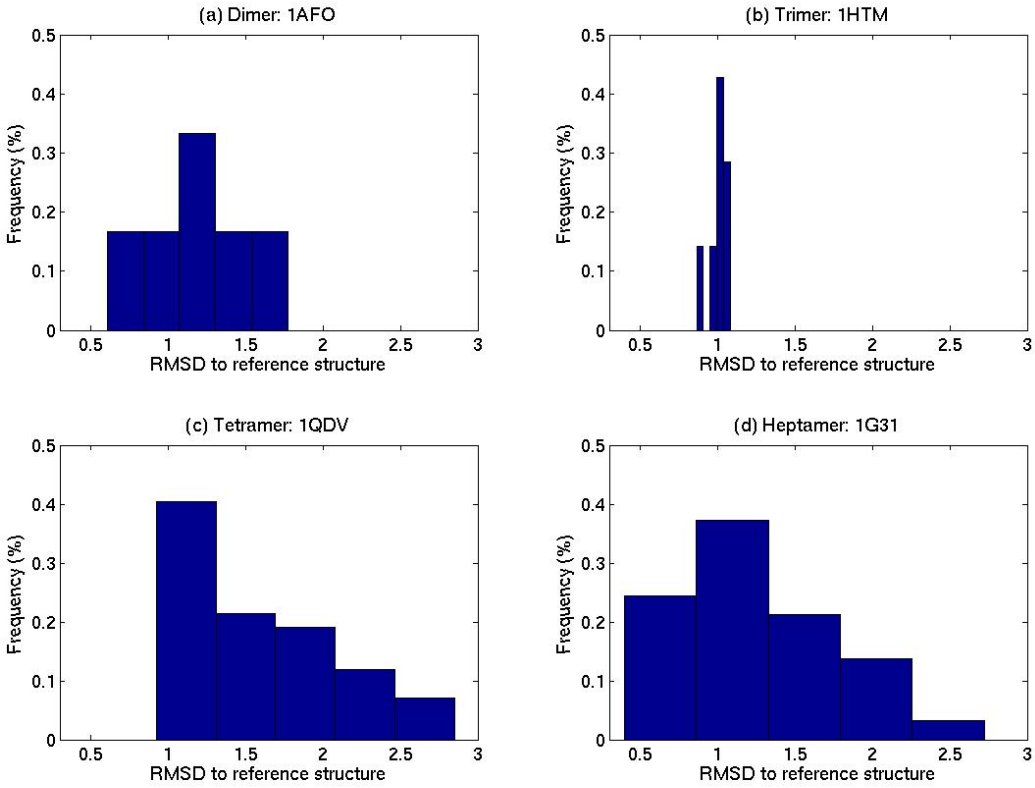
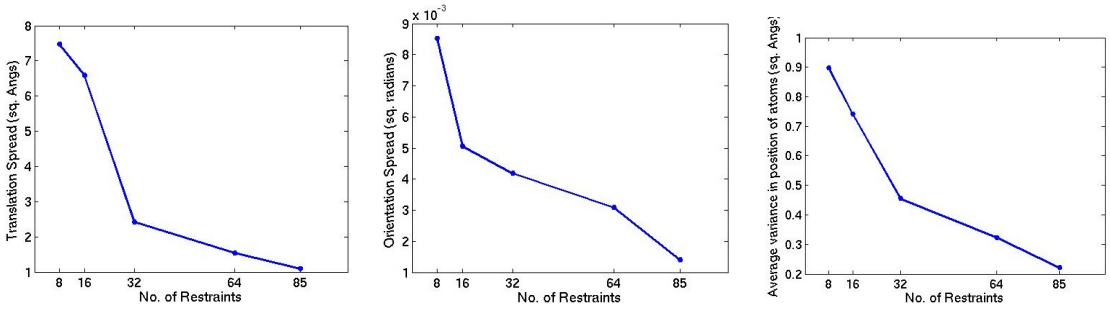


Fig. 3.17: Histogram of backbone RMSD to the reference structure for the WPS structures returned by our approach for (a) human glycophorin A (dimer:1AFO), (b) influenza haemagglutinin (trimer:1HTM), (c) Kv1.2 potassium channel (tetramer:1QDV), and (d) Gp31 co-chaperonin (heptamer:1G31).



(a) (b) (c)

Fig. 3.18: Change with number of simulated restraints of (a) translation spread of WPS regions, (b) orientation spread of WPS regions, and (c) average variance in atomic positions for influenza haemagglutinin (1HTM).

of haemagglutinin as the number of relatively independent restraints changed. We refer to a set of restraints as *relatively independent* if the restraints have no common atoms and are far apart from each other. We choose the set of independent restraints by the following heuristic. We consider midpoints of each possible restraint and represent distances between the restraints by distances between their midpoints. We choose as the first two members of the set, the two restraints that have their midpoints farthest apart. We choose the third restraint as the one whose midpoint has the largest minimum distance from the first two members. We continue this process iteratively to choose the remaining restraints. Figure 3.18 illustrates the change in the translation spread, orientation spread, and average variance in position of atoms with an increasing number of independent restraints. The average variance in the atom positions decreases from 0.9 $Å^2$ to 0.2 $Å^2$ as the number of independent restraints increases. This analysis helps us quantify the minimum number of

independent restraints required to determine the structure of the homo-oligomer up to a specified precision. For example, if we are willing to tolerate approximately 1 Å^2 uncertainty in the positions of the atoms, about eight independent restraints will be sufficient. On the other hand, if we want high precision and are not willing to tolerate uncertainty above 0.3 Å^2 in the positions of atoms, we need as many as 64 independent restraints.

We also applied our method for determination of oligomeric number to the four proteins. For glycophorin A, haemagglutinin, and the Kv1.2 potassium channel we could determine the oligomeric number with high certainty because the set of WPS structures was empty for oligomeric numbers other than the correct one. Glycophorin A is especially interesting since we determined the correct oligomeric number using just six restraints. For the co-chaperonin, oligomeric numbers of hexamer, heptamer, and octamer have WPS structures with $E_l(6) = -69.09$ kcal/mol, $E_l(7) = -72.19$ kcal/mol, and $E_l(8) = -68.92$ kcal/mol. In this case too, since $E_l(7)$ is the lowest packing score, we correctly conclude that the oligomeric number is 7.

We continued our study of the effect of number of independent restraints for haemagglutinin (1HTM). When the number of restraints chosen was 85, 64, or 32, WPS structures were absent for oligomeric numbers other than 3, allowing us to determine with high confidence that haemagglutinin is a trimer. With only 16 or 8 restraints, we obtained WPS structures for oligomeric numbers other than 3. As expected, as the available experimental data increases, our confidence in determining the oligomeric number increases.

3.3 Conclusions

In this chapter, the core algorithm of our complete, data-driven approach for structure determination of symmetric homo-oligomers was presented. We showed that our approach identifies all conformations that are consistent with NOE restraints and that display high-quality vdW packing from our results on five test cases. From our results on phospholamban which had 9 experimental NOE restraints, we showed that our approach is particularly important in sparse-data cases, where relying on an incomplete, biased search (as in the classical approaches) may result in missing well-packed, satisfying conformations. By being complete and data-driven, our approach enables objective evaluation of the amount of structural uncertainty provided by data versus by data and modeling. We show that the average variance in structures is decreased from 12.32 Å^2 to 6.80 Å^2 for phospholamban and 0.97 Å^2 to 0.47 Å^2 for glycophorin A when data alone versus data and packing are considered. By first searching for regions of conformation space consistent with the NOE restraints, and then filtering these regions according to predicted quality of packing, our approach makes good use of the relatively greater discriminatory information content in sparse NOEs to focus on plausible conformations for subsequent analysis by relatively finer-grained packing metrics.

The approach discussed in this chapter assumes that the inter-subunit NOEs are unambiguously assigned. In the next chapter, we extend the approach discussed in this chapter to handle ambiguity in NOEs in a robust manner and perform NOE assignment.

4. RESOLVING AMBIGUITY FOR INTER-SUBUNIT NOE ASSIGNMENT

This chapter extends the core algorithm described in Chapter 3, relaxing the assumption that the NOEs are correctly assigned, and dealing with the ambiguity present in assigning NOEs. As mentioned in Chapter 2, the sources of ambiguity in inter-subunit NOE assignment are *intra- vs. inter-subunit ambiguity*, whether the restrained atoms are within the same subunit or in different subunits, *subunit ambiguity*, the subunits to which the restrained atoms of an inter-subunit NOE belong, *atom ambiguity*, the identities of the restrained atoms among those with chemical shifts similar to the NOE peak. *Intra- vs. inter-subunit* ambiguity can be resolved experimentally by labeling strategies while both subunit ambiguity and atom ambiguity require computational solutions. In this chapter, we develop a complete approach that simultaneously resolves subunit and atom ambiguity in inter-subunit NOE assignment and determines C_n symmetric homo-oligomeric structures, given the subunit structure.

As discussed, a configuration space representation allows us to be complete, data-driven and efficient. We extend the configuration space to include not only symmetry axis parameters, but also all possible assignments of the NOEs in terms of atom and subunit identities.

Our algorithm considers NOEs sequentially, i.e., one after the other, pruning parts of the configuration space in which axes and assignments are mutually inconsistent. Furthermore, the results of our algorithm are deterministic and not affected by the order in which NOEs are considered. Pruning occurs only due to provable inconsistency and thereby avoids the pitfall of local minima that could arise from best-first sampling-based approaches. Ultimately, we return a mutually-consistent set of conformations and NOE assignments. The developed algorithm is complete, as in the previous chapter, in that it identifies structures representing, to within a user-defined similarity level, every structure consistent with the available data (ambiguous or not). However, it avoids explicit enumeration of the exponential number of combinations of possible assignments. It requires time only linear in the number of ambiguous NOEs, the number of possible assignments for each, and the set of 4D cells in the SCS. Our algorithm can draw two types of conclusions not possible under existing methods: (a) that different assignments for an NOE would lead to different structural classes, or (b) that it is not necessary to assign an NOE since the choice of assignment would have little impact on structural precision. Note that our approach focuses on inter-subunit NOE assignment and currently cannot be applied to intra-subunit NOE assignment for monomer structure determination.

4.1 Problem Formulation

The problem that we must solve is the following. We are given a set of inter-subunit NOEs (each possibly atom- and subunit-ambiguous), the subunit structure and the oligomeric

number. Our goal is to determine a mutually-consistent set of NOE assignments and regions in the SCS (symmetry configuration space), such that the regions represent all conformations consistent with the assignments, and the assignments are all those consistent with the regions. Recall (Section 2.2), that a subunit assignment of the NOE specifies the relative positions of the subunits and atom assignment of the NOE specifies the atoms involved in the NOE. Let us first formalize the representation of subunit and atom assignments.

Subunit assignment: An NOE could capture an interaction between any pair of protons in any pair of subunits. Under C_n symmetry, each interaction is "mirrored", i.e., holds between each pair of subunits that are indistinguishable under C_n rotation. Thus a subunit assignment for an NOE specifies the difference (1 to $n-1$) in the positions of the involved subunits within the symmetric order.

Atom assignment: In typical approaches to NMR structure determination, the chemical shift assignments of the various atoms are obtained in an earlier analysis, and are then used to determine which atoms' interactions might have generated the NOE peaks. The possible identities of the interacting protons are determined by comparing the chemical shift coordinates of an NOE peak with the assigned chemical shifts (the same for each subunit, by C_n symmetry). In the 3D spectra used here, there are two coordinates (^{15}N/^{13}C and ^{1}H) by which to determine one side of the interaction and one coordinate (^{1}H) for the other. Since two coordinates generally result in a unique assignment, we here consider only one-sided atom ambiguity, although our approach can readily be generalized. Thus an atom assignment for an NOE specifies the identity of the ambiguous atom, from the entire set A of atoms (here we consider only protons) in a monomer. Typically a user-specified match

tolerance ε (e.g., 0.04 ppm) limits the possible protons to those with similar chemical shift to the peak coordinate.

In summary, the assignment for an NOE specifies a pair $(i, j) \in \mathbb{Z}_{n-1} \times A$ where i assigns the relative positions of the subunits in the complex and j assigns the proton in the subunit. Here $\mathbb{Z}_{n-1}$ represents $\mathbb{S}_n$, the symmetric group of n elements. Each element of $\mathbb{S}_n$ is represented by an integer in $\{1, \ldots, n-1\}$.

Recall from Chapter 3 that the SCS is the space of all symmetry axis parameters, each defining a homo-oligomeric complex structure. Thus we seek to simultaneously assign NOEs (both subunit and atom assignment) and determine structure (by identifying regions in SCS), in a manner that guarantees that we find all consistent assignments and structures. In order to make such a guarantee, we must carefully formulate what it means for an NOE, or one of its possible assignments, to be *inconsistent* with respect to a given set of NOEs and the structures that they define. In the process, we are also able to formulate what it means to be *redundant*. In addition to helping us prove (see Section 4.3) that our approach is correct, these formulations allow us to characterize (Section 4.5) the information content provided by a set of NOEs regarding a homo-oligomeric structure.

Let R be a set of (possibly ambiguous) NOEs. Let r_{kl} represent the l^{th} possible assignment of k^{th} NOE of R, and let $s_{kl} \subset S^2 \times \mathbb{R}^2$ be the region of the SCS in which r_{kl} is satisfied. A determined structure must satisfy all ambiguous NOEs; it satisfies each by satisfying one or more of its assignments. In the SCS, this translates into finding the region

$p(R)$ defined as:

$$p(R) = (s_{11} \cup s_{12} \cup s_{13} \cup \ldots) \cap (s_{21} \cup s_{22} \cup \ldots) \cap \ldots = \bigcap_{k=1}^{\|R\|} \left(\bigcup_{l=1}^{t_k} s_{kl} \right) \tag{4.1}$$

where t_k is the number of possible assignments for the k^{th} NOE.

Using Equation (4.1), we can give the definition for *redundant* and *inconsistent* NOEs and assignments. Let $Q \subset R$ be a set of NOEs. Let r_k be an (arbitrary) ambiguous NOE (not in Q) with t_k assignments. As above, let r_{kl} be its l^{th} possible assignment and s_{kl} be the region of the SCS in which the distance restraint r_{kl} is satisfied. We say r_{kl} is *redundant with respect to* Q if and only if $p(Q) \subset s_{kl}$. We say r_{kl} is *inconsistent with respect to* Q if and only if $p(Q) \cap s_{kl} = \emptyset$. We can extend these notions from assignments to NOEs as follows. Let $S_k = \bigcup_{l=1}^{t_k} s_{kl}$ be the region of the SCS which satisfies at least one assignment of r_k (the k^{th} NOE). We then declare r_k to be *redundant with respect to* Q if and only if $p(Q) \subset S_k$, and declare r_k to be *inconsistent with respect to* Q if and only if $p(Q) \cap S_k = \emptyset$. Our algorithm eliminates inconsistent NOEs and assignments, and is able to detect redundant ones.

4.2 Ambiguity Resolution Algorithm

Although we would like to compute exactly the region $p(R)$ (Equation 4.1) defining all consistent structures, it is a semi-algebraic set characterized by high-degree polynomials that are expensive to solve exactly. Our ambiguity resolution algorithm computes a superset of $p(R)$ by using the conservative analytical bounds developed in Chapter 3 to test satisfaction of possible NOE assignments. We say that our method is *complete* because ev-

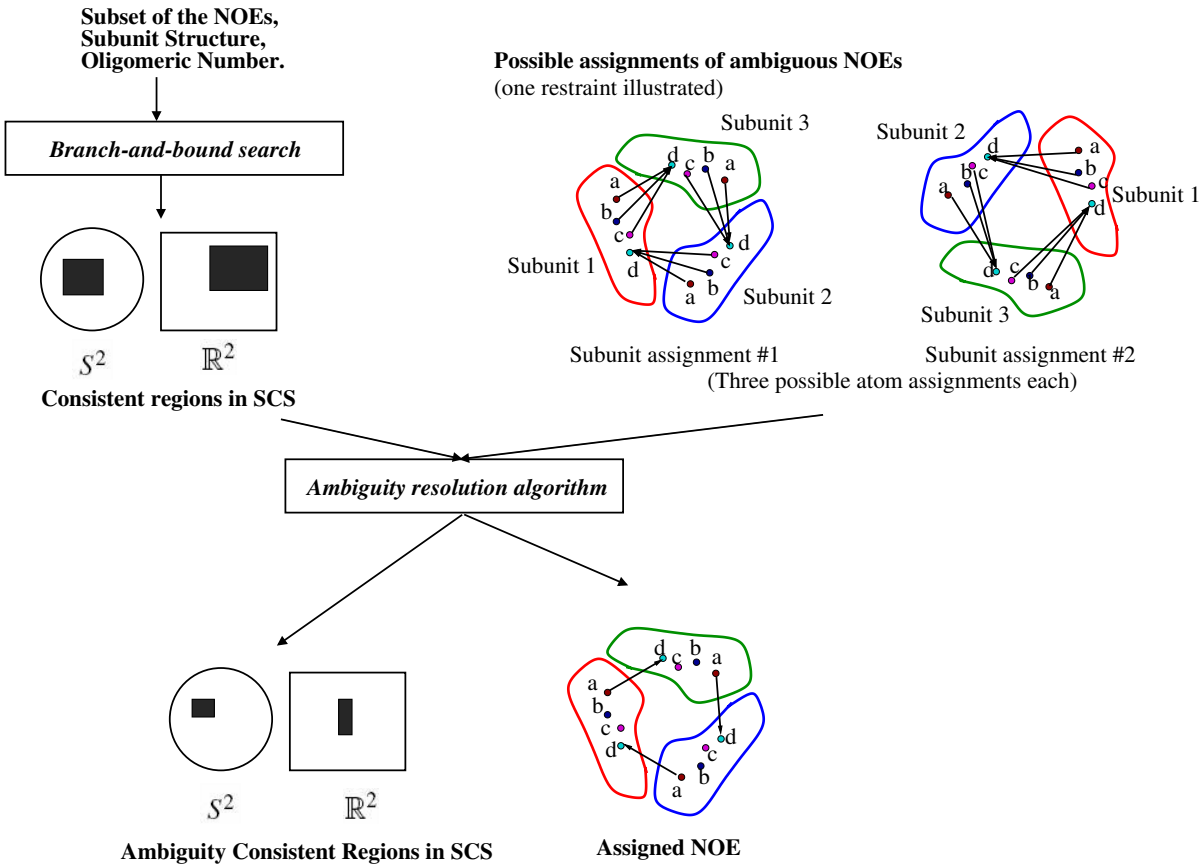


Fig. 4.1: Resolving ambiguous NOEs in structure determination of symmetric homo-oligomers. Different possible assignments (arising from subunit and atom ambiguity) for one ambiguous NOE are illustrated for a trimer. The NOE could be ordered between subunits as 1-2-3 or 1-3-2, and has chemical shift degeneracy between atoms a, b, and c. *Phase 1*: Consistent regions in the symmetry configuration space (SCS), the space of symmetry axis parameters, are obtained using our previously developed branch-and-bound algorithm (Chapter 3), taking a subset of available NOEs (typically atom-unambiguous NOEs if any) as input. The 4-dimensional SCS is cartoon-represented as two 2D regions, a sphere representing the orientation space S^2 and a rectangle representing the translation space $\mathbb{R}^2$. *Phase 2:* Our ambiguity resolution algorithm takes as input the consistent regions output from our branch-and-bound algorithm and possible assignments of the ambiguous NOEs to determine a mutually-consistent, complete set of NOE assignments and ambiguity consistent regions (ACR) in SCS.

ery element of $p(R)$ is provably guaranteed to be in the superset returned by our algorithm. More specifically, our method takes as input an initial region of the SCS (represented as a set of non-overlapping cells, each of which is a hyper-cuboid region in $S^2 \times \mathbb{R}^2$, in a hierarchical decomposition), and outputs the intersection between that region and a superset of $p(R)$. In this way, it is possible to sequentially compute a superset of $p(R)$ by sequentially "intersecting" more and more (possibly ambiguous) NOEs.

Our approach to identifying the mutually-consistent, complete set of NOE assignments and SCS regions is summarized in Figure 4.1. Our algorithm has two phases. In the first phase, branch-and-bound, we perform a complete search of the SCS, as in Section 3.1.1, using a subset of the NOEs (typically atom-unambiguous NOEs). (We note that subunit ambiguity *always* exists for oligomers with more than two subunits.) This phase identifies the consistent regions (a subset of the SCS), which contain symmetry axes defining all complexes consistent with the chosen subset of the NOEs. In the second phase, ambiguity resolution, we apply our ambiguity resolution algorithm (Figure 4.2) that considers each of the remaining ambiguous NOEs sequentially, that is, one after the other. At each step, we use our *ordering criterion* (next section) to choose the NOE considered most informative. Each SCS cell of the consistent regions is checked, using our conservative bound, to ensure that some assignment of the NOE is consistent with the structures represented by the cell. A cell is eliminated if every assignment of any of the NOEs is violated by all the conformations represented by that cell. We then continue with the next NOE chosen using our ordering criterion and the remaining cells. This process of choosing, updating, and resolving is continued until all the NOEs have been considered. The remaining cells

Input:
R: Set of NOEs
$U \subset R$: Subset of the NOEs (typically atom-unambiguous NOEs)
C: Set of cells from branch-and-bound with U
N:$R \to \mathbb{Z}_{n-1} \times A$: Possible assignments for each NOE
due to subunit ambiguity (subunit offset 1 to $n-1$)
and atom ambiguity (proton in the set A).

Output:
$C' \subset C$: Remaining cells
$N' \subset N$: Remaining assignments

Algorithm:
initialize $R' \leftarrow R - U,\ C' \leftarrow C$
while $R' \neq \emptyset$
// Determine the next NOE to assign according to our criterion
$S \leftarrow$ representative structures for C'
$q \leftarrow \arg\max_{r \in R'} \text{average}_{a \in N(r)} |\{s \in S \mid r \text{ under assignment } a \text{ is violated in } s\}|$ (*)
$R' \leftarrow R' - \{q\}$

// Remove inconsistent cells
$C' \leftarrow \{c \in C' \mid \exists a \in N(q)$ s.t. c is consistent with q under assignment $a\}$
end while

// Remove inconsistent assignments
$N' \leftarrow \{(r, a) \mid r \in R, a \in N(r), \exists c \in C'$ s.t. c is consistent with r under assignment $a\}$

return N', C'

Fig. 4.2: Ambiguity resolution algorithm.

form the *ambiguity-consistent regions* (ACR) and contain symmetry axes defining all complexes consistent with some assignment for each NOE. Finally, we resolve the ambiguity in the NOEs by eliminating assignments inconsistent with the ACR. As we formally defined above, this process allows us to label assignments as inconsistent (violated everywhere in the ambiguity-consistent regions) or redundant (does not provide any additional constraint for the regions). Our ambiguity resolution algorithm hence identifies every mutually consistent set of assignments and ambiguity-consistent regions.

We represent both consistent regions and ambiguity-consistent regions with cells in a hierarchical decomposition of the symmetry configuration space. Thus the volume of these cells allows us to characterize the progress in structure determination, as the algorithm focuses on consistent symmetry axes. We generate representative conformations from these cells such that any structure in any cell is within τ_0 Å RMSD (the user-defined similarity level as before) of at least one representative structure. As in Chapter 3, we evaluate the representatives for quality of NOE restraint satisfaction, and then energy-minimize them and evaluate them for quality of van der Waals packing, ultimately identifying the set of well-packed satisfying structures. The average variance in atomic positions in these conformations allows us to evaluate the structural precision attained by the two phases of the algorithm.

4.3 NOE Ordering Criterion

We use a particular ordering criterion to choose the next NOE in our sequential strategy. Ordering of the NOEs only affects the efficiency of our approach, and not the result.

Claim 4.1 *Re-ordering the NOEs in R does not affect the completeness of our sequential algorithm.*

Proof sketch: This is a direct consequence of the fact that set intersection is commutative, and the fact that our method returns a superset of $p(R)$ (Equation 4.1). □

However, the order does impact the efficiency and we want to identify the NOE that causes the largest decrease in the SCS, in order that fewer consistency tests are performed. Identifying such an NOE will decrease the number of satisfaction checks required for the remaining assignments, and hence decrease the overall time complexity. To find the next NOE to assign, we first generate a set of representative structures, one from each cell. For each assignment of an NOE, we compute its scores as the number of representative structures that violate the NOE under the assignment. This is indicative of the portion of the SCS that would be eliminated under the assignment. Each NOE's score is then the average of the scores of all of its possible assignments (the line marked (*) in Figure 4.2). We choose the next NOE to assign as the one with the maximum score. Although our ordering criterion does not guarantee optimal efficiency, our results show that it does very well in practice and allows for the identification of redundant NOEs. We also tried evaluating an NOE by the minimum of its assignment scores, but found this approach to be inferior (results not shown).

4.4 Computational Complexity

When there are r NOEs with a assignments each, we have a^r possible combinations of assignments. An advantage of our algorithm is that it avoids the potentially exponential enumeration of all a^r possible combinations of NOE assignments.

We first run the core branch-and-bound algorithm with a subset of the available NOEs (typically atom-unambiguous NOEs) as input. The output from the search is then input to our ambiguity resolution algorithm to ultimately return a mutually consistent set of ACR and NOE assignments.

Claim 4.2 *Given c cells as consistent regions from the branch-and-bound search with a subset of the NOEs, the time complexity of our algorithm is $O(rac)$ where r is the number of ambiguous NOEs, and a is the number of assignments for each NOE.*

Proof sketch: Let t be the time complexity for checking satisfaction of an NOE in one cell. Assigning a particular NOE requires a satisfaction test in each of the cells for each of the assignments (i.e., whether some conformation in the cell is consistent with the assigned NOE) and thus a total time of $O(atc)$ in the worst case where no assignments and no cells have been eliminated. The time complexity for a satisfaction test is constant, $t = O(1)$ (see Theorem 3.1). Therefore, assigning one NOE has a time complexity of $O(ac)$. Our algorithm considers a total of r NOEs, and in the worst case each is independent of the others (i.e., no pruning occurs), for a total time of $O(rac)$. □

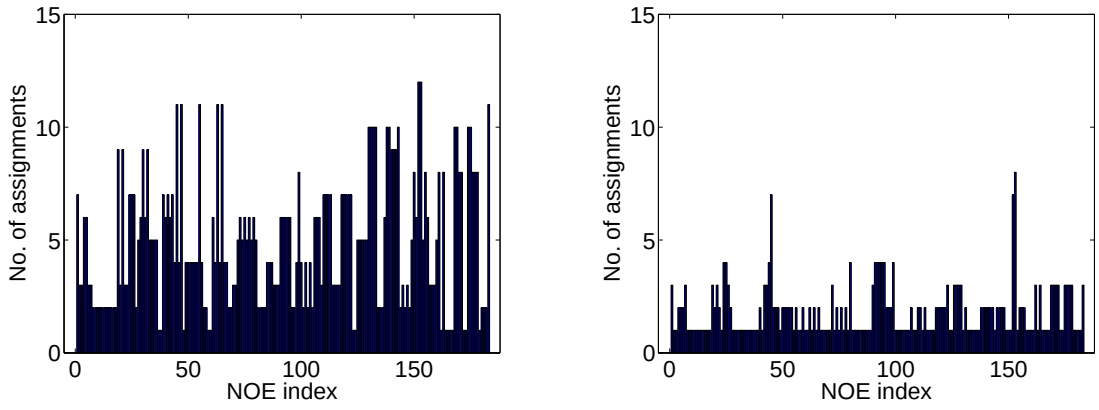
The number of cells, c, depends on the complexity of the well-packed satisfying regions in the SCS. These regions are bounded by algebraic hyper-surfaces. In principle, a worst

case bound could be obtained based on the combinatorial complexity of this 4-dimensional arrangement [42]. In practice, we have characterized c empirically for the test cases in this chapter ($c = 70$ for the homo-dimeric MinE and $c = 4006$ for the homo-trimeric CCMP with atom-unambiguous NOEs as input) and report that our algorithm is efficient in practice. In general for ambiguous NOE assignment problems, a bottleneck in combinatorial complexity lies in the potentially exponential number a^r of possible assignments, and for this reason our methodology in this chapter has focused on developing an algorithm that can provably make correct and consistent assignments while only considering a *linear* number of possible assignments, ar.

4.5 Results

4.5.1 Ambiguity Resolution and Structure Determination of Homo-dimeric MinE

The first test case is the homo-dimeric topological specificity domain of MinE. This complex has no subunit ambiguity, since subunit ambiguity is absent in homo-dimers. Atom ambiguity was simulated for each of the inter-subunit NOE restraints, according to the deposited chemical shifts, using a chemical shift match tolerance for ^{1}H of $\varepsilon = 0.04$ ppm. Our approach for simulating atom ambiguity (Section 1.2.3) resulted in atom ambiguity for 168 of the 183 inter-subunit NOEs, as illustrated in Figure 4.3(a). Note that there is ambiguity as high as 12 (i.e., 12 possible assignments) for two of the NOEs.



(a) (b)

Fig. 4.3: Ambiguity resolution for homo-dimeric MinE (1EV0). Plotted is the number of assignments for each NOE (a) after the branch-and-bound algorithm but before ambiguity resolution and (b) after ambiguity resolution.

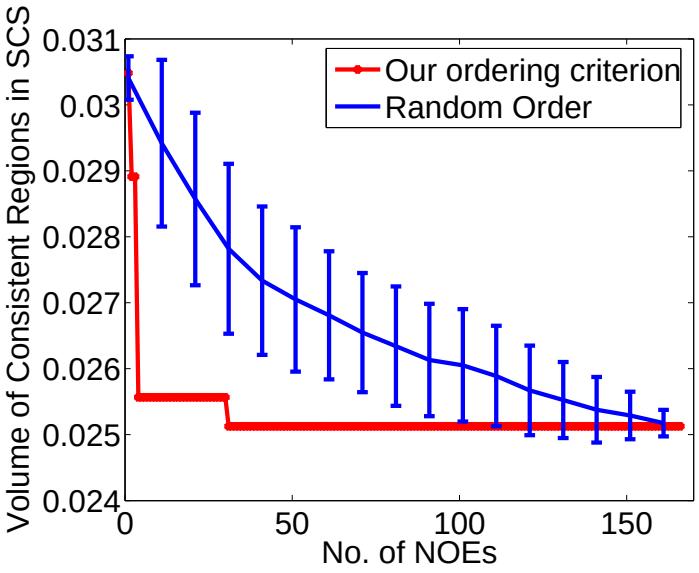


Fig. 4.4: Progress in ambiguity resolution for homo-dimeric MinE (1EV0). The NOE considered at each step is chosen either by our ordering criterion (red) or at random (blue: mean and error bars over 100 random trials). Plotted is the decrease in total volume of the SCS cells with number of NOEs chosen.

We first ran our branch-and-bound algorithm on homo-dimeric MinE for a complete SCS search with the 15 atom-unambiguous NOEs. The consistent regions output from this step included a volume of 0.030 $Å^2$ rad^2 represented by 70 cells in SCS. We then ran our ambiguity resolution algorithm. Figure 4.4 shows the decrease in the SCS volume over the number of NOEs chosen. The running time for the branch-and-bound search (Chapter 3) on a Intel Pentium Linux desktop machine with a 3.20GHz CPU was under 9 hours, while that for the ambiguity resolution algorithm was under 1 hour. The SCS volume after 30 NOEs were chosen using our ordering criterion was 0.025 $Å^2$ rad^2. No further decrease in SCS volume occurred after this, indicating that the NOEs chosen later are redundant with respect to the chosen ones. Note that about 90% of the reduction in volume occurred after three NOEs were chosen, an indication that our ordering criterion does focus on more informative NOEs. Our ambiguity resolution algorithm eliminates the inconsistent assignments and Figure 4.3(b) shows the number of possible assignments remaining for each NOE after all the inconsistent assignments have been eliminated. On comparison with Figure 4.3(a), we see that the ambiguity is considerably reduced, from an average of 5.0 assignments per NOE to an average of 1.9.

Figure 4.5 illustrates the effect of ambiguity resolution on the resulting conformations (WPS structures): the average variance is decreased from 0.34 $Å^2$ (WPS structures after branch-and-bound but before ambiguity resolution) to 0.22 $Å^2$ (after ambiguity resolution), and the average RMSD to the reference structure is reduced from 0.95 Å to 0.73 Å. There are no experimental NOEs observed between atoms in the N- and C-termini; consequently, the termini display higher uncertainty. Note that ambiguity resolution does decrease the

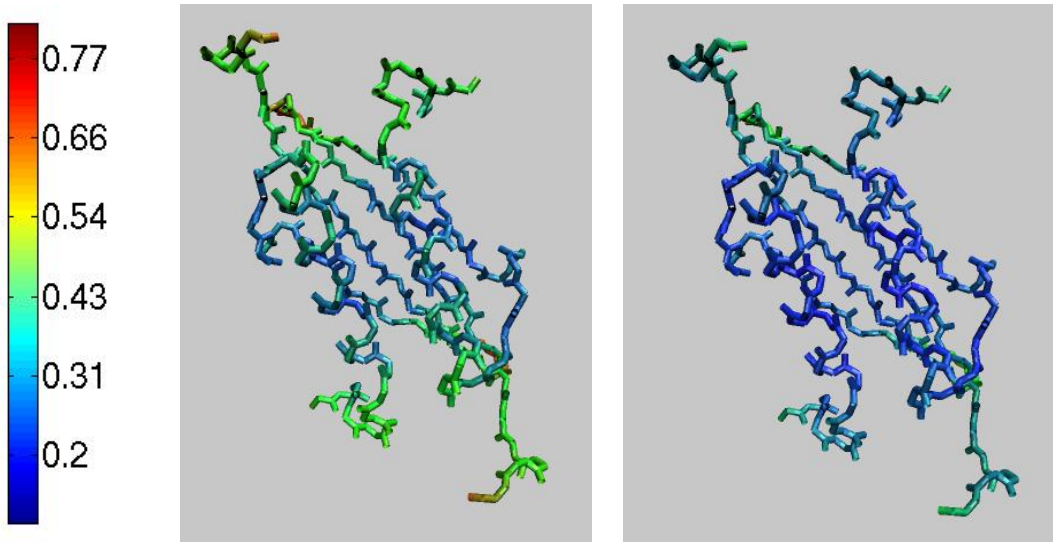


Fig. 4.5: Variance in WPS structures of homo-dimeric MinE (1EV0): (a) after the branch-and-bound algorithm but before ambiguity resolution and (b) after ambiguity resolution. Each backbone atom is colored by the variance in the position of the atom according to the shown color scale.

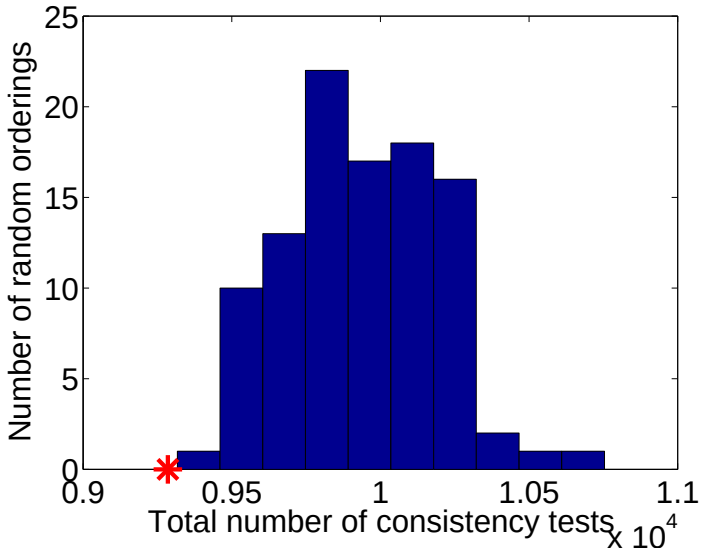


Fig. 4.6: Histogram over 100 random trials of the computational cost of ambiguity resolution for homo-dimeric MinE (1EV0). The red star indicates the number of consistency tests made using our ordering criterion.

uncertainty in these regions, by way of restraint from NOEs involving other atoms. While the assignment is not unique (Figure 4.3(b) shows NOEs with as many as eight possible assignments remaining), the remaining assignments are consistent with the determined set of WPS structures and their effect is minimal.

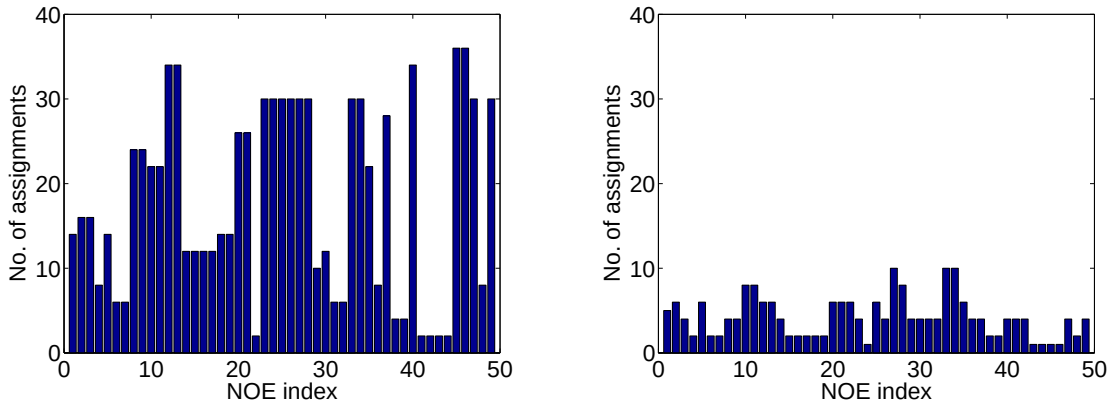
Our algorithm employs a particular criterion to determine the order in which to consider NOEs. We showed that the order does not affect the results, but only the efficiency in obtaining them (Claim 4.1). To evaluate the efficiency of our ordering criterion, we compared it against a strategy in which the order of NOEs is chosen randomly. Figure 4.4 shows that our criterion prunes the SCS much faster than random ordering. Our ordering criterion has

fully constrained the SCS volume with the first 30 NOEs, earlier than all but one of 100 random runs. We showed (Claim 4.2) that the time spent for each NOE is linear in the number of cells tested for consistency. Our ordering criterion prunes inconsistent cells much earlier than random ordering, thereby allowing us to spend time on the regions most likely to be part of the ACR. Figure 4.6 shows a histogram of the total number of consistency tests for the 100 random orderings. The number of consistency tests using our ambiguity resolution algorithm is also shown. The figure illustrates that our ordering criterion requires many fewer consistency tests than most of the 100 random trails, and is thus much faster.

4.5.2 Ambiguity Resolution and Structure Determination of Homo-trimeric CCMP

The second test case, the trimeric coiled-coil domain of chicken cartilage matrix protein (CCMP), displays subunit ambiguity in addition to atom ambiguity. The subunit ambiguity was previously resolved by manual identification of a self-consistent subunit identity for all 49 inter-subunit restraints. We ignored the subunit assignment of the authors [148], and simulated atom ambiguity as in the homo-dimer case (with $\varepsilon = 0.04$ ppm). Figure 4.7(a) illustrates the number of assignments that arise from subunit and atom ambiguity for each of the 49 inter-subunit NOEs. Two of the NOEs have as many as 36 possible assignments, when considering the combination of subunit and atom ambiguity.

Complete SCS search was performed for the homo-trimeric CCMP with the five NOEs that had no atom ambiguity. The search identified a volume of 0.27 Å^2 rad^2 represented by



(a) (b)

Fig. 4.7: Ambiguity resolution for homo-trimeric CCMP (1AQ5). Plotted is the number of assignments for each NOE (a) after the branch-and-bound algorithm but before ambiguity resolution and (b) after ambiguity resolution.

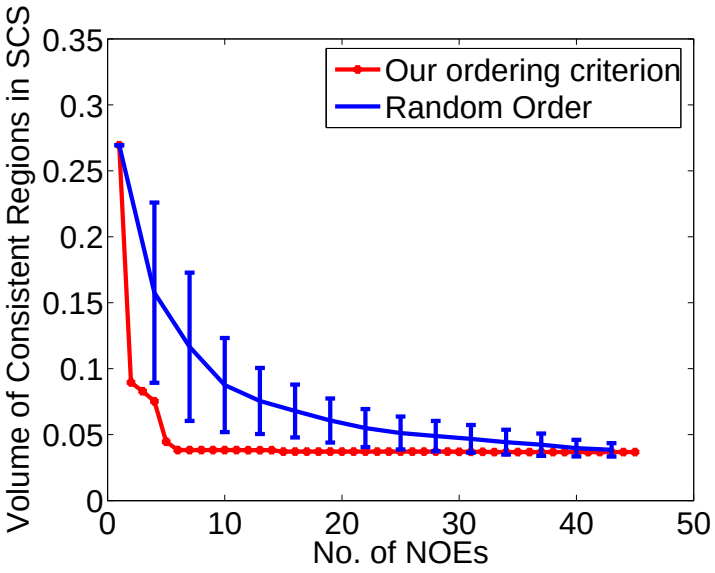


Fig. 4.8: Progress in ambiguity resolution for homo-trimeric CCMP (1AQ5). The NOE considered at each step is chosen either by our ordering criterion (red) or at random (blue: mean and error bars over 100 random trials). Plotted is the decrease in total volume of the SCS cells with number of NOEs.

4006 cells in SCS. Note that the presence of subunit ambiguity caused a larger set of cells in comparison to the dimer case. We then ran our ambiguity resolution algorithm. Our ambiguity resolution algorithm reduced the volume to 0.037 $Å^2$ rad^2. The running time for the branch-and-bound search on a Intel Pentium Linux desktop machine with a 3.20GHz CPU was around a couple of days, while that for the ambiguity resolution algorithm was around 7 hours. As Figure 4.8 shows, no further decrease in SCS volume occurred after 32 NOEs were chosen, with the bulk of the reduction (99%) occurring after 5 NOEs were chosen. The flat regions in Figure 4.8 for number of chosen NOEs 5–10 and 14–28 indicate that the NOEs considered then were redundant in the presence of the already handled NOEs. Our ambiguity resolution algorithm eliminated all the inconsistent assignments and Figure 4.7(b) illustrates the surviving assignments, significantly reduced (average of 18.2 to average of 4.3) from Figure 4.7(a).

Figure 4.9 illustrates the impact of ambiguity resolution on WPS structures. Even though many NOEs have not been resolved to a single assignment, the final set of WPS structures has high precision: the average variance is decreased from 1.92 $Å^2$ (after the branch-and-bound but before ambiguity resolution) to 0.54 $Å^2$ (after ambiguity resolution), and the average backbone RMSD to the reference structure is reduced from 1.82 Å to 0.6 Å. As Figure 4.9(b) shows, our ambiguity resolution algorithm reduced the uncertainty in the positions of all the atoms to less than 1 $Å^2$. Our algorithm reduced the high uncertainty that was present at the N- and C-termini of the chains after the branch-and-bound algorithm to below 1 $Å^2$.

Figure 4.8 compares the number of consistency tests under our ordering criterion vs.

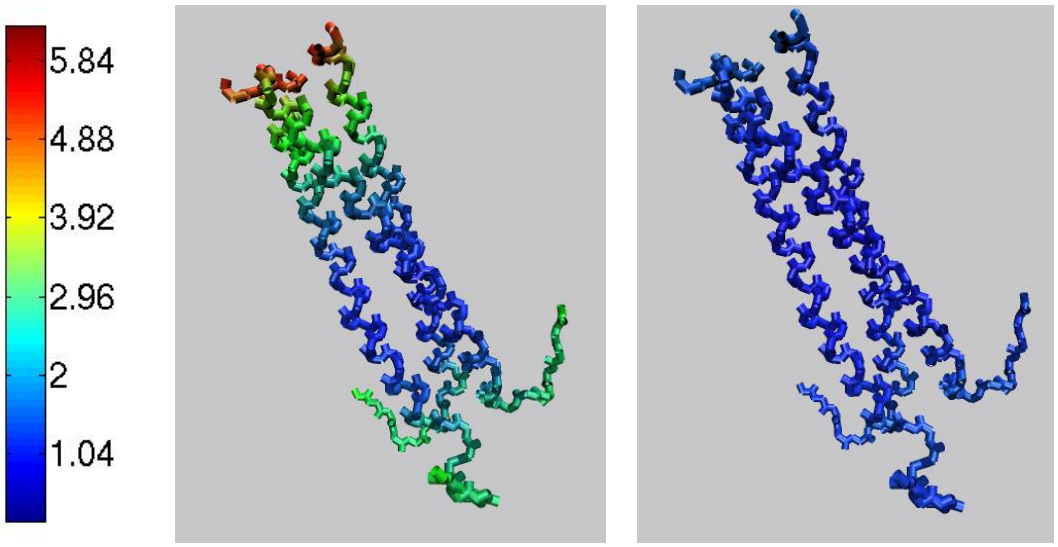


Fig. 4.9: Variance in WPS structures of homo-trimeric CCMP (1AQ5): (a) after the branch-and-bound algorithm but before ambiguity resolution and (b) after ambiguity resolution. Each backbone atom is colored by the variance in the position of the atom according to the shown color scale.

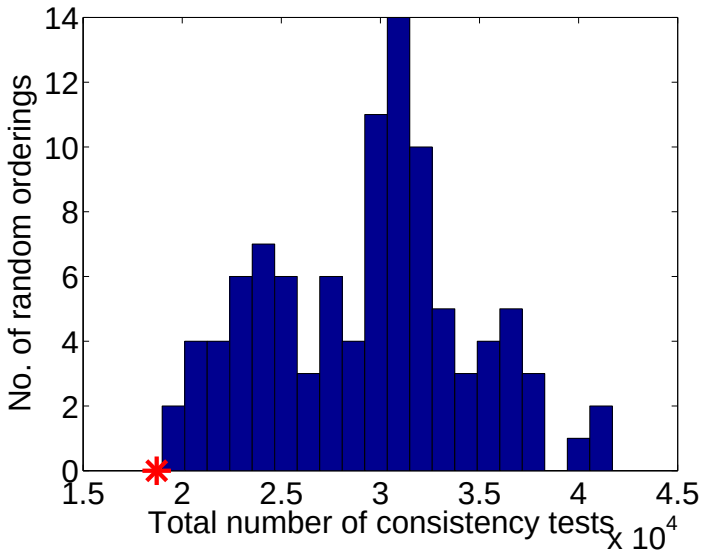


Fig. 4.10: Histogram over 100 random trials of the computational cost of ambiguity resolution for homo-trimeric CCMP (1AQ5). The red star indicates the number of consistency tests made using our ordering criterion.

under a random choice, and Figure 4.10 shows the distribution of number of cells checked for consistency. As with the homo-dimer, the ordering criterion leads to significantly less work than most of the random trails.

4.5.3 Effect of Maximum Ambiguity Level in Branch-and-Bound Input

The results so far have used only the atom-unambiguous NOEs as input to the branch-and-bound phase of our algorithm. Here we study the use of NOEs with varying amounts of ambiguity for that phase. Note that using any different set of NOEs as input to the branch-and-bound phase will always ultimately lead to the same results. It affects efficiency, not correctness. We define the set of NOEs with *maximum ambiguity level k* to include all NOEs having k or fewer possible assignments. We would expect that, with increasing maximum ambiguity level, more work would be done in the branch-and-bound search (since it has to deal with more ambiguity) and less work would be done in the ambiguity resolution algorithm (since the branch-and-bound search has eliminated some ambiguity).

Figure 4.11 shows that this is indeed the case for MinE, measuring the amount of work as the number of consistency tests performed. The figure also shows that the extra work in the branch-and-bound search with increased maximum ambiguity level is not worth it—the number of tests summed across both phases increases with the maximum ambiguity level. When all NOEs are used as input to the branch-and-bound search, the total number of consistency tests required is 7.8 times more than the number required when only atom-unambiguous NOEs are used. It is more efficient to start with only the atom-unambiguous

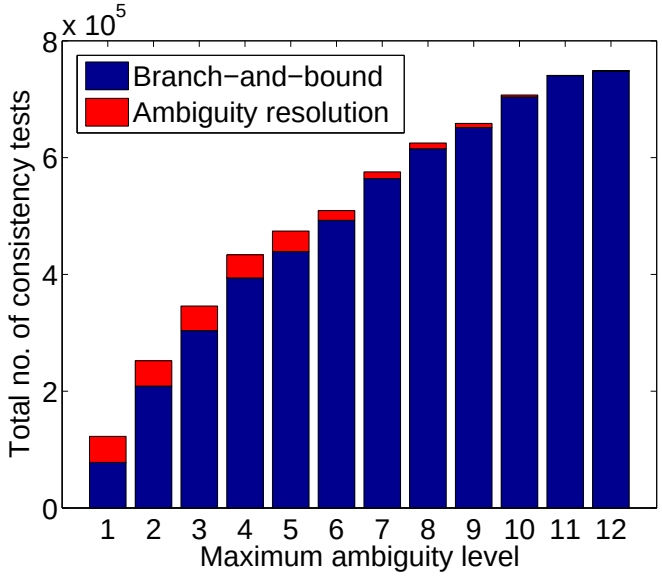


Fig. 4.11: Effect of the amount of ambiguity in the branch-and-bound search, for MinE (1EVO). The maximum ambiguity level limits the NOEs used in the search phase. The plot shows the total number of consistency tests during the branch-and-bound search and during the ambiguity resolution algorithm.

NOEs.

Our approach is applicable even when no atom-unambiguous NOEs are available. To demonstrate this, we ignored the 15 atom-unambiguous NOEs for MinE, and ran the branch-and-bound search with the 34 (out of 168) NOEs with maximum ambiguity level 2. The output from this step was an SCS volume of 0.042 $Å^2$ rad^2 represented by 94 cells. We then ran our ambiguity resolution algorithm with the remaining 134 NOEs. This reduced the SCS volume to 0.034 $Å^2$ rad^2. The ambiguity in the 168 NOEs was reduced from an average of 5.3 assignments per NOE to an average of 2.0. The average variance in WPS structures was decreased from 0.37 $Å^2$ (after branch-and-bound but before ambiguity resolution) to 0.22 $Å^2$ (after ambiguity resolution). The final structural precision after the ambiguity resolution algorithm, 0.22 $Å^2$ is the same as when we started with the 15 atom-unambiguous NOEs. Thus there was enough redundant information in the atom-ambiguous restraints to compensate for the loss of the 15 atom-unambiguous ones.

4.5.4 Effect of Spurious NOEs

So far, we have assumed that each NOE peak is true; that is, that every NOE has at least one correct assignment. However, our approach is readily extended to handle a small number of spurious NOEs, NOEs for which all assignments are wrong. In this case, we eliminate a cell when at least a specified number (which is one in the presented algorithm) of the NOEs are violated. An NOE is identified as spurious if it is inconsistent with all the remaining cells of the ACR. Note that we may not be able to identify those spurious NOEs that are

consistent with the remaining cells of the ACR.

We tested this approach with MinE, simulating spurious NOEs from the reference structure by randomly choosing pairs of protons that were more than 10 Å apart. Atom ambiguity was simulated for each of the spurious NOE as before (with ^{1}H chemical shift match tolerance $\varepsilon = 0.04$ ppm). We performed 100 random runs for each number of spurious NOEs, from one to five.

Figure 4.12(a) illustrates the detection rate for the spurious NOEs. In the case when one spurious NOE is added, the detection rate is 93%. This means that in 93% of the cases, the spurious NOE has no effect on the remaining cells in the ACR or the remaining assignments. The sets of mutually consistent cells and assignments are the same as that with no spurious NOEs. The failure of detection in the remaining seven cases is due to the presence of an atom assignment for the spurious NOE that is consistent with the remaining conformations. The detection rate decreases to 62% as the number of spurious NOEs is increased from one to five.

However, as illustrated in Figures 4.12(b), (c) and (d), even if a spurious NOE cannot be detected, it has minimal effect on the volume in SCS, the structural precision (indicated by average variance) and the structural quality (indicated by the average backbone RMSD to the reference structure). Note that even with as many as five spurious NOEs, the mean of the average variance after the ambiguity resolution algorithm is around 0.30 Å^2 and is less than the average variance of the ACR before use of the ambiguity resolution algorithm (0.34 Å^2). This indicates that we can obtain information and improve structural precision, in spite of the presence of spurious NOEs.

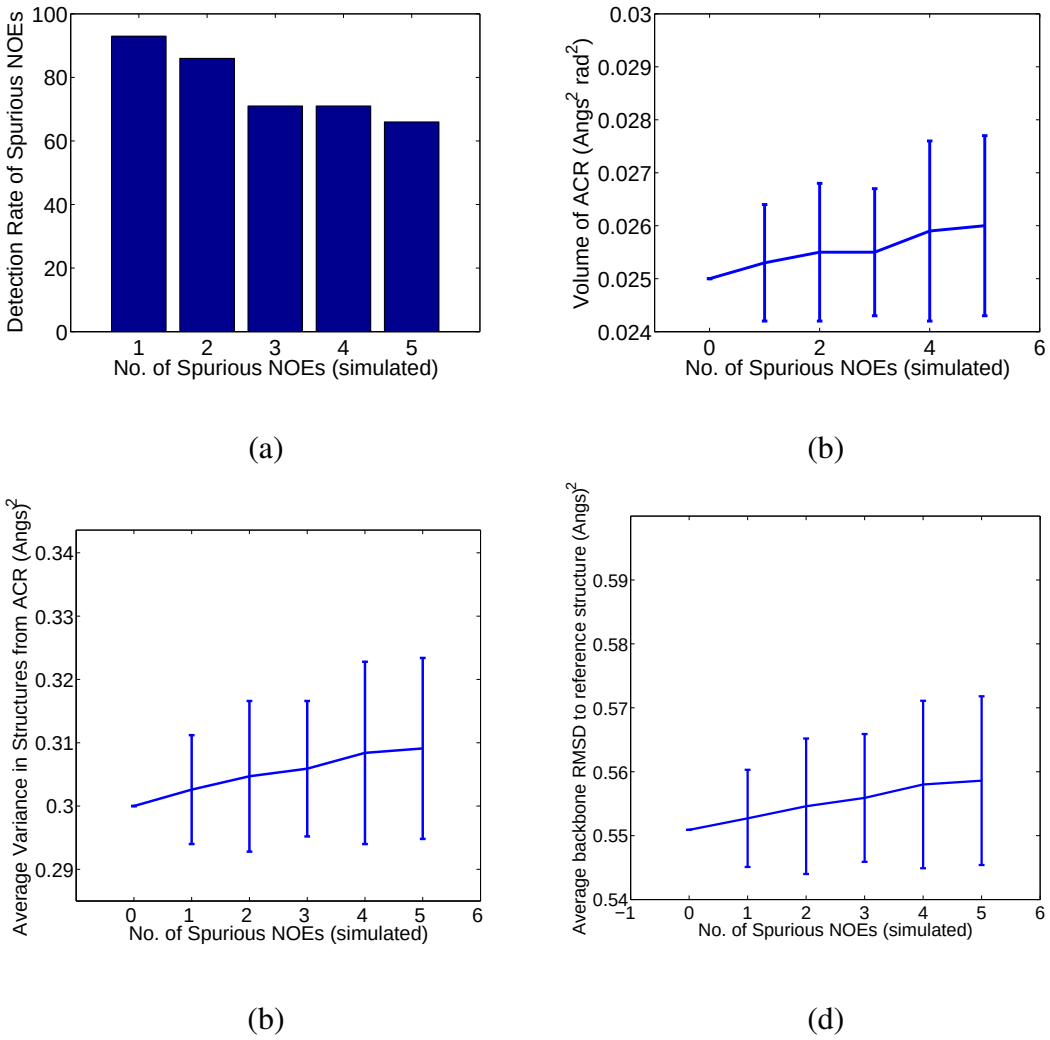


Fig. 4.12: Effect of spurious NOEs on ambiguity resolution for MinE (1EVO). The effect was studied on an increasing number of simulated spurious NOEs, with 100 random runs each. (a) The detection rate of spurious NOEs. (b) The mean and error bars for the volume of ACR. (c) The mean and error bars for the average variance of representative structures of ACR. (d) The mean and error bars for the average backbone RMSD of representative structures of ACR to the reference structure.

4.6 Conclusions

In this chapter, we have extended the core algorithm to resolve the atom and subunit ambiguity inherent in NOEs, and used the extended approach to simultaneously assign NOE peaks and determine structures of symmetric homo-oligomers. We showed from results on two test cases that our approach reduces ambiguity in NOE assignment and increases precision in structures. Using our approach, atom ambiguity was reduced by a factor of 2.6 in our first test case (MinE) and combined subunit/atom ambiguity by a factor of 4.2 in our second test case (CCMP). Incorrect NOE assignments tend to be mutually inconsistent, and therefore are eliminated by our algorithm. Using our ambiguity resolution algorithm we obtained structures with high precision: the average variance in positions of atoms is reduced to 0.22 Å^2 for MinE and 0.54 Å^2 for CCMP. Our sequential strategy to choose NOEs that appear informative allowed us to prune a significant number of bad cells (about 90%) early, and hence results in greatly reduced time complexity. Our sequential strategy also helped elucidate the information content in ambiguous NOEs—many NOEs may be redundant with other (previously handled) NOEs. We identified that 138 of the 183 NOEs in MinE and 12 of the 49 NOEs in CCMP were redundant with respect to the chosen NOEs and caused no further improvement in structural precision. We also showed that our approach can work with NOEs at different maximum ambiguity levels provided as input to the branch-and-bound search, and can handle a reasonable number of spurious NOEs.

Assignment of NOE data is a key bottleneck in structure determination by NMR and NOE assignment can be difficult to fully automate. At the same time, the structure de-

termination of symmetric proteins by NMR can be challenging. It is interesting that by *combining* these two difficult problems, we have obtained an algorithm that solves both simultaneously, and that enjoys guarantees on its completeness and complexity.

In this chapter and the previous chapter, the subunit structure was assumed to be known correctly without any error. In order to make our approach broadly applicable, we extend our approach to handle subunit side-chain uncertainty in the next chapter.

5. SIDE-CHAIN UNCERTAINTY

This chapter extends the core algorithm to handle uncertainty in the side-chain conformations of the subunit. Side-chain uncertainty could arise either from insufficiency of the data in accurately determining the side-chain conformations or from flexibility in the side-chains. In the previous chapters, no form of side-chain uncertainty was considered during the search through the SCS. Uncertainty in the side-chains was incorporated only during the energy-minimization stage and energy minimization was used only to *refine* the existing side-chain conformations. The heuristic techniques used for energy minimization might miss sampling low-energy side-chain conformations that require crossing high energy barriers. This leads to energy minimization being useful only for fine-tuning existing side-chain conformations. In this chapter, we develop an approach that handles side-chain uncertainty during the search through SCS and that allows for sampling of low-energy conformations across high-energy barriers. Since we incorporate side-chain uncertainty during the search, our approach is also applicable to cases where we know the subunit structure only in its apo form and no change in backbone occurs between the apo and holo forms.

We handle uncertainty in side-chains by discretizing the space of side-chain conformations using rotamers (Section 2.3) and extending the configuration space accordingly.

Rotamers provide "classes" of side-chain conformations, that is, a sampling of side-chain conformations across high-energy barriers, which can later be fine-tuned by energy minimization. We incorporate side-chain uncertainty during the search through the SCS by extending the configuration space to include rotamers. We handle side-chain uncertainty *simultaneously* with "docking" uncertainty, that is, the uncertainty in the docked position of the backbone of the adjacent subunit with respect to the fixed subunit. This is in contrast to existing docking-based approaches which follow a two-stage approach of identifying backbone conformations by rigid docking before considering side-chain flexibility (see Section 2.3). By considering both side-chain and docking uncertainty simultaneously we take into account the dependency between the backbone and the side-chains.

We maintain the guarantee of completeness, identifying all conformations that are within a user-defined similarity level (backbone RMSD) to native structures even when there is uncertainty in subunit side-chain conformations. We also guarantee that all rotamers that are part of valid conformations are identified. In symmetric membrane proteins accurate determination of side-chain conformations is especially crucial for analysis of transmembrane helical packing [101, 119]. By being complete, as before, we avoid false precision in the set of determined structures and are able to accurately quantify the structural constraint from the data. We are also able to determine all pair-wise interactions between residues on adjacent subunits. Such interactions suggest candidates for single or double mutant experiments to better understand the function of the complex.

5.1 Algorithm

Given the subunit structure of a symmetric homo-oligomer, we augment the existing side-chain conformations of each residue with the several rotameric conformations from the rotamer library. (Note that our approach is applicable to any rotamer library and we currently use the penultimate rotamer library of Lovell et al. [100].) Thus, the four-dimensional symmetry configuration space (SCS), is extended to form the *symmetry configuration space with rotamers (SCSR)*, $S^2 \times \mathbb{R}^2 \times R_1 \times R_2 \times \ldots \times R_n$, where $R_1, R_2, \ldots, R_n$ represent the set of rotamers considered for residues $1, 2, \ldots, n$.

We search through the SCSR by hierarchically subdividing the SCS, eliminating regions of the configuration space that provably do not represent valid conformations. We apply a set of pruning operators (discussed below) to regions of the configuration space to prune rotamers that either violate restraints, or lead to steric clashes (that is, "significantly" high vdW energies). We say a restraint is violated if all pairs of positions (arising from rotamers and symmetry axes) of the corresponding atoms violate the restraint. Pruning all possible rotamers for any residue implies that no valid conformation exists in the cell; hence the configuration region is eliminated.

At the end of the hierarchical subdivision, we have a set of leaves that could have at least one valid conformation (of backbone and side-chains), that is, a conformation that satisfies all the restraints and has no steric clashes. We choose representative structures from the centers of the leaf cells. Each representative structure has sets of possible side-chain conformations for each of the residues, arising from the remaining rotamers. Due

to the conservative nature of our bounds, we could have kept rotamers that cannot be part of valid conformations. So, we test rotamers for restraint satisfaction and steric clashes in each of the representative structures, pruning rotamers that cannot be part of at least one valid conformation. On each of the representative structures, we also apply the DEE criterion developed by Desmet et al. [37] to eliminate rotamers that provably cannot lead to energies within a specified threshold of the lowest possible energy. We currently use the Goldstein DEE criterion [56], but can readily use any other DEE criterion. Note that elimination of rotamers could lead to elimination of some of the representative structures. We refer to the set of remaining representative structures as satisfying structures. We then apply an A* search [96] to choose the rotamers that give the lowest energy for the backbone of each satisfying structure. We finally obtain, as in previous chapters, a set of well-packed satisfying structures (WPS) by clustering the satisfying structures according to backbone RMSD, and subjecting each of the clustered structures to energy minimization.

In addition to generating conformations, we use our approach to obtain an *interface map* indicating possible pairwise interactions between residues on adjacent subunits. These interactions could then be used to plan mutagenesis experiments that would provide insights into the function of the complex. Given a set of structures, we obtain the interface map as follows. We say that a pair of residues is in contact when any of their atoms are within a specified distance of each other. For each pair of residues on adjacent subunits, we identify the frequency that the pair is in contact in the given set of structures. The interface map is a matrix of rows and columns with each block indicating the contact frequency for the corresponding residues on adjacent subunits.

Restraint-based pruning operator:

Input:

$K \times R_{K1} \times R_{K2} \times \ldots \times R_{Kn}$, where $K \subset S^2 \times \mathbb{R}^2$ and R_{Ki} is the set of rotamers allowed for residue i in cell K

D: Set of inter-subunit distance restraints

Method:

do

for each $\|p - q'\| \leq d_{\text{NOE}} \in D$, where $p \in$ residue i of fixed subunit and $q' \in$ residue j of adjacent subunit (q' corresponds to q in the fixed subunit)

$R_{Ki} \leftarrow \{r \in R_{Ki} \mid \exists s \in R_{Kj}$ where $\forall\, z \in K, \|r(p) - \mathcal{T}_z(s(q))\| \leq d_{\text{NOE}}\}$

$R_{Kj} \leftarrow \{s \in R_{Kj} \mid \exists r \in R_{Ki}$ where $\forall\, z \in K, \|r(p) - \mathcal{T}_z(s(q))\| \leq d_{\text{NOE}}\}$

while (any R_K is updated)

Backbone steric-based pruning operator:

Input:

$K \times R_{K1} \times R_{K2} \times \ldots \times R_{Kn}$, where $K \subset S^2 \times \mathbb{R}^2$ and R_{Ki} is the set of rotamers allowed for residue i in cell K

B: Set of backbone atoms of the fixed subunit

Method:

for each residue i in the fixed subunit

$R_{Ki} \leftarrow \{r \in R_{Ki} \mid \forall$ atom $a \in$ residue $i,\ \forall$ atom $b \in B,\ \exists z \in K$ such that $\|r(a) - \mathcal{T}_z(b)\| > d_{\text{steric}}\}$

for each residue j in the adjacent subunit

$R_{Kj} \leftarrow \{s \in R_{Kj} \mid \forall$ atom $b \in$ residue $j,\ \forall$ atom $a \in B,\ \exists z \in K$ such that $\|a - \mathcal{T}_z(s(b))\| > d_{\text{steric}}\}$

Fig. 5.1: Description of pruning operators

Side-chain steric-based pruning operator:
Input:
$K \times R_{K1} \times R_{K2} \times \ldots \times R_{Kn}$, where $K \subset S^2 \times \mathbb{R}^2$ and R_{Ki} is the set of rotamers allowed for residue i in cell K
Method:
do
for each residue i in the fixed subunit
$R_{Ki} \leftarrow \{r \in R_{Ki} \mid \forall$ residue j in the adjacent subunit, $\forall s \in R_{Kj}$, $\forall$ atom $b \in$ residue j, $\forall$ atom $a \in$ residue i, $\exists z \in K$ such that $\|r(a) - \mathcal{T}_z(s(b))\| > d_{\text{steric}}\}$
for each residue j in the adjacent subunit
$R_{Kj} \leftarrow \{s \in R_{Kj} \mid \forall$ residue i in the fixed subunit, $\forall r \in R_{Ki}$, $\forall$ atom $a \in$ residue i, $\forall$ atom $b \in$ residue j, $\exists z \in K$ such that $\|r(a) - \mathcal{T}_z(s(b))\| > d_{\text{steric}}\}$
while (any R_K is updated)

Fig. 5.2: Description of pruning operators

5.1.1 Description of the Pruning Operators

Figures 5.1 and 5.2 describe the pruning operators. Here, q is the identity of the atom, $s(q)$ is the position, in the fixed subunit, of atom q with respect to rotamer s, and $\mathcal{T}_z(s(q))$ is the position of atom q in the adjacent subunit with respect to rotamer s under the symmetry axis z. d_{NOE} denotes the maximum distance between an NOE-restrained pair of atoms and d_{steric} denotes a minimum steric allowed distance.

Restraint-based pruning operator: The restraint-based operator eliminates rotamers that violate at least one of the inter-subunit NOEs in all the conformations represented by the cell. Consider a restraint of the form $\|p - q'\| \leq d_{\text{NOE}}$ where p is an atom of residue i on the fixed subunit and q' is an atom of residue j on the adjacent subunit. The set of possible positions for p is determined by the remaining rotamers in the cell K, $r \in R_{Ki}$. For q', the set of all possible positions is obtained as the union of our geometric bound

over remaining rotamers $s \in R_{Kj}$. The operator eliminates a rotamer $r \in R_{Ki}$ when $r(p)$ violates the restraint with all possible positions of q'. Similarly, the operator eliminates a rotamer $s \in R_{Kj}$ when all possible positions of $s(q)$ on the adjacent subunit violate the restraint with all possible positions of p. The operator eliminates rotamers according to each restraint individually and repeatedly loops over all restraints until no further rotamers are eliminated.

Backbone steric-based pruning operator: The backbone steric-based operator eliminates rotamers that would be involved in a steric clash with backbone atoms of another subunit for any conformation represented by a cell. The operator considers each rotamer of each residue on the adjacent subunit with respect to a given SCS cell. The possible positions of each atom on the adjacent subunit for a given rotamer s are obtained using our geometric bounds. The operator eliminates s if all positions of any of its atoms are too close to any of the backbone atoms in the fixed subunit. Similarly, each rotamer of each residue on the fixed subunit is considered and tested for steric clash with the possible positions of the backbone atoms of the adjacent subunit.

The backbone steric-based operator must test all pairs of atoms on adjacent subunits for steric clash twice: once for eliminating rotamers on the fixed subunit and once for eliminating rotamers on the adjacent subunit. In order to avoid all atom pairwise tests between atoms on the adjacent subunit and backbone atoms of the fixed subunit, we pre-calculate the region occupied by the backbone of the fixed subunit. This region, the *subunit map* M, is a union of balls of radius equal to the steric clash threshold. The operator then prunes a rotamer if the geometric bound on positions of any of its atoms on the adjacent subunit lies

entirely within M. Obtaining such a subunit map for the backbone of the adjacent subunit is not possible, since the map changes depending on the cell under consideration. Due to the cost involved in explicitly testing all pairs, in practice we do not perform these tests during the hierarchical subdivision but perform them only for the representative structures at the leaves.

Side-chain steric-based pruning operator: The side-chain steric-based operator eliminates rotamers that would be involved in a steric clash with side-chain atoms of the subunit for any conformation in a cell. The operator independently considers each rotamer of each residue in the adjacent subunit. As in the case of the backbone steric-based operator, the set of all possible positions of each atom on the adjacent subunit for a given rotamer s is obtained using our geometric bound. The operator eliminates s if all possible positions of any of its atoms are too close to all rotameric positions of any of the side-chain atoms in the fixed subunit. Similarly, each rotamer of each residue on the fixed subunit is considered and tested for steric clash with the possible positions of the atoms of the adjacent subunit.

Note that the side-chain operator requires tests with all possible positions (arising from different rotamers) of side-chain atoms in both the subunits and is hence more expensive than the backbone steric-based operator. As we did for the backbone operator, in order to avoid all pairwise tests, we pre-calculate the occupied regions, here by the rotamers of each of the residues of the fixed subunit. These regions, formed by the union of balls of radius equal to the steric clash threshold, are referred to as *rotamer maps*. We denote the rotamer map of rotamer $r \in R_{Ki}$ by Q_{Kir}. The operator then prunes a rotamer if the bound on any of its atoms lies entirely within the Q_{Kir} of any residue i, $\forall r \in R_{Ki}$. As in the case of

the backbone operator, obtaining such rotamer maps for atoms on the adjacent subunit is not possible. So, tests to eliminate rotamers on the fixed subunit that are involved in steric clashes with the side-chain atoms of the adjacent subunit are performed only at the leaf level.

5.1.2 Correctness of the Operators

In this section we show the correctness of the operators in eliminating invalid rotamers and keeping valid rotamers. For this, we first define what we mean by invalid rotamers, then we define what we mean by blockages and show that when blockages are absent and the bounds are perfect, the operators are guaranteed to eliminate all invalid rotamers.

Let $\mathcal{O}$ represent the operator under consideration. When $\mathcal{O}$ represents the restraint-based operator, we seek to enforce the constraint that a rotamer satisfy all the distance restraints. For the backbone-based steric operator, the constraint is that the rotamer does not have a steric clash with the backbone of another subunit. For the side-chain steric-based operator, the rotamer must not have a steric clash with the side-chains of another subunit. We say a rotamer r is an $\mathcal{O}$-*valid* rotamer in a given cell K if it is the side-chain conformation of at least one backbone conformation represented in K that satisfies the constraint specific to $\mathcal{O}$. Every rotamer that is not $\mathcal{O}$-valid is called an $\mathcal{O}$-*invalid* rotamer. We say a cell K is an $\mathcal{O}$-*invalid cell* when all rotamers of some residue are $\mathcal{O}$-invalid rotamers.

Let us consider a graph for each operator $\mathcal{O}$ in each cell, with vertices for the residues

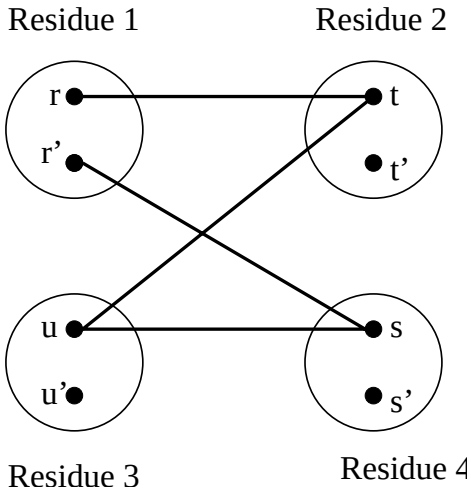


Fig. 5.3: Blockages between rotamers.

on the fixed and an adjacent subunit in the protein complex. In a given cell of the configuration space, let each vertex have sub-vertices corresponding to the $\mathcal{O}$-valid rotamers of the residues in that cell. An edge exists between a sub-vertex a of vertex A on the fixed subunit and sub-vertex b of vertex B on the adjacent subunit when a and b are consistent with each other with respect to the operator $\mathcal{O}$, and can be part of the same valid conformation. Existence of a simple path between two rotamers implies that there exists a valid conformation involving all the rotamers in the path. Let us denote a simple path between two rotamers r and r' as $r \sim r'$. We say that we have a *blockage* (Figure 5.3) when given two sub-vertices r and r' of a vertex, there exists a simple path $r \sim r'$ but not $r \sim r$.

Claim 5.1 *In the absence of blockages and assuming that the bounds are perfect, the restraint-based operator will keep a rotamer in a given cell iff it is $\mathcal{O}$-valid in the cell.*

Proof: *only if:* Let us consider an $\mathcal{O}$-valid rotamer, r. From the definition of an $\mathcal{O}$-valid rotamer, for each constraint that r is involved in, there exists a rotamer on the adjacent

subunit with which r satisfies the constraint. Then, it follows from the definition of the operator represented by $\mathcal{O}$ that r is not eliminated by the operator.

if: (Proof by induction). We say a rotamer r is k-valid when it is $\mathcal{O}$-valid for the first k constraints. (Recall that the constraints vary depending on the operator that $\mathcal{O}$ represents.) Let us consider the base case, $k = 1$, where the constraint involves residues i and j. The operator $\mathcal{O}$ keeps r only when there exists a rotamer $t \in R_j$ with which r satisfies the restraint. Since only one constraint exists, r and t are part of a valid conformation making them $\mathcal{O}$-valid rotamers.

For the inductive hypothesis, let the rotamer r be k-valid. To complete the proof, we need to prove that on adding the $(k+1)^{\text{th}}$ constraint, the operator $\mathcal{O}$ still only keeps rotamers that are $\mathcal{O}$-valid. Let us first consider the case where the $(k+1)^{\text{th}}$ constraint is independent of the existing k constraints, that is, it does not involve any of the residues in the k constraints. Since the operator $\mathcal{O}$ checks each constraint independently, it follows from a similar argument as in the base case that the operator keeps only $\mathcal{O}$-valid rotamers.

Let us now consider the other case where the $(k+1)^{\text{th}}$ constraint involves residues in the previous k constraints (recall that we require that it cannot lead to blockages). Let the $(k+1)^{\text{th}}$ constraint be between residues i and j. Note that from the inductive hypothesis it follows that i and j have only $\mathcal{O}$-valid rotamers. The operator $\mathcal{O}$ now keeps a rotamer $r \in R_i$ only when there exists an $\mathcal{O}$-valid rotamer, $t \in R_j$, with which r satisfies the $(k+1)^{\text{th}}$ constraint. If r and t were part of at least one valid conformation before the $(k+1)^{\text{th}}$ constraint was considered, then they continue to be $\mathcal{O}$-valid rotamers. On other hand, suppose that r and t did not belong to the same valid conformation in the first k

constraints (that is, a simple path $t \sim r$ does not exist). Then, since t is an $\mathcal{O}$-valid rotamer, it would be part of a valid conformation, say C', including some other rotamer, say r', of the residue to which r belongs to. Since r is kept because of t, there would exist a simple path $r \rightarrow t \sim r'$, but not $r \rightarrow t \sim r$, causing a blockage. This would contradict our assumption that no blockages exist. Thus it must be the case that the operator $\mathcal{O}$ only keeps rotamers that are $\mathcal{O}$-valid.

□

5.1.3 Cost of the Operators

Claim 5.2 *The time complexity for applying the restraint-based operator in an SCS cell is $O(n_d n_r^2)$, where n_d is the number of distance restraints and n_r is the maximum number of rotameric positions for atoms involved in a restraint.*

Proof: Testing satisfaction of a restraint between a pair of atomic coordinates takes constant time (see Claim 3.1). The satisfaction tests must be performed between all rotameric pairs for each restraint, for a total cost of $O(n_r^2)$. Testing for all the distance restraints yields a total cost of $O(n_d n_r^2)$. □

Claim 5.3 *The time complexity for applying the backbone steric-based pruning operator in an SCS cell to eliminate rotamers on the adjacent subunit, is $O(n_a n_t t_b)$, where n_a is the number of atoms in the subunit, n_t is the maximum number of rotameric positions for atoms*

in the subunit and t_b is the time complexity for testing if a rotameric coordinate of an atom is involved in a steric clash with the backbone.

Proof: The time for testing whether an atom, a, is involved in a steric clash requires (a) computing bounds on all possible positions of a and (b) testing if the bounds lie within the subunit map. Computing bounds on all possible positions of an atom takes constant time (Claim 3.1). Testing whether a bound lies within the map takes time t_b, which is the time for testing if a convex hull lies completely inside a union of balls. We perform the test for inclusion by voxelizing the subunit map and the convex hull and testing if all the voxels of the convex hull lie inside the map. This makes $t_b = O(V_b V_h)$, where V_b is dependent on the backbone atoms of the subunit and V_h is dependent on the volume of the hull which is in turn dependent on the cell under consideration. V_h in the worst case is obtained from the largest hull possible and is dependent on the input translation and orientation space considered. The steric clash test must be performed for all rotameric positions of all the atoms in the subunit making the complexity $O(n_a n_t t_b)$. □

Claim 5.4 *The time complexity for applying the side-chain steric-based pruning operator in an SCS cell to eliminate rotamers on the adjacent subunit is $O(n_a n_r n_t^2 t_s)$, where n_a is the number of atoms in the subunit, n_r is the number of residues in the subunit, n_t is the maximum number of rotameric positions for atoms in the subunit and t_s is the time complexity for testing if a rotameric coordinate of an atom is involved in a steric clash with the side-chain.*

Proof: The time for testing whether an atom, a, is involved in a steric clash requires (a)

computing bounds on all possible positions of a and (b) testing if the bounds lie within maps of all rotamers of all residues. Computing bounds on all possible positions of an atom takes constant time (Claim 3.1). Testing whether a bound lies within a single map takes time t_s, which is the time for testing if a convex hull lies completely inside a union of balls. The time t_s is similar to t_b above except that V_b is now dependent on the number of side-chain atoms forming the map. The number of maps that need to be considered is $n_r n_t$, making the time complexity for testing whether an atom is involved in a clash $n_r n_t t_s$. The steric clash test must be performed for all rotameric positions of all the atoms in the adjacent subunit, making the complexity $O(n_a n_t n_r n_t t_s) = O(n_a n_r n_t^2 t_s)$. □

5.1.4 Conservativeness of the Operators

The presence of blockages could prevent elimination of $\mathcal{O}$-invalid rotamers and hence invalid cells. Partitioning a region of the SCSR along one of the residue's set of rotamers is the only way to eliminate these invalid rotamers. So, given sets of rotamers for each residue in a SCS cell, we choose one of the residues, say i and partition the space such that the two children have the same SCS cell, but differ in the rotamers for i. Some of the rotamers of i are now in the first child and the remaining are in the second child. In practice, we find that blockages are not very common and so do not test for them during the hierarchical subdivision so as to avoid the additional cost of the test. By assuming that there are no blockages, we could accept cells that would otherwise been eliminated, causing us to be conservative in accepting cells.

Claim 5.1 states that the operators ensure that all invalid rotamers and cells are eliminated as long as we have perfect bounds. Since our bounds are conservative, we continue to subdivide the invalid cells, leading to exploration of additional nodes. The invalid rotamers and hence all sub-cells of the invalid cells are eliminated at the level that the conservative bounds get tight enough to eliminate them. Please refer to Section 3.1.1 for a discussion on the number of additional levels that need to be explored due to the conservative nature of our bound.

5.2 Results

We validated our approach on two test cases: homo-trimeric CCMP and homo-dimeric MinE. For each of the test cases, we assumed that no subunit or atom ambiguity exists and test our approach to handle side-chain uncertainty independent of subunit and atom ambiguity. The algorithms developed here and in Chapter 4 can be combined to handle both side-chain uncertainty and ambiguity in NOE data simultaneously. We used the threshold for deciding whether two C_β atoms are in contact as 8 Å [46] and that for deciding whether any pair of atoms are in contact as 4.5 Å [88].

For each test case we show that: (1) the rotamer approach avoids false precision in the set of satisfying structures and (2) applying the operators during the hierarchical subdivision is more efficient than applying them at the end of the search.

5.2.1 Results on Homo-dimeric MinE

Our first test case is the homo-dimeric topological specificity domain of *Escherichia coli* MinE [89]. The subunit structure, the set of inter-subunit distance restraints and the rotamers from the Lovell rotamer library [100] are taken as input to our algorithm.

To verify that it is necessary to account for side-chain uncertainty, we first tested our core algorithm on the subunit structure with incorrect side-chains. We randomly chose sets of rotamers for the side-chain conformations of the subunit and then ran our core algorithm of Chapter 3. We generated 100 such structures, and found that the set of consistent structures was empty for each of them. In fact, of the 7.50327×10^{44} different combinations of rotamers, our results below show that 7.50331×10^{44} are invalid. This implies that the core approach would fail to identify any consistent structures for any fixed subunit structure using one of these 7.50331×10^{44} rotamer combinations.

Figure 5.4 shows the satisfying regions obtained using the rotamers approach and compares these regions with those obtained from the core approach. As the figure shows, all the cells returned by the core approach are completely included in the rotamers approach, indicating that the rotamers approach does not miss any valid solutions. The volume of the SCS taken as input to the hierarchical subdivision was 22619.47 Å^2-rad^2, out of which 0.01335 Å^2-rad^2 is returned as satisfying regions, indicating the constraint that the 183 restraints are providing.

Figure 5.5 shows the effectiveness of our approach in eliminating invalid rotamers. The plot compares the average number of rotamers for each residue before and after identifying

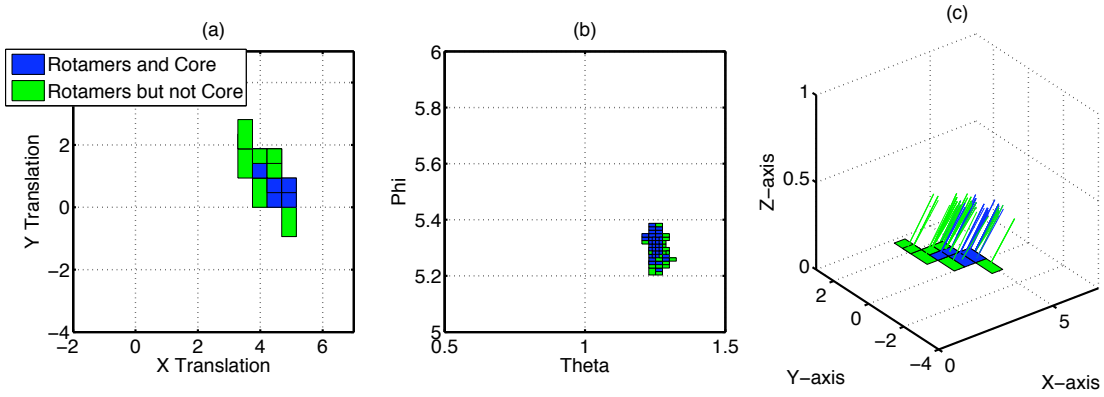


Fig. 5.4: Comparison of satisfying regions for MinE dimer (1EV0) using the rotamers approach and the core algorithm. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis.

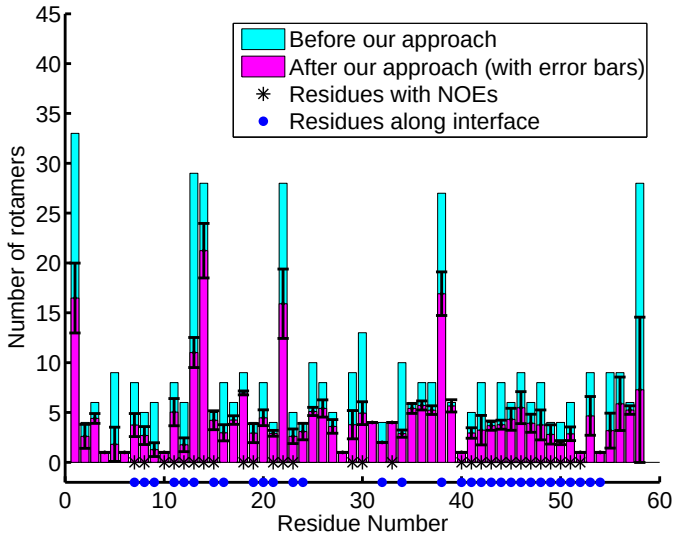


Fig. 5.5: Average number of rotamers for each residue for the homo-dimeric MinE (1EV0) before and after the branch-and-bound search.

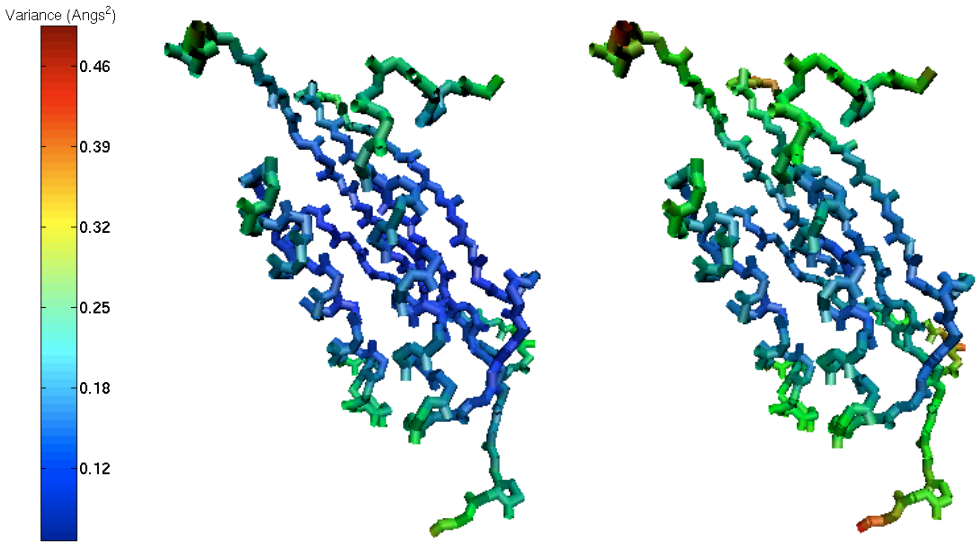


Fig. 5.6: Backbone variance in the set of satisfying structures for the homo-dimeric MinE (1EV0): (a) core algorithm (b) with rotamers when operators are applied during the search. Each backbone atom is colored by the variance in the position of the atom according to the color scale on the left.

the set of satisfying structures. Note that the number of rotamers remaining varies depending on the backbone conformation of the satisfying structure. The error bar for each residue indicates the variance in the number of rotamers remaining for that residue over the set of satisfying structures. Most of the residues for MinE have several rotamers pruned. For residues that either have distance restraints or lie along the interface, as many as half of the rotamers are pruned. All the rotamers remaining at the end of our approach are valid and belong to at least one conformation that is not involved in steric clashes and satisfies all the data.

Figure 5.6 compares the backbone variance in the set of satisfying structures with the

core algorithm versus the rotamers approach. The side-chains in the core algorithm are fixed and are obtained from the subunit structure chosen. The figure shows that the backbone variance is increased by considering rotamers. The average backbone variance increases from 0.13 $Å^2$ using the core algorithm to 0.19 $Å^2$ using the rotamers approach. This illustrates that considering uncertainty in the *side-chain* during the configuration space search allows identification of *backbones* that would have otherwise been missed and hence avoids false precision in the set of satisfying structures.

In the rotamers approach, we apply the pruning operators at each cell of the hierarchical subdivision. Figure 5.7 illustrates the advantage of applying the pruning operators during the hierarchical subdivision. Not applying an operator in a cell leads to no rotamers being eliminated due to that operator in that cell. This causes a cell to be eliminated only when all rotamers of any residue either violate a restraint or are involved in steric clash. This implies that a cell could be kept even when it has no valid conformation, leading to more cells being explored and the output being more conservative. As Figure 5.7(a) shows, the number of nodes explored increases from 9888 when operators are applied to 14646 when no operators are applied. Figures 5.7(b) and (c) compare the average backbone variance and average backbone RMSD in the satisfying structures, and illustrate the conservative nature of the output when no operators are applied. The average backbone variance is 1.7 times more and average backbone RMSD is 1.32 times more when no operators are applied. Figure 5.7(d) illustrates the average number of rotamers remaining in the set of satisfying structures. All rotamers taken as input are remaining (average of 8.3 per residue) when no operators are applied whereas on a average 4.6 rotamers per residue are remaining when

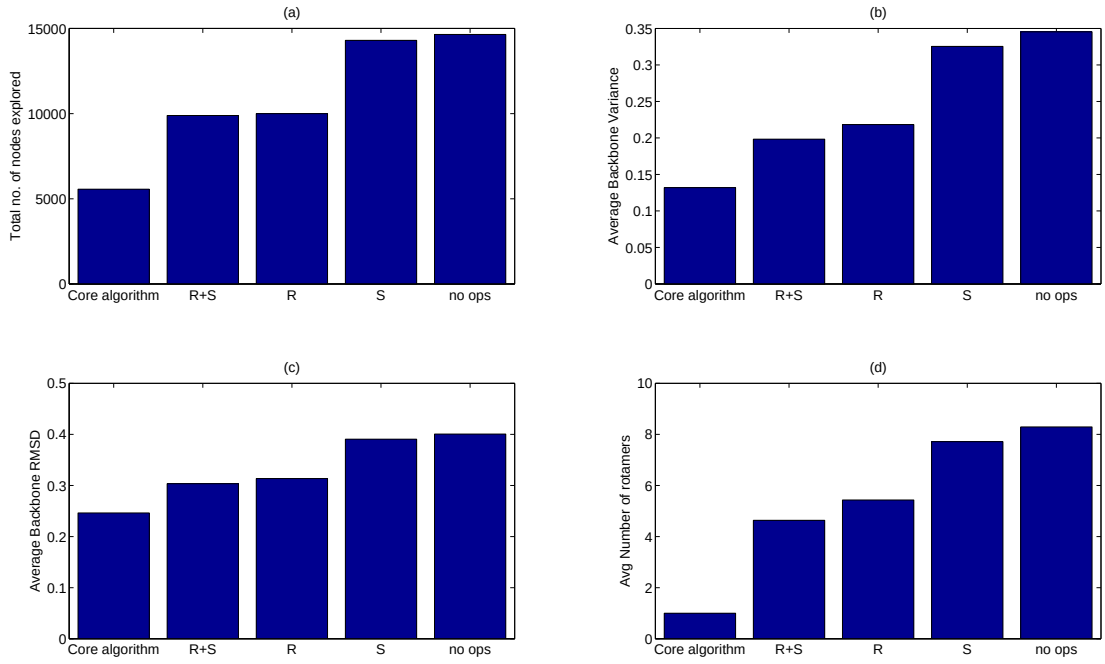


Fig. 5.7: Empirical analysis of the effect of the operators for the homo-dimeric MinE (1EV0) on (a) total number of nodes explored, (b) average backbone variance in the set of satisfying structures, (c) average backbone RMSD in the set of satisfying structures, (d) average number of rotamers over all residues and structures at the end of the hierarchical subdivision. Here *R* indicates the restraint-based operator, *S* indicates the steric-based operators, and "no ops" indicates no operators.

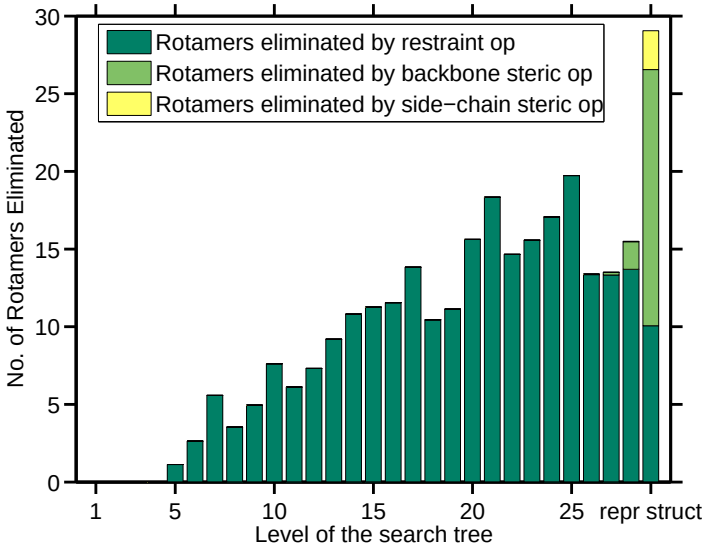


Fig. 5.8: Empirical analysis of the number of rotamers eliminated by the operators at each level of the search tree for the homo-dimeric MinE (1EV0). The number of rotamers eliminated at each level is calculated by finding, for each residue, the average number of rotamers eliminated in all nodes at that level, and then summing over all residues. The last level indicates the number of rotamers eliminated by applying the operators on the representative structures.

operators are applied. These results clearly indicate that applying the operators during the search is more efficient than applying them at the end of the search. A comparison between the core algorithm and the other cases shows that the core algorithm, though efficient in terms of the nodes explored, leads to false precision in the set of satisfying structures.

Figures 5.8 and 5.9 illustrate the effectiveness of the operators in eliminating rotamers and nodes at each level of the search tree. The plots show how many rotamers and nodes are pruned by each operator at each level of the search tree. As shown, the restraint-based

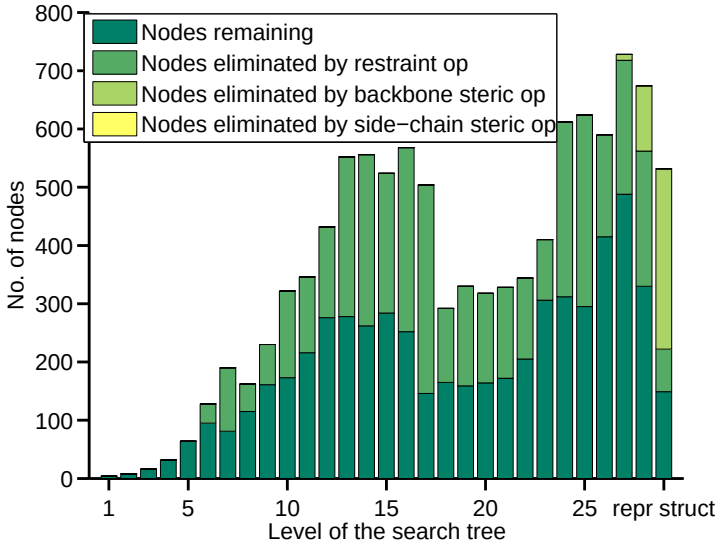
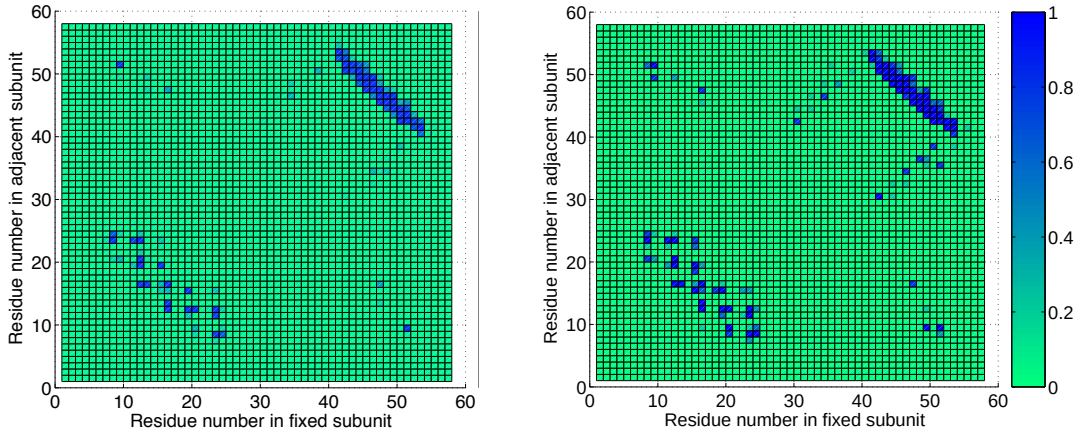


Fig. 5.9: Empirical analysis of the number of nodes eliminated by the operators at each level of the search tree for the homo-dimeric MinE (1EV0). The last level indicates the number of nodes eliminated by applying the operators on the representative structures.

operator is the most effective operator in eliminating rotamers and cells higher up in the tree. The steric operators start eliminating rotamers at level 25 when the average volume of the cell is around 3.5×10^{-5} Å^2-rad^2 and the average backbone RMSD in the structures represented by each cell is 1.05 Å. The side-chain steric-based operator eliminates rotamers only at the leaf level and for the representative structures. In general, there is a linear increase in the number of rotamers eliminated as the level of the tree increases, due to tighter bounds at deeper levels of the tree. Enough nodes are eliminated by the operators at each level to prevent the exponential increase in the number of nodes to be explored at the next level. This is clear from the nodes remaining at each level of the tree as shown in Figure 5.9.

As mentioned in Section 5.1, the DEE criterion is used to eliminate rotamers that lead to conformations with energy worse than a threshold within that of the best. The threshold was chosen as 10 kcal/mol (about 1% of the total vdW energy of the complex) for the current test cases. The average number of rotamers per residue eliminated by the DEE criterion in the set of representative structures was 2.1 while that for the restraint-based operator (over all levels) was 5.4. A larger energy threshold would lead to fewer rotamers being eliminated. Note that the DEE criterion does not cause elimination of nodes since the criterion ensures that the lowest energy (this includes the pseudo-energy term from restraint satisfaction) conformation is always present in each representative structure.

Figure 5.10 illustrates the interface maps for MinE for the set of satisfying and WPS structures. Recall that the set of satisfying structures includes sets of valid rotamers for each residue, while the set of WPS structures includes one rotamer (as decided by the A*-



(a) (b)

Fig. 5.10: Inter-subunit interface map for the homo-dimeric MinE (1EV0). The color scale indicates the frequency of contact between pairs of residues on adjacent subunits in the set of (a) satisfying structures and (b) WPS structures.

search and energy minimization) for each residue. In the set of satisfying structures, a pair of rotamers is considered to be in contact if any of their corresponding atoms are in contact. For each pair of residues in each structure, we first calculate the fraction of the rotameric combinations in which the pair of residues is in contact. We then average these values over all satisfying structures. Figure 5.10(a) illustrates the interface map for the set of satisfying structures. Twenty-eight pairs of residues are always in contact (in all structures). These pairs of residues indicate the regions of the subunits that interact with each other and could help understand the function of the complex. Further, thirty-four pairs of residues have a contact frequency of greater than or equal to 0.5. These pairs of residues could be targets for mutagenesis experiments to study their effect in the functionality of MinE. Unfortunately, there was no experimental mutagenesis data available with which to correlate our results.

The interface maps help gain insight into the arrangement of secondary structures at the interface of the complex. From the interface maps, it is clear that the pairs of residues that are in contact are mainly present in two regions. One of these regions involving residues 41 to 53 forms an anti-parallel beta sheet while the other region involving residues 12 to 23 forms an anti-parallel alpha helix. The anti-parallel nature is evident from the cross-diagonal nature of these blue(dark)-colored regions.

Figure 5.10(b) illustrates the interface map for the set of WPS structures. As before, a pair of residues is considered to be in contact if any of their corresponding atoms are in contact. The difference of this interface map from the interface map of the satisfying structures is that here we have only one rotamer per residue per structure. We now calculate the frequency of contact for each pair of residues as the ratio between the number of WPS structures in which the residues are in contact to the total number of WPS structures. As Figure 5.10(b) shows, 59 pairs of residues are always in contact. The 28 pairs of residues identified to be in contact in the set of satisfying structures are all included in the 59 pairs. This places greater confidence in the results obtained from the set of satisfying structures. The remaining 31 pairs do not have a contact frequency of 1 in the set of satisfying structures, since there exist valid rotamers for these pairs which are not in contact.

5.2.2 Results on Homo-trimeric Chicken Cartilage Matrix Protein (CCMP)

Our second test case is the homo-trimeric coiled-coil domain of the chicken cartilage matrix protein (CCMP). The subunit structure, the set of inter-subunit distance restraints and the

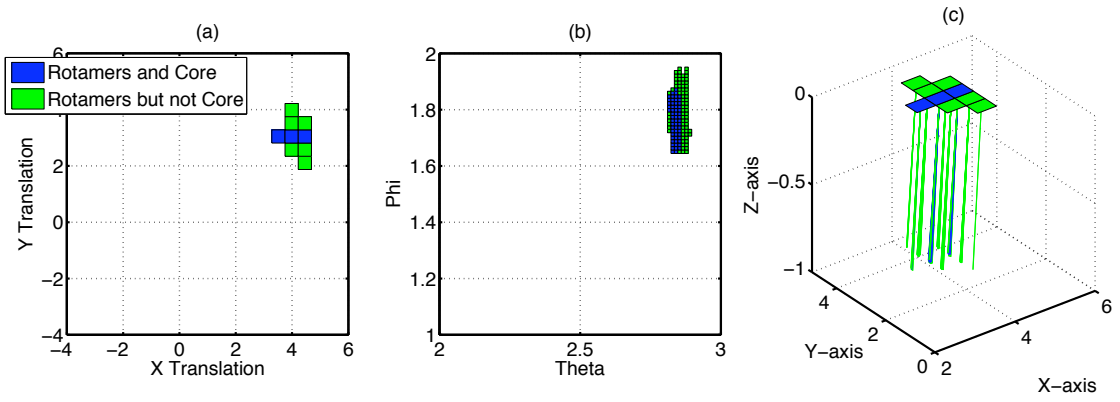
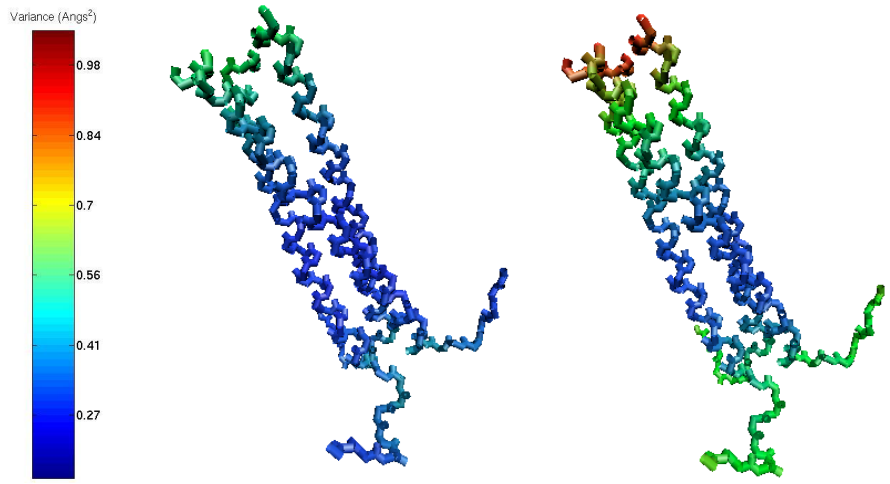


Fig. 5.11: Results on CCMP trimer (1AQ5): Comparison of satisfying regions using the rotamers approach and the core algorithm with the 183 experimental restraints. (a) Translation parameters. (b) Orientation parameters when projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis.

rotamers from the Lovell rotamer library are taken as input to our approach. We also used the set of intra-subunit NOEs and the χ-angle restraints available for CCMP and eliminated rotamers *a priori* that violate any of these restraints.

As before, we studied the importance of handling side-chain uncertainty by running our core algorithm on subunits with randomly chosen sets of rotamers for the side-chain conformations. No consistent structures were found for any of 100 such generated structures. From our results below, we find that 2.7933×10^{31} of the 2.7926×10^{31} possible different rotameric conformations are invalid. Hence, the core approach would fail to identify any consistent structures for any subunit structure with one of the 2.7933×10^{31} combinations.

Figures 5.11 and 5.12 compare the satisfying regions and structures obtained using the rotamers approach and the core approach. The volume taken as input was 45238.93 Å^2-



(a) (b)

Fig. 5.12: Backbone variance in the set of satisfying structures for the homo-trimeric CCMP (1AQ5): (a) core algorithm (b) with rotamers when operators are applied during the search. Each backbone atom is colored by the variance in the position of the atom according to the color scale on the left.

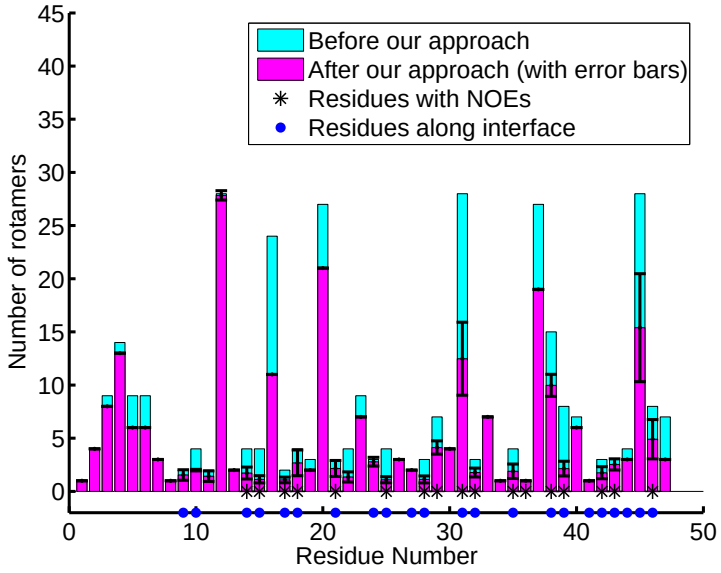


Fig. 5.13: Average number of rotamers for each residue for the homo-trimeric CCMP (1AQ5) before and after the branch-and-bound search.

rad^2. The satisfying regions are in one continuous region occupying a volume of 0.00358 Å^2-rad^2 and the satisfying structures have an average variance of 0.30 Å^2, indicating the constraint that the 49 restraints are providing. The average variance in the satisfying structures using the core approach was only 0.15 Å^2, showing the false precision that the core approach causes. The satisfying regions of the core approach are completely contained in the satisfying regions of the rotamers approach, showing the correctness of the rotamers approach.

Figure 5.13 compares the average number of rotamers for each residue before and after identifying the set of satisfying structures. Note that for some of the residues such as residues 1 and 2, no pruning occurs. This is because these residues are not involved in any restraints and do not lie along the inter-subunit interface. Neither the data nor vdW packing

affect side-chain conformations of these residues. For residues that either have distance restraints or lie along the interface, as many as half the rotamers are pruned. All invalid rotamers are eliminated and the rotamers remaining at the end of our approach belong to at least one valid conformation.

Figure 5.14 illustrates the advantage of applying the pruning operators during the hierarchical subdivision. As the Figure 5.14 shows, the number of nodes explored is almost three times more, the average variance is 3.4 times more, the average backbone RMSD is 1.8 times more and on a average 2.5 rotamers remain per residue when no operators are applied. These results clearly indicate that applying the operators during the search is more efficient than applying them at the end of the search. On comparing the results between the core algorithm and the other cases, we see that the core algorithm leads to false precision in the set of satisfying structures though it is efficient in terms of the nodes explored.

Figures 5.15 and 5.16 show the levels in the search tree at which each of the operators is effective in pruning invalid rotamers and nodes. The restraint-based operator starts eliminating rotamers at level 6 whereas the steric operators start eliminating rotamers at level 24 when the average volume of the cell is around 7×10^{-5} Å^2-rad^2 and average backbone RMSD in the structures represented by the cell is 1.32 Å. The nodes eliminated prevent an exponential increase in the number of nodes explored. Only 16736 nodes are explored at the leaf level instead of the 2.6×10^8 number of nodes that would have been explored without nodes being eliminated. The average number of rotamers per residue eliminated by the DEE criterion (threshold chosen as 10 kcal/mol as in the case of MinE) in the set of representative structures was 0.96 while that for the restraint-based operator (over all

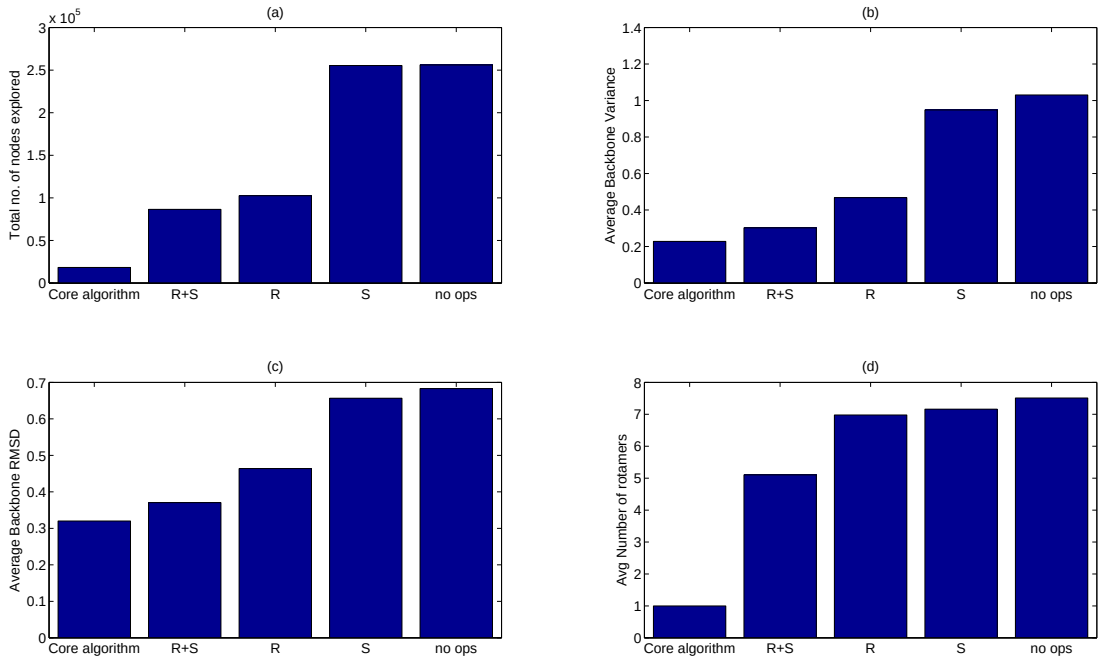


Fig. 5.14: Empirical analysis of the effect of the operators for the homo-trimeric CCMP (1AQ5) on (a) total number of nodes explored, (b) average backbone variance in the set of satisfying structures, (c) average backbone RMSD in the set of satisfying structures, (d) average number of rotamers over all residues and structures at the end of the hierarchical subdivision. Here *R* indicates the restraint-based operator, *S* indicates the steric-based operators, and "no ops" indicates no operators.

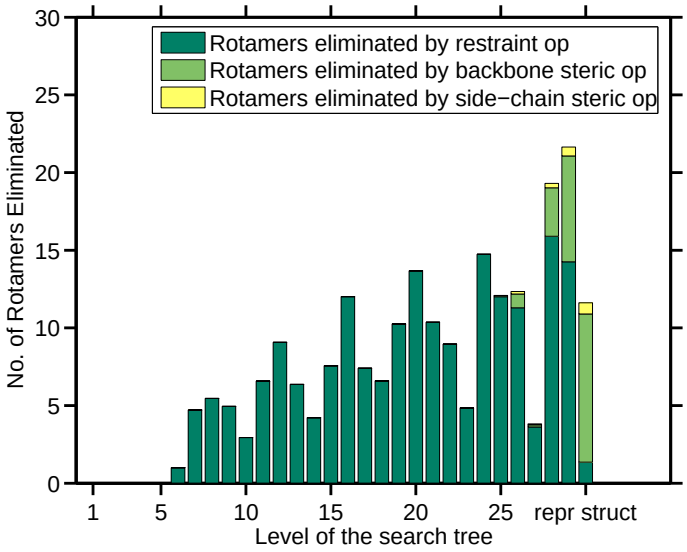


Fig. 5.15: Empirical analysis of the sum of the average number of rotamers per residue eliminated by the operators at each level of the search tree for the homo-trimeric CCMP (1AQ5). The last level indicates the number of rotamers eliminated by applying the operators on the representative structures.

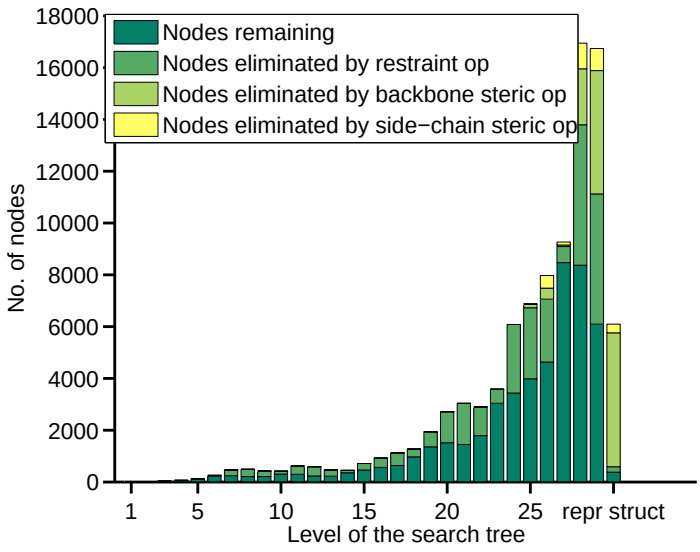
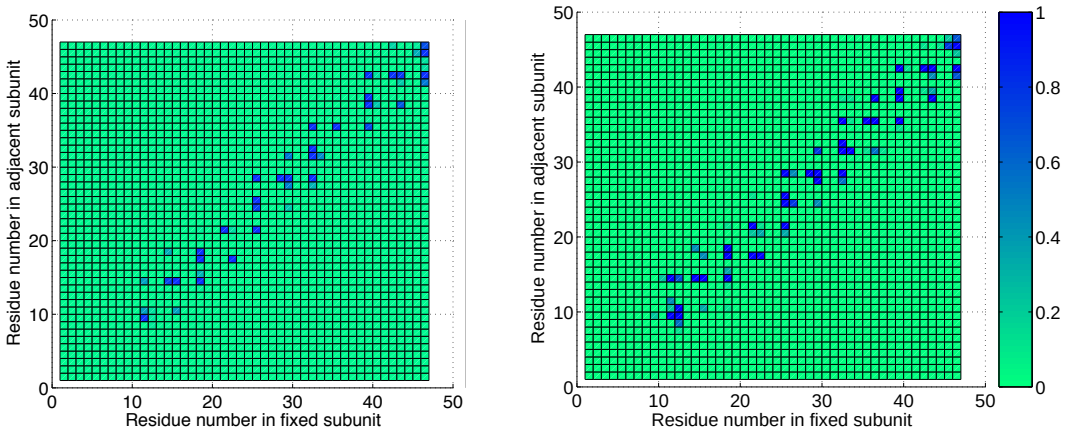


Fig. 5.16: Empirical analysis of the number of nodes eliminated by the operators at each level of the search tree for the homo-trimeric CCMP (1AQ5). The last level indicates the number of nodes eliminated by applying the operators on the representative structures.



(a) (b)

Fig. 5.17: Inter-subunit interface map for the homo-trimeric CCMP (1AQ5). The color scale indicates the frequency of contact between pairs of residues on adjacent subunits in the set of (a) satisfying structures and (b) WPS structures.

levels) was 4.3.

Figure 5.17 illustrates the interface maps for CCMP for the set of satisfying and WPS structures. Figure 5.17(a) illustrates the interface map for the satisfying structures. Sixteen pairs of residues are always in contact. One pair among these is residue 11 on the fixed subunit and residue 9 on the adjacent subunit. It has been shown experimentally that these two cysteine residues are indeed in close contact. Each pair of the cysteines on adjacent subunits form a ring of three (since CCMP is a trimer) inter-subunit disulfide bonds [70]. Seventeen other pairs of residues have a frequency of contact greater than or equal to 0.5. All these pairs are located along the diagonal of the interface map confirming the coiled-coil fold of the trimer. These pairs of residues could be targets for mutagenesis experiments to study their effect in the function and structure of the coiled coil.

Figure 5.17(b) illustrates the interface map of all pairs of residues in the set of WPS structures. As the figure shows, 36 pairs of residues have the frequency of contact as 1. The 16 pairs of residues identified to be always in close contact in the set of satisfying structures are all included in the 36 pairs. The remaining twenty pairs do not have a contact frequency of 1 in the set of satisfying structures, since there exist valid rotamers for these pairs which are not in contact.

5.3 Conclusions

In this chapter, we presented an approach to handle uncertainty in side-chain conformations of the input subunit structure. The developed algorithm extends the core algorithm of Chapter 3 and continues to have the features of being complete and data-driven. We defined a set of operators to eliminate invalid rotamers and cells during the search. We showed from our results on two test cases that our complete, data-driven approach is now applicable to scenarios where the subunit structure is not exactly known in complex. In both test cases, we showed that considering side-chain uncertainty during the search through the SCS avoids false precision in the output. Using our approach we identify that on an average 3.7 rotamers per residue for MinE and 2.5 rotamers per residue for CCMP are invalid. We also showed that applying the pruning operators to eliminate invalid rotamers during the search is more efficient than applying them at the end of the search. By applying the operators during the search rather than at the end of the search, the number of nodes explored is 1.5 times less for MinE and 3 times less for CCMP.

In this chapter and the previous chapters, the presented algorithms assumed limited or no uncertainty in the NOE data. In the next chapter, we present a probabilistic framework that is not only robust to noise and uncertainty in the NOE data, but also enables structural inference.

6. HOMO-OLIGOMERIC STRUCTURAL INFERENCE

In this chapter, we develop algorithms that enable structural inference on symmetric homo-oligomers, so that we can reason probabilistically about the degree of agreement between possible conformations and both data and biophysical constraints. Our approach for symmetric homo-oligomeric structural inference is not only complete, but also robust to experimental noise and uncertainty. The algorithms developed in the previous chapters assumed that there was no uncertainty in the available NOE data, that is, the upper bound on the distance between the restrained atoms provided a hard threshold. The presence of a few noisy (or spurious) NOEs (false positives), was considered in Chapter 4. As we show in the results, that approach would fail to identify consistent structures when the number of false positives in the data is greater than a preset number. Also, in previous chapters, we assumed that well-packed satisfying structures were all equally likely and evaluated them for restraint satisfaction and vdW packing independently. In this chapter, we handle spurious NOEs in a robust manner and assign a posterior probability to each structure indicating how well it satisfies both the data and the biophysical constraints. The likelihood of the structure captures the consistency of the structure to the data, while the prior captures the biophysical constraints.

Our approach is closely related to the Inferential Structure Determination (ISD) method of Wolfgang et al. [149], developed for protein structure determination by NMR (Section 2.4). The posterior probability of each conformation is obtained as the product of the likelihood of the data and the prior support for the structure. Due to the difficulties in estimating the posterior density over the entire conformation space, the ISD approach simulates the posterior density over possible conformations using a Markov Chain Monte Carlo (MCMC) algorithm based on the replica-exchange method [65]. These simulation techniques could miss native conformations. Our approach, on the other hand, takes a configuration space viewpoint. We perform a complete analysis of the configuration space and guarantee that native conformations are not missed, hence allowing for a more reliable structural inference than the ISD approach. Our complete approach identifies all cells in the SCS that are likely to have high posterior density. Regions are rejected only when they are not supported by the data or have extremely bad steric clashes, and thus would have negligible posterior density.

6.1 Algorithm

Given a particular conformation x and set of NOE restraints R, from Bayes theorem, the posterior probability of conformation x is given by

$$p(x, \sigma \mid R) \propto p(R \mid x, \sigma)\ p(x)\ p(\sigma), \tag{6.1}$$

where σ indicates the error in the system and includes both the experimental noise and systematic effects such as internal dynamics [99] and spin diffusion [103]. The σ is a

nuisance parameter redundant, so we obtain the marginal posterior distribution $p(x \mid R)$ by marginalizing over σ. Hence,

$$p(x \mid R) \;=\; \int p(x, \sigma \mid R)\, d\sigma . \tag{6.2}$$

From Bayes theorem again,

$$\begin{aligned} p(x \mid R) &= \int p(x, \sigma \mid R)\, \mathrm{d}\sigma & (6.3)\\ &\propto \int p(x \mid R, \sigma)\; p(\sigma)\, \mathrm{d}\sigma & (6.4)\\ &\propto \int p(R \mid x, \sigma)\; p(x)\; p(\sigma)\, \mathrm{d}\sigma & (6.5) \end{aligned}$$

In our approach, since we take a configuration space viewpoint, we consider the posterior for a point k in the configuration space and obtain it as,

$$p(k \mid R) \propto \int p(R \mid k, \sigma)\; p(k)\; p(\sigma)\, \mathrm{d}\sigma \tag{6.6}$$

As in the ISD approach [149], we assume that the restraints in R are conditionally independent given the structure and that they follow a log-normal distribution with mean d_i and variance σ for restraint r_i. The log-normal distribution was shown to be a natural choice for modeling errors in NOE distances by Rieping et al. [127]. The log-normal distribution is restricted to positive values and hence gives equal weight to over- or under-estimating the true value. This is not the case for error distributions defined on both positive and negative values, such as a Gaussian, which assign a non-vanishing probability to unobservable negative distances.

Under the assumptions of conditional independency and log-normal distribution, the

likelihood for a conformation represented by a point k in the SCS, $p(R \mid k, \sigma)$ is given by

$$p(R \mid k, \sigma) = \prod_{i=1}^{n} p(r_i \mid k, \sigma) \tag{6.7}$$

$$= \prod_{i=1}^{n} \frac{1}{\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\|\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} \log^2\left(\frac{\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\|}{d_i}\right)\right\} \tag{6.8}$$

where n is the number of restraints, r_i: $\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\| \leq d_i$ is the i^{th} restraint, $\mathbf{p}_i$ is the atom on the fixed subunit and $\mathcal{T}_k(\mathbf{q}_i)$ is the atom on the adjacent subunit corresponding to $\mathbf{q}_i$ in the fixed subunit in the homo-oligomer conformation represented by k.

As in the ISD approach [149], assuming that experiments are carried out at a constant temperature T and following the principle of Maximum Entropy [76], the prior $p(k)$ is obtained by

$$p(k) = \frac{1}{Z(T)} \exp\{-E(k)/(BT)\}, \tag{6.9}$$

where $E(k)$ is the physical energy of conformation represented by k, B is the Boltzmann's constant. Employing Jeffrey's prior [78], the commonly used prior for nuisance parameters such as σ, we set $p(\sigma) = \sigma^{-1}$.

In our approach for structural inference, we would like to estimate the posterior density of the conformation space of a given protein complex. We do this by searching through the configuration space of the complex and eliminating regions (and hence sets of conformations) that provably have low posterior density. We hierarchically subdivide the configuration space (as in the previous chapters), and in each cell, K, estimate the maximum and minimum posterior probability among all the conformations represented by the cell. Let $p_{\max}(K \mid R)$ and $p_{\min}(K \mid R)$ denote the maximum and minimum posterior of a single

conformation possible in the cell K given the set of restraints R. In other words,

$$p_{\max}(K \mid R) = \max_{k \in K} p(k \mid R) \tag{6.10}$$

$$p_{\min}(K \mid R) = \min_{k \in K} p(k \mid R). \tag{6.11}$$

The value of $p(k \mid R)$ can be obtained from Equation 6.6. One way to compute the exact values of $p_{\max}(K \mid R)$ and $p_{\min}(K \mid R)$ in a cell K is to enumerate all the conformations represented by K. This would be computationally infeasible. Hence we get an overestimate for $p_{\max}(K \mid R)$ and an underestimate for $p_{\min}(K \mid R)$ (shown below).

In the hierarchical subdivision, we would like to identify all structures that have their posterior within a threshold of the best. Let Δ (a fractional value less than 1) denote this user-defined threshold. We use two bounds (discussed below) to prune a cell in the hierarchical subdivision. The order in which the cells are explored affects the tightness of the bounds and hence when cells with low posterior are eliminated. In order to make the bound tight as early as possible, we bias the search towards cells with high posterior by exploring the nodes in the order of their $p_{\max}(K \mid R)$, maintained in a priority queue. As in the core algorithm, we stop subdividing cells and accept a cell when all structures represented by a cell are within a user-defined similarity level, τ_0 Å.

Bounds: The two bounds used to eliminate cells are as follows. The first bound eliminates a cell K_1 when $p_{\max}(K_1 \mid R)$ is Δ times worse than the $p_{\min}(K \mid R)$ of any K. In other words, we

$$\text{Eliminate } K_1 \text{ when, } p_{\max}(K_1 \mid R) < \Delta \cdot \max_{K_2 \in W} p_{\min}(K_2 \mid R) \tag{6.12}$$

where W is the set of cells explored so far. The second bound is obtained as follows.

For a leaf cell (an accepted cell), the posterior of the representative structure is considered representative of all the structures in the cell. Let us denote by p_{best} the maximum posterior probability among the representative structures of the accepted cells. On accepting a cell, p_{best} provides the second bound that can be used to eliminate cells which is

$$\text{Eliminate } K_1 \text{ when, } p_{\max}(K_1 \mid R) < \Delta \cdot p_{\text{best}} \tag{6.13}$$

We found that in practice, the first bound is useful to eliminate cells until a cell is accepted, and once a cell is accepted, the second bound provides for a tighter bound.

Estimates of maximum and minimum posterior in a cell: The over- and under-estimates for the posterior are obtained by considering the over- and under-estimates of the likelihood and the prior separately and then using Equation 6.6 to obtain the posterior.

Estimates for prior: The over- and under-estimates for the prior for a cell K are given by

$$p_{\max}(K) = \begin{cases} 1 - f_{\text{stericMax}}(K), \text{ when } R(K) > \tau_0 \text{Å} \\ \exp\{-E(x)/k_B T\}, \text{ when } R(K) \leq \tau_0 \text{Å} \end{cases} \tag{6.14}$$

$$p_{\min}(K) = \begin{cases} 1 - f_{\text{stericMin}}(K), \text{ when } R(K) > \tau_0 \text{Å} \\ \exp\{-E(x)/k_B T\}, \text{ when } R(K) \leq \tau_0 \text{Å} \end{cases} \tag{6.15}$$

where $R(K)$ denotes the maximum backbone RMSD among all the structures represented in K, $E(x)$ denotes the minimized energy of the structure $x \in X$ that is representative of all the structures in K, and τ_0 is the usual user-defined similarity level. For cells that are higher up in the search tree (RMSD greater than τ_0), the prior is the probability that the represented conformations are involved in steric clashes. The cell gets a prior of zero if all the conformations represented by the cell are involved in a steric clash. Conformations

having steric clashes could satisfy the restraints better, which could lead to incorrect bounds and elimination of valid cells. Accounting for steric clash during the search prevents this problem and avoids accepting conformations with steric clashes. Due to the cost involved in performing energy minimization at each cell during the search, we perform it only for the representative structures of the accepted cells. We use the same energy functions as in the previous chapters.

The functions $f_{\text{stericMin}}(K)$ and $f_{\text{stericMax}}(K)$ are obtained as follows.

$$f_{\text{stericMin}}(K) = \min\left(1, \frac{|\{a \in A: \exists b \in B\ \exists k \in K, \|\mathcal{T}_k(a) - b\| \leq d_{\text{steric}}\}|}{\delta}\right) \quad (6.16)$$

$$f_{\text{stericMax}}(K) = \min\left(1, \frac{|\{a \in A: \exists b \in B\ \forall k \in K, \|\mathcal{T}_k(a) - b\| \leq d_{\text{steric}}\}|}{\delta}\right) \quad (6.17)$$

where A is the set of atoms on the adjacent subunit, B is the set of atoms on the fixed subunit, $\|x - y\|$ denotes the distance between x and y, d_{steric} denotes the threshold distance for declaring a pair of atoms to be involved in a steric clash and δ denotes the user-defined value which indicates the minimum number of atoms that must be involved in a clash for a structure to be eliminated due to steric clash. The numerator of $f_{\text{stericMin}}(K)$ denotes the number of atoms that are involved in a steric clash in *at least one of* the conformations represented in K, while the numerator of $f_{\text{stericMax}}(K)$ denotes the number of atoms that are involved in a steric clash in *all* the conformations represented in K. Dividing these numerators by δ gives estimates for the fraction of atoms that are involved in steric clashes. We take a linear form for these functions since the larger the number of atoms clashing in a conformation, the more likely that the conformation will have high vdW energy and be eliminated due to steric clash.

Estimates for likelihood: The overestimate on the likelihood for a cell K is obtained as follows. We have

$$p_{\max}(R \mid K, \sigma) \leq \prod_{i=1}^{n} p_{\max}(r_i \mid K, \sigma) \tag{6.18}$$

where n is the number of restraints and for each restraint r_i,

$$\begin{aligned} p_{\max}(r_i \mid K, \sigma) &= \max_{k \in K} p(r_i \mid k, \sigma) && (6.19) \\ &= \max_{k \in K} \frac{1}{\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\| \sqrt{2\pi}\,\sigma} \cdot \\ &\quad \exp\left\{ -\frac{1}{2\sigma^2} \log^2 \left(\frac{\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\|}{d_i} \right) \right\} && (6.20) \end{aligned}$$

where each restraint r_i is of the form $\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\| \leq d_i$ with $\mathbf{p}_i$ as an atom on the fixed subunit and $\mathcal{T}_k(\mathbf{q}_i)$ as an atom on the adjacent subunit in the conformation represented by k. (Note that the form of Equation 6.20 is same as that of Equation 6.8.) Let $\min_i$ and $\max_i$ denote the minimum and maximum possible values of $\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\|$ in K. Estimates for $\min_i$ and $\max_i$ are obtained using our geometric bound (convex hull) on the possible positions of $\mathcal{T}_k(\mathbf{q}_i)$, $\forall k \in K$ (see Section 3.1.1). Since the mode of a log-normal distribution with mean $\log d_i$ and variance σ is at $\exp(\log d_i - \sigma^2)$, the maximum possible probability for any log-normal distribution is obtained by substituting this value for $\|\mathbf{p}_i - \mathcal{T}_k(\mathbf{q}_i)\|$ in the above equation. Hence the $p_{\max}(r_i \mid K, \sigma)$ is obtained as

$$p_{\max}(r_i \mid K, \sigma) = \begin{cases} \frac{1}{\exp(\log d_i - \sigma^2)\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \log^2 \left(\frac{\exp(\log d_i - \sigma^2)}{d_i} \right) \right\}, \\ \qquad \text{when } \min_i \leq \exp(\log d_i - \sigma^2) \leq \max_i \\ \frac{1}{\min_i \sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \log^2 \left(\frac{\min_i}{d_i} \right) \right\}, \text{when } \exp(\log d_i - \sigma^2) < \min_i \\ \frac{1}{\max_i \sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \log^2 \left(\frac{\max_i}{d_i} \right) \right\}, \text{when } \exp(\log d_i - \sigma^2) > \max_i \end{cases} \tag{6.21}$$

Similarly the underestimate on the likelihood is obtained as

$$p_{\min}(R \mid K, \sigma) \geq \prod_{i=1}^{n} p_{\min}(r_i \mid K, \sigma) \tag{6.22}$$

where

$$p_{\min}(r_i \mid K, \sigma) = \min \left(\frac{1}{\min_i \sqrt{2\pi}\,\sigma} \exp \left\{ -\frac{1}{2\sigma^2} \log^2 \left(\frac{\min_i}{d_i} \right) \right\}, \right. \\ \left. \frac{1}{\max_i \sqrt{2\pi}\,\sigma} \exp \left\{ -\frac{1}{2\sigma^2} \log^2 \left(\frac{\max_i}{d_i} \right) \right\} \right) \tag{6.23}$$

Moments of the posterior distribution

At the end of the hierarchical subdivision, we have a set of leaf cells, where each leaf cell represents structures with high posterior probability. Recall that each point in these leaf cells represents a conformation. Let us denote the regions represented by the union of the leaf cells as P. We are interested in obtaining the moments of the posterior distribution over the configurations, the expected value and variance in the positions of the atoms. Let us consider the position of a particular atom, m. Since each point in P represents one possible position for m, we obtain the expected value and variance in m by integrating over P. Hence,

$$\begin{aligned} E(m \mid R) &= \int_{k \in P} E(\mathcal{T}_k(m) \mid k, R)\ p(k \mid R)\, \mathrm{d}k \\ &= \int_{k \in P} \mathcal{T}_k(m)\ p(k \mid R)\, \mathrm{d}k \end{aligned} \tag{6.24}$$

$$\mathrm{var}(m \mid R) = \int_{k \in P} (\mathcal{T}_k(m) - E(m \mid R))^2\ p(k \mid R)\, \mathrm{d}k \tag{6.25}$$

where $\mathcal{T}_k(m)$ denotes the position of m in the structure represented by k. The $p(k \mid R)$ above is obtained from Equation 6.6. Note that $\mathcal{T}_k(m)$ for the atom m on the adjacent

subunit is obtained by using the vector equation (Section 3.1.1)

$$\mathcal{T}_k(m) = ((\mathbf{m}-\mathbf{t})\cdot\mathbf{a})\mathbf{a} + (\sin\alpha)(\mathbf{a}\times(\mathbf{m}-\mathbf{t})) + (\cos\alpha)((\mathbf{m}-\mathbf{t}) - ((\mathbf{m}-\mathbf{t})\cdot\mathbf{a})\mathbf{a}) + \mathbf{t} \quad (6.26)$$

where $\mathbf{m}$ denotes the position of m in the fixed subunit, α denotes the angle of symmetry, $\mathbf{a}$ denotes the vector corresponding to the orientation parameters and $\mathbf{t}$ denotes the vector corresponding to the translation parameters of $k = (\mathbf{a}, \mathbf{t}) \in \mathbf{P}$. Obtaining an analytical expression for the integrals in Equations 6.24 and 6.25 is hard, given the values for $\mathcal{T}_k(m)$ and $p(k \mid R)$ as mentioned.

Since obtaining the integrals is hard, we approximate it by sampling. We could obtain the set of samples of the posterior distribution by choosing one point from each of the leaf cells. The problem with this is that the nature of the hierarchical subdivision (specifically the branching) could place points representing similar conformations in different cells. This would cause such a sampling to be biased. We avoid this bias by clustering the leaf cells based on similarity of their representative conformations and choosing representatives from each of the clusters (as in the core algorithm). The number of samples of the conformations is hence equal to the number of clusters. This sampling does not have similar conformations being overly represented and is hence not biased. As in the previous chapters, we refer to this clustered set of structures as well-packed satisfying (WPS) structures. Note that this sampling is different from the stochastic sampling of the ISD approach [149]. Our samples come with the guarantee that a native conformation (of sufficient posterior) is within a user-defined similarity to at least one of the samples. Let us denote the set of points in the SCS

representing the samples by C. The above equations now become

$$E(m \mid R) = \sum_{c \in C} \mathcal{T}_c(m)\ p(c \mid R) \tag{6.27}$$

$$\text{var}(m \mid R) = \sum_{c \in C} (\mathcal{T}_c(m) - E(m \mid R))^2\ p(c \mid R) \tag{6.28}$$

6.2 Results

We validated our approach on two test cases: homo-dimeric MinE (1EV0) and homo-trimeric CCMP (1AQ5). We extended the core algorithm of Chapter 3 for structural inference, assuming that there is no ambiguity in the NOE data and no uncertainty in the subunit structure. We chose as subunit structure the first chain of the most representative conformer of the ensemble of structures deposited in the Protein Data Bank [17]. The experimental error, σ, was obtained from the deposited NOE data. The value of Δ, the threshold for rejecting cells (Equations 6.12 and 6.13), was chosen for the following tests as 10^{-3} (that is, we eliminate cells 1000 times worse than the optimal). In order to deal with extremely small values of probabilities, we replace probabilities by log (base e) probabilities in all the equations.

6.2.1 Results on Homo-dimeric MinE

Figure 6.1 shows the satisfying regions obtained using our structural inference approach with the 183 available inter-subunit NOEs. As seen, the 183 restraints lead to a very small satisfying region indicating the constraint that they provide. Figure 6.2 compares the sat-

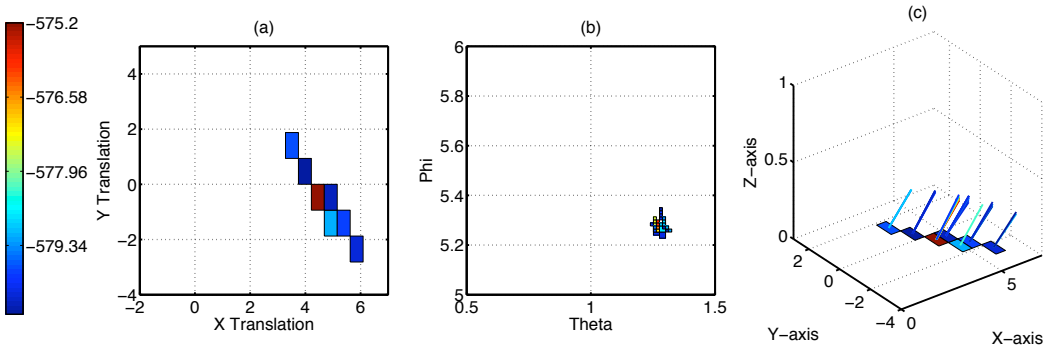


Fig. 6.1: Satisfying regions for the MinE dimer (1EV0) using the inference approach. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis. Regions are colored by the log(likelihood) of the representative structures according to the color scale shown on the left.

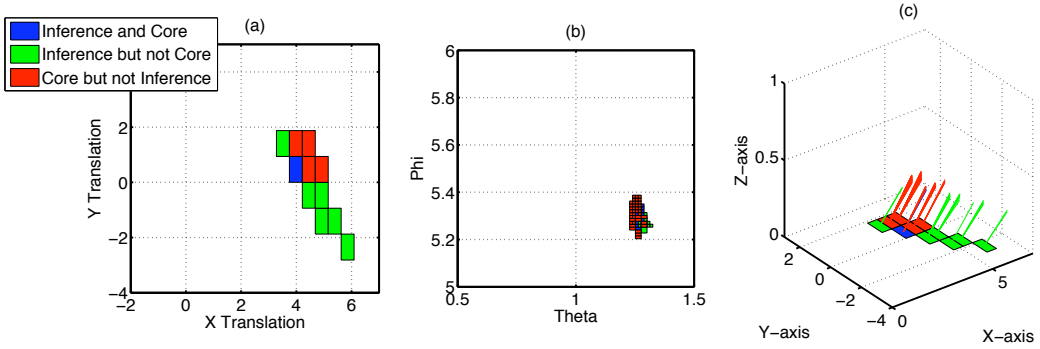


Fig. 6.2: Comparison of satisfying regions for the MinE dimer (1EV0) using the inference approach versus the core algorithm. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis.

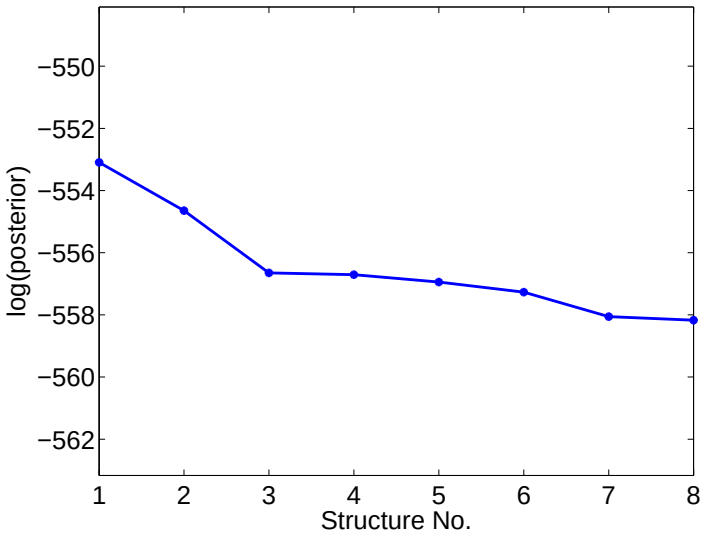


Fig. 6.3: Log posterior probabilities of the set of structures for the MinE dimer (1EV0).

isfying regions obtained using the inference approach and the core algorithm (Chapter 3). The volume of the satisfying regions using the inference approach is 0.0032 $Å^2$-rad^2 while that for the core approach is 0.0094 $Å^2$-rad^2. The overlap in the volume returned by the two approaches is 0.0006 $Å^2$-rad^2. The inference approach eliminates the remaining cells (volume 0.0088 $Å^2$-rad^2) since their posterior is more than $\log_e 10^3$ times worse than the maximum posterior.

Figure 6.3 illustrates the posterior probabilities of the set of WPS structures under the inference approach. The maximum log posterior probability is -553.1. As the figure shows, after the first two structures with maximum posterior, the distribution is almost flat until all structures within $\log_e 10^{-3}$ of the maximum posterior are identified. This indicates that the first two structures mostly determine the position of the atoms. These posterior probabilities are used to calculate the expected value and variance in the position of the atoms.

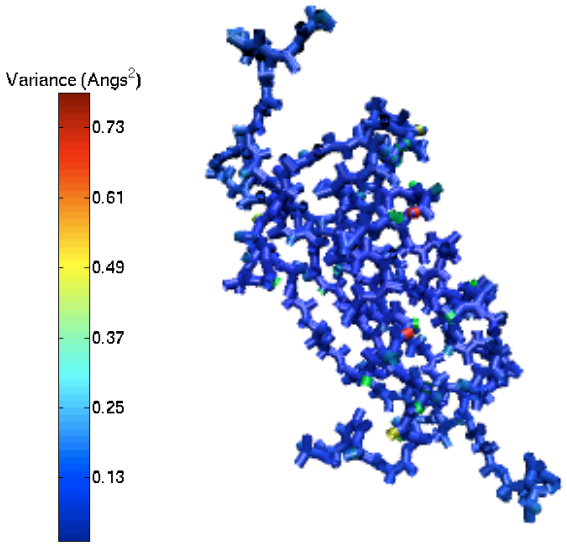


Fig. 6.4: Mean positions of the backbone atoms in the set of WPS structures for MinE dimer (1EV0). The variance in the position of each atom is illustrated by the color scale shown on the left.

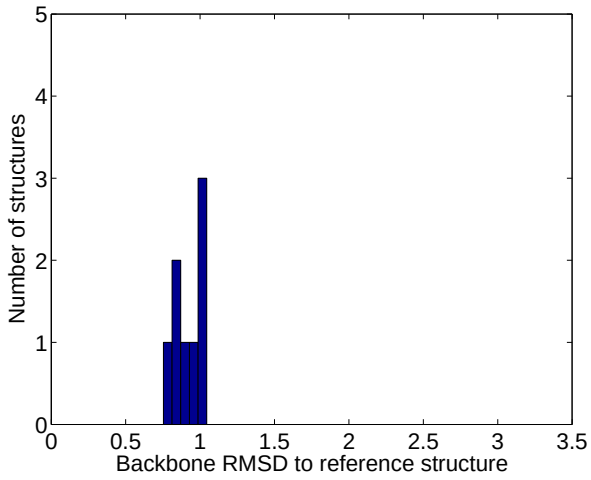


Fig. 6.5: Histogram of backbone RMSD for the MinE dimer (1EV0) between the reference structure and the set of structures obtained using the structural inference approach.

The structural inference approach enables us to infer the posterior probability of every possible conformation given the set of experimental restraints and the prior information (incorporating biophysical modeling). Using the inferred probability we estimate the expected value and variance in the positions of the atoms (Equations 6.27 and 6.28). Figure 6.4 shows the expected value and variance in the position of the backbone atoms in the set of WPS structures. As the figure illustrates, the 183 NOE restraints constrain the position of most of the atoms, leading to an average backbone RMSD of 0.28 Å and an average backbone variance of 0.05 $Å^2$. In the core algorithm we evaluate each conformation by a satisfaction score and a packing score, and consider a conformation to be valid if it has good scores. The well-packed satisfying structures were considered equally likely. This resulted in an average variance in the positions of the atoms of 0.26 $Å^2$, compared to 0.05 $Å^2$ under the inference approach. This indicates that by considering all conformations as being equally likely, we are being overly conservative in evaluating the structural constraint provided by the data and packing.

Figure 6.5 shows the distribution of the backbone RMSD of the structures to the reference structure. As the histogram shows, the peak lies below 1 Å and we identify structures as close as 0.75 Å to the reference structure. The tightness of the distribution is a clear indication of the structural constraint that the 183 experimental restraints provide.

Tests for Robustness to Noise:

In this section we test how our approach performs when false positives and experimental error exist in the NOE data. We simulate noise in the NOE data in the following way. We

have an ensemble $\mathcal{S}$ of structures determined by NMR. We simulate a set of inter-subunit distance restraints, R, from one structure $s_1 \in \mathcal{S}$. We identify the subset of the R restraints, T, that are violated by the subunit of another structure $s_2 \in \mathcal{S}$. We then use our structural inference approach to determine the structure of the complex using as input the subunit structure from s_2 and the experimental restraints augmented with T. Since the set T is not exactly satisfied by s_2, adding these restraints to the experimental restraints simulates false positives and uncertainty in the input data. This kind of a simulation is closer to real-world scenarios and better than adding random noise.

We applied this approach to the MinE dimer, choosing as s_1 model 9 from the ensemble of 15 deposited structures and as s_2 the reference structure (most representative conformer). We chose model 9 since it had the maximum all-atom RMSD of 3.74 Å to the reference structure. We simulated inter-subunit NOE restraints from model 9 by finding pairwise distances between protons on adjacent monomers. Every pair that had a distance between 5 Å and 6 Å was chosen as a restraint and an uncertainty of ± 1 Å was added. Twenty-two of these restraints were violated in the reference structure (16 by more than 1 Å and 2 by about 19 Å). The set of 183 experimental restraints was augmented with these 22 restraints to yield a set of 205 restraints.

We first tested the performance of a simple extension of the core approach (Chapter 3). The original core algorithm produced an empty set of solutions since a cell in the hierarchical subdivision is eliminated if even one NOE is violated. We extended that approach to handle false positives, as in Chapter 4, accepting a cell until a fixed number of NOEs is violated. Let us denote by δ the number of NOEs that can be violated in a cell. Figure 6.6

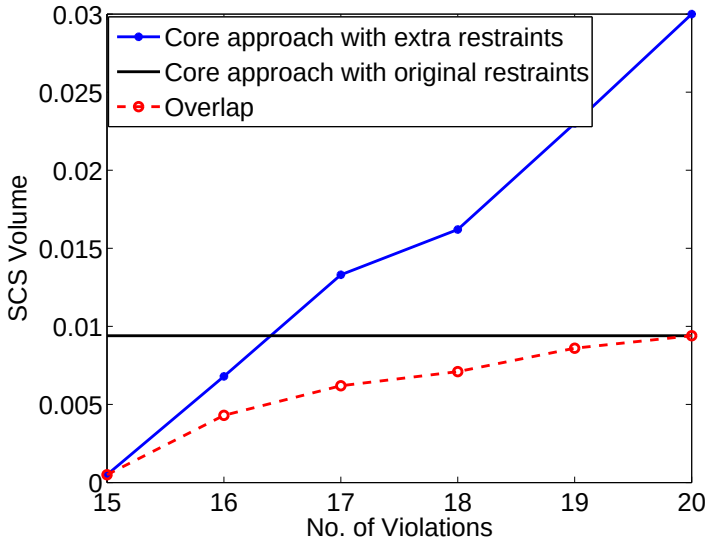


Fig. 6.6: SCS volume for the MinE dimer (1EV0) using the core algorithm, with an increasing number of allowed violations.

shows the change in SCS volume as δ is increased from 15 to 20. (No solutions were obtained when fewer than 15 NOE violations were considered.) When $\delta = 15$, no structure within 1 Å RMSD to the reference structure is identified. As δ is increased, we do get structures close to the reference structure, but along with them we also accept several structures that might not be valid. This is indicated by the non-overlapping SCS regions between the original core algorithm using the 183 experimental restraints versus the extended core algorithm with 205 restraints. The non-overlapping volume increases from 0.0025 $Å^2$-rad^2 to 0.0206 $Å^2$-rad^2 as δ increases from 15 to 20. Depending on the choice of δ, the correctness of the solutions varies and the best value for δ is unclear *a priori*.

We then tested the robustness to noise of the inference approach. Figure 6.7 illustrates the satisfying regions obtained using our structural inference approach with the noisy

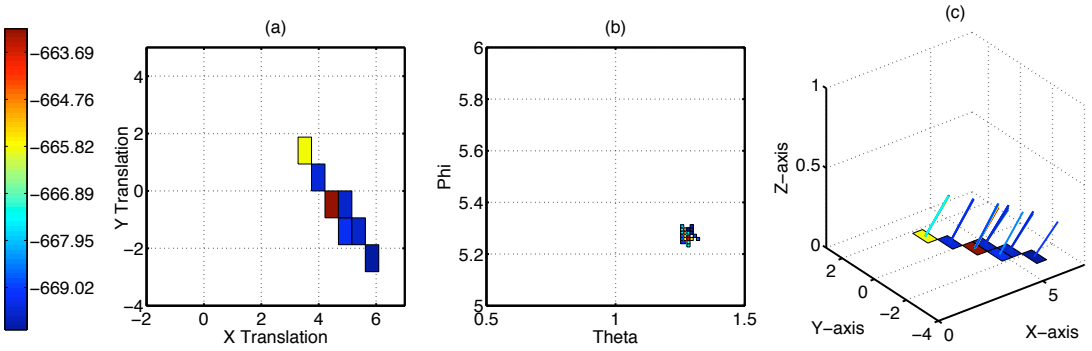


Fig. 6.7: Satisfying regions for the MinE dimer (1EV0) with 183 experimental and 22 noisy restraints. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis. Regions are colored by the log(likelihood) of the representative structures according to the color scale shown on the left.

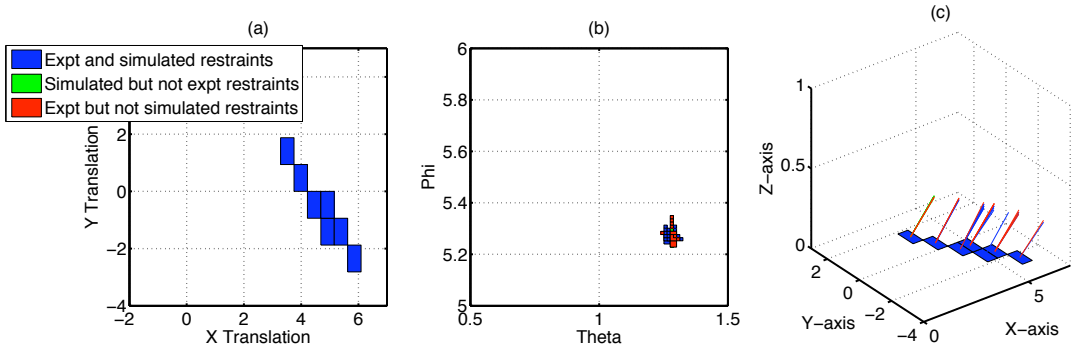


Fig. 6.8: Comparison of satisfying regions for the MinE dimer (1EV0) using just the 183 experimental restraints versus the 183 experimental + 22 noisy restraints. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis.

restraints (183 experimental and 22 simulated restraints). The small satisfying regions returned indicate the constraint provided by the data. The distribution of backbone RMSD to the reference structure (not shown) peaks at below 1Å, as expected from the overlap in SCS volume between the regions obtained using the 183 experimental restraints versus the those using the 205 restraints (discussed below). We are able to identify structures with backbone RMSD of 0.75 Å to the reference structure, in spite of the noise present in the data.

Figure 6.8 compares the satisfying regions obtained with the 183 experimental restraints versus the 205 restraints. As seen, most of the regions obtained with the 183 restraints are also obtained with the 205 restraints. The volume of the satisfying regions using the 183 restraints is 0.0032 $Å^2$-rad^2 while that using the 205 restraints is 0.0023 $Å^2$-rad^2. The overlap in the volume returned by the two approaches is 0.0019 $Å^2$-rad^2. This indicates that in spite of the noise, the regions identified are almost all valid. Also, all satisfying regions obtained using the 183 restraints and having log posterior within 10^{-4} of the maximum posterior are included in the satisfying regions obtained using the 205 restraints. This indicates the robustness of our approach to noise in the NOE data. The structural inference approach does not need parameters such as δ and enables us to identify the structures that best satisfy the available data.

6.2.2 Results on Homo-trimeric CCMP

Our second test case was the homo-trimeric CCMP, which had 49 experimental restraints. Figure 6.9 shows the satisfying regions obtained using our structural inference approach with the 49 available inter-subunit NOEs. The satisfying regions, though small, are not as small as the MinE, indicating the relatively smaller constraint that the 49 experimental restraints provide. Figure 6.10 compares the satisfying regions obtained using the inference approach and the core algorithm. As seen, most of the regions obtained by the core algorithm are also obtained with the inference approach. The volume of the satisfying regions using the inference approach is 0.019 $Å^2$-rad^2 while that for the core approach is 0.0058 $Å^2$-rad^2. The overlap in the volume returned by the two approaches is 0.0051 $Å^2$-rad^2. The inference approach does not identify the remaining cells (with volume 0.0007 $Å^2$-rad^2) since these cells have a likelihood more than $\log_e 10^3$ times worse than the maximum likelihood.

Figure 6.11 illustrates the log posterior probabilities of the set of WPS structures using the inference approach. The maximum posterior probability is -119.0. As the figure shows, the distribution is smooth and has a small slope indicating the gradual decrease in the posterior. This causes the concern that the threshold chosen does not capture the posterior distribution well enough. But, note that (as mentioned in the previous paragraph) a volume of 0.0051 $Å^2$-rad^2 of the 0.0058 $Å^2$-rad^2 obtained using the core approach is also obtained by the inference approach indicating that almost all of the valid conformations are captured.

Figure 6.12 shows the expected value and variance in the position of the backbone

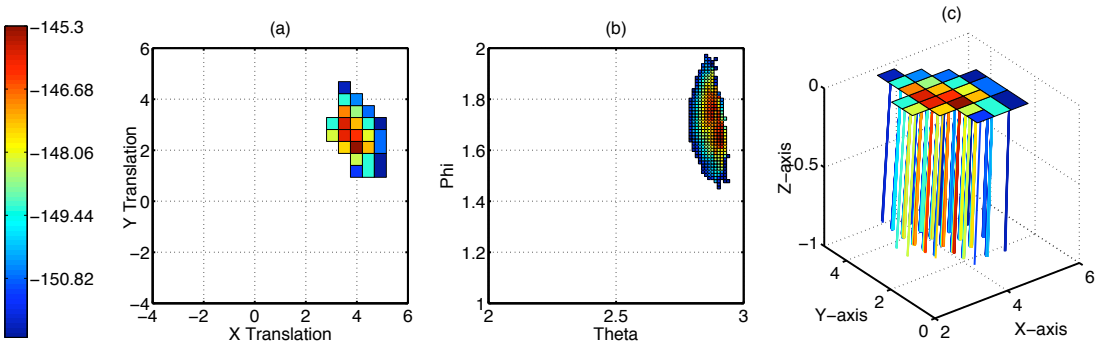


Fig. 6.9: Satisfying regions for the CCMP trimer (1AQ5) with 49 experimental restraints. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis. Regions are colored by the log(likelihood) of the representative structures according to the color scale shown on the left.

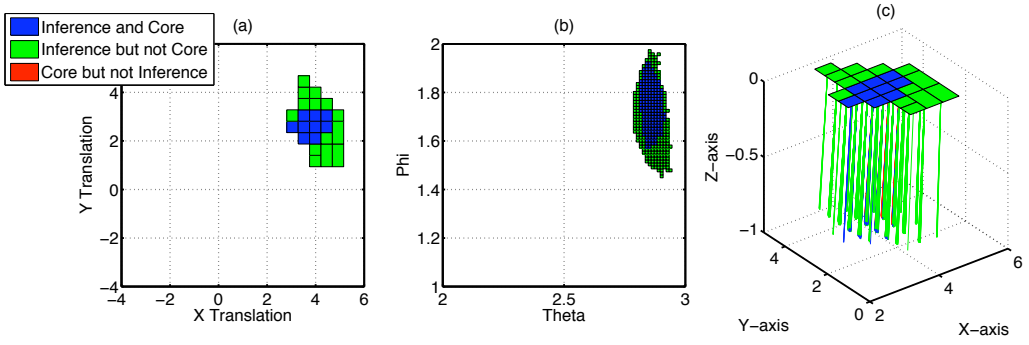


Fig. 6.10: Comparison of satisfying regions for the CCMP trimer (1AQ5)using the inference approach versus the core algorithm. (a) Translation parameters. (b) Orientation parameters, when projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis.

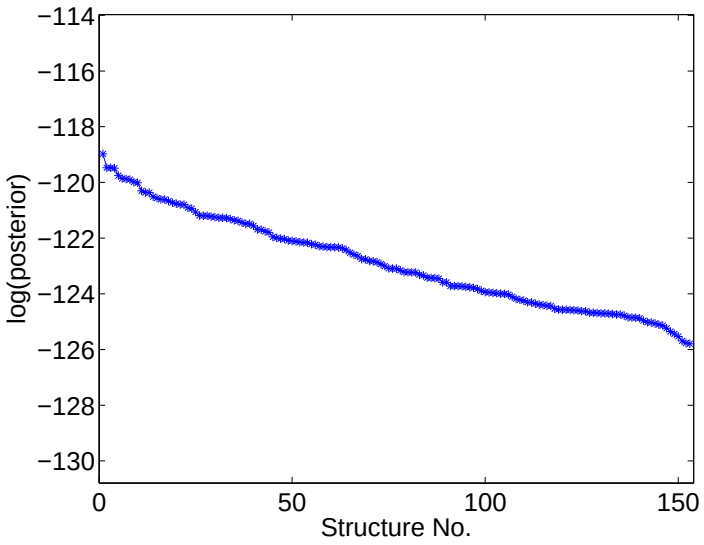


Fig. 6.11: Log posterior probabilities of the set of structures for the CCMP trimer (1AQ5).

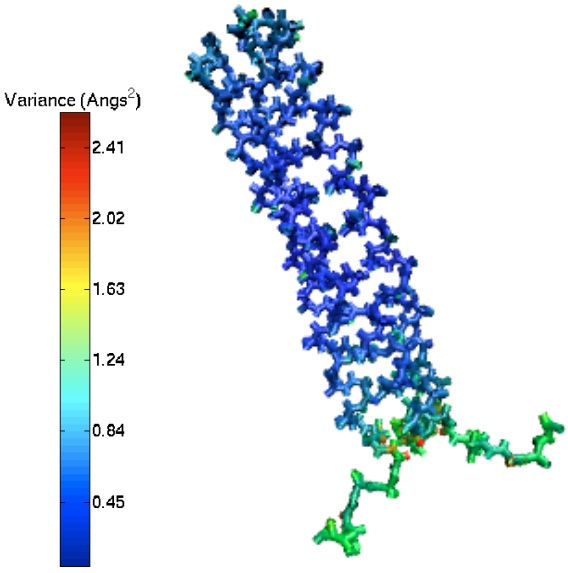


Fig. 6.12: Mean positions of the backbone atoms in the set of WPS structures for the CCMP trimer (1AQ5) using the inference approach with the 49 experimental restraints. The variance in the position of each atom is illustrated by the color scale shown on the left.

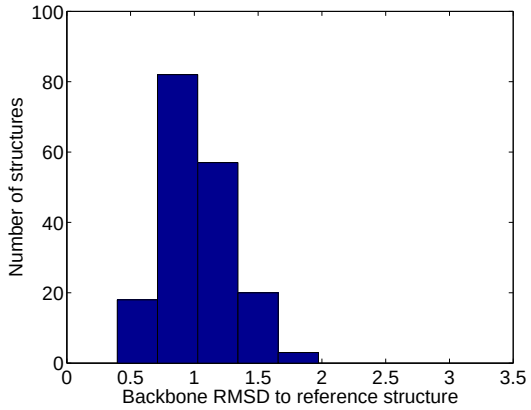


Fig. 6.13: Histogram of backbone RMSD for the CCMP trimer (1AQ5) between the reference structure and the set of structures obtained using the structural inference approach.

atoms in the set of WPS structures. As the figure illustrates, the 49 NOE restraints constrain the positions of most of the atoms, leading to an average backbone RMSD of 0.61 Å and an average backbone variance of 0.51 $Å^2$. The average variance in the positions of the atoms when weighting by probability versus the core approach when no weighting is done increases from 0.51 $Å^2$ to 0.94 $Å^2$. This indicates, as before, that structural inference enables a more precise characterization of structural uncertainty than the core algorithm.

Figure 6.13 shows the distribution of the backbone RMSD of the structures to the reference structure. As the histogram shows, the peak lies below 1 Å and we identify structures as close as 0.4 Å to the reference structure. The distribution ranges from 0.5 Å to 2 Å indicating the constraint provided by the 49 experimental restraints.

Tests for Robustness to Noise:

We simulated noise as in the case of the MinE dimer. We chose as s_1 model 10 from the

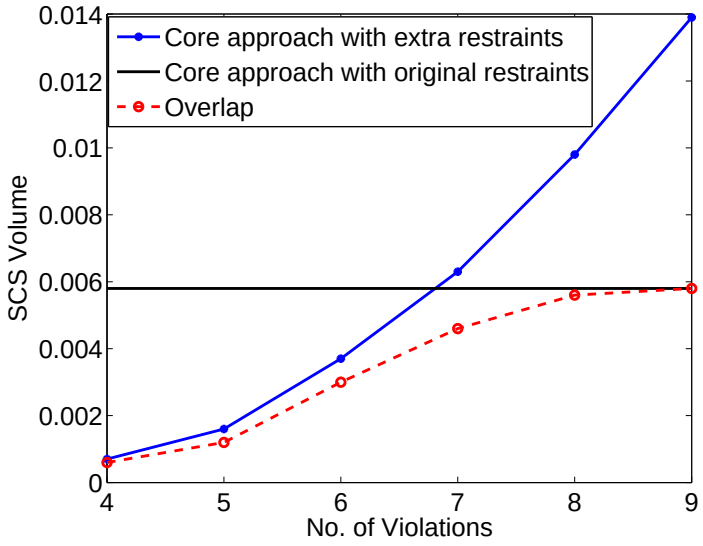


Fig. 6.14: SCS volume for the CCMP trimer (1AQ5) using the core algorithm, with an increasing number of allowed violations.

ensemble of 20 deposited structures and as s_2 the reference structure (most representative conformer). We chose model 10 since it had the maximum all-atom RMSD of 1.46 Å to the reference structure. The set of 49 experimental restraints was augmented with the 11 restraints from model 10 violated in the reference structure, yielding a set of 60 restraints.

As in the case of MinE, we first tested the performance of a simple extension of the core approach. The original core algorithm with no violations, that is, $\delta = 0$, produced an empty set of solutions. Solutions were obtained when the core algorithm was extended to allow for $\delta > 0$. Figure 6.14 shows the change in SCS volume as δ is increased from 4 to 9. (No solutions were obtained when fewer than four NOE violations were considered.) When $\delta = 4$, no structure within 1 Å RMSD to the reference structure is identified. The non-overlapping volume between the original core algorithm with 49 restraints and the core

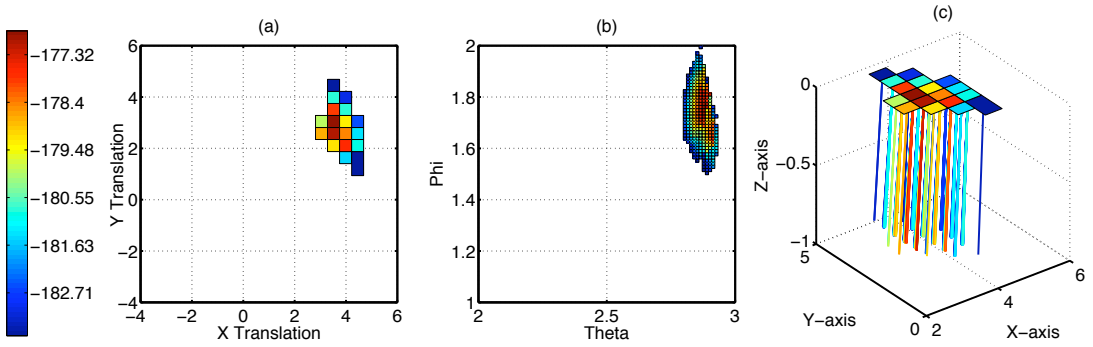


Fig. 6.15: Satisfying regions for the CCMP trimer (1AQ5) with 49 experimental and 11 noisy restraints. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis. Regions are colored by the log(likelihood) of the representative structures according to the color scale shown on the left.

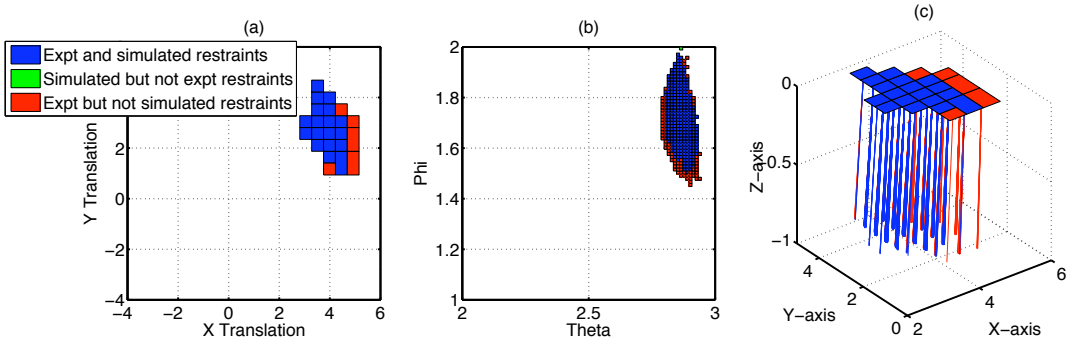


Fig. 6.16: Comparison of satisfying regions for the CCMP trimer (1AQ5) using just the 49 experimental restraints versus 49 experimental + 11 noisy restraints. (a) Translation parameters. (b) Orientation parameters, projected onto a plane of theta and phi angles. (c) Translation parameters with a line from the center of each cell indicating the orientation of the symmetry axis.

algorithm with δ violations and 60 restraints increases from 0.0004 Å^2-rad^2 to 0.0081 Å^2-rad^2 as δ is increased from 4 to 9. This indicates as δ increases, the conservative nature of the core approach increases. The results clearly show that the output varies depending on the choice of δ.

We then tested the robustness to noise of the inference approach. Figure 6.15 illustrates the satisfying regions with 49 experimental and 11 noisy restraints. The distribution of backbone RMSD to the reference structure (not shown) using the inference approach continues to peak at below 1 Å, in spite of the noise. The small satisfying regions and the tightness of the RMSD distribution indicate that the constraint from the data is not obscured by the noise.

Figure 6.16 compares the satisfying regions obtained with the 49 experimental restraints versus the 60 restraints. As seen, most of the regions obtained with the 49 restraints are also obtained with the 60 restraints. The volume of the satisfying regions using the 49 restraints is 0.0188 Å^2-rad^2 while that using the 60 restraints is 0.0111 Å^2-rad^2. The volume overlap is 0.0110 Å^2-rad^2 and all regions within 10^{-4} of the maximum posterior using the 49 restraints are present in the satisfying regions obtained using the 60 restraints. This indicates the robustness of our approach to noise in the NOE data.

6.3 Conclusions

In this chapter we have developed an approach for structural inference of symmetric homo-oligomers that is complete and robust to noise and uncertainty in the experimental data.

We showed from our results on MinE that our approach is robust to as many as 22 false positives in the data. We also showed that our structural inference approach allows us to give a probability measure to each structure, hence allowing for a more accurate assessment of structural constraint from the data. Using our approach we can identify all conformations within a threshold of the conformation that best satisfies the data and biophysical constraints.

7. SUMMARY AND FUTURE WORK

This thesis has developed a method for structure determination of symmetric homo-oligomers that is (1) complete in that it identifies all conformations (within a user-defined similarity level) which are consistent with NOE restraints and which display high-quality vdW packing; (2) efficient in that by following a configuration space-based approach, it avoids explicit enumeration of all conformations; (3) data-driven in that the structural constraint from data and modeling can be separately quantified; (4) robust in that it can deal with noise and ambiguity in the input. We showed that our approach is particularly important in cases where the data available is sparse and ambiguous, where relying on an incomplete, biased search may result in missing well-packed, satisfying conformations. Extending our core algorithm of searching the four-dimensional symmetry configuration space to handle ambiguity in NOE data, uncertainty in the side-chains of subunit structure and noise or uncertainty in the NOE data makes our approach applicable to the structure determination of a broad range of homo-oligomeric complexes. By using our approach, one can accurately identify the information content in the data. We show from our results on test cases that the number of restraints is not a very good indication of the information in the data. Apart from identifying the three-dimensional positions of the atoms, our approach also enables

identification of pair-wise interactions between atoms on adjacent subunits.

The developed approach allows for a number of directions for future research which we present in the remainder of this chapter.

7.1 Future Work

7.1.1 Complete SCS Search

Currently, we perform a hierarchical subdivision of the entire configuration space. Other approaches to search through the space might be possible. A "flood-fill" kind of an approach in the configuration space could be used if we know a point in the configuration space (a "seed point") representing a structure that is consistent with both data and biophysical constraints. Regions around the seed point are tested until cells with no solution are obtained. This approach would work if all WPS structures are guaranteed to be in one contiguous region in the configuration space surrounding the structure. If not, one seed point from every contiguous region must be known. It might be possible to obtain these seed points using heuristic approaches such as simulated annealing.

Our current analytical bounds provide bounds in conformation space, given a region in the configuration space. Another way to look at this is to try to identify bounds in configuration space, given a region in conformation space. This kind of an *inverse map* could enable for more efficient bounds and better information content analysis. Given a restraint of the form $\|\mathbf{p} - \mathbf{q}'\| \leq d$, where $\mathbf{p}$ is an atom on the fixed subunit and $\mathbf{q}'$ is an

atom on the adjacent subunit, the set of possible positions for $\mathbf{q}'$ to satisfy the restraint is in a ball of radius d around $\mathbf{p}$. The inverse map aims to find regions in configuration space that would place $\mathbf{q}'$ in the ball. The inverse map would hence enable identification of regions in the configuration space corresponding to each restraint and would allow for quantifying the information content in each restraint.

Our branch-and-bound approach currently uses conservative analytical bounds and simple branching techniques. Tighter bounds and better partitioning techniques will allow a more efficient search of SCS. At each node in our search, our current tests for restraint satisfaction are conservative, in that we individually test whether for each restraint there exists at least one point in the cell, such that the structure represented by the point satisfies the restraint. A better bound would be to test for *simultaneous* restraint satisfaction by checking whether there exists at least one point in the cell, that simultaneously satisfies all the restraints.

In our current method to determine oligomeric number (Chapter 3), we run our branch-and-bound search on each different putative oligomeric number. One way to extend this is to incorporate the oligomeric number into the search space and perform our search in the ESCS, $\mathbb{Z}_9 \times S^2 \times \mathbb{R}^2$ rather than using 8 sequential searches of the SCS, $S^2 \times \mathbb{R}^2$.

Currently we provided extensions of the core algorithm to handle ambiguity, noise and side-chain uncertainty independent of the others. Algorithms that simultaneously consider all kinds of uncertainty in input and perform structural inference could be developed by combining the extensions.

7.1.2 Extensions to Other Protein Complexes

The developed approach is applicable to any C_n symmetric homo-oligomer irrespective of its size, given the subunit structure. Our approach could be extended to handle other kinds of symmetry, such as a dimer of dimers, a trimer of dimers, or improper symmetry, by defining appropriate configuration spaces and searching them in an analogous manner. For instance, in the case of dimer of dimers (D_2 symmetry), the space of symmetry axes would be represented by seven dimensions (four dimensions for one symmetry axis and three dimensions for the symmetry axis perpendicular to the first), and bounds analogous to the bounds we have previously obtained for the 4-dimensional case (Chapter 3) would have to be developed. The developed bounds would then be used in the hierarchical subdivision of the configuration space.

Similarly, we could apply our approach to docking subunits that have no symmetry. In this case, we have a six-dimensional configuration space. By formulating the six-dimensions as three translation and three rotation parameters, with an axis-angle representation for the rotation parameters we can extend and use our developed analytical bounds (Section 3.1.1). As the number of dimensions increase, the complexity increases and calls for more efficient branching and tighter bounding techniques.

7.1.3 Extensions to Other Experimental Data

The only experimental data the developed approach uses is inter-subunit NOEs. Proximity information from other experiments such as mutagenesis and chemical shift perturbation

could be used instead. The proximity information can be formulated as a set of "or" distance restraints—every atom will have a distance restraint to one of several atoms rather than just to one atom. We then eliminate cells based on violation of any of the "or" restraints.

Residual dipolar couplings (RDCs) are another type of experimental information that could be used. When the dynamics of a symmetric homo-oligomer allow determination of high-quality RDCs, the orientation of the symmetry axis lies along the director of the alignment tensor and can be easily obtained. The structure determination problem then reduces from a 4D search in $S^2 \times \mathbb{R}^2$ to a search in the 2D translation space, $\mathbb{R}^2$, and can be efficiently solved.

7.1.4 Other Extensions

Our approach for quantifying information content in the experimental data could be used to plan future experiments. Our approach enables us to identify the structural constraint provided by a set of restraints. Using this, we could identify residues in need of additional constraint to obtain better precision in the structures, and plan experiments involving these residues using isotopic labeling strategies.

Currently, we assume that we know the backbone of the subunit structure when it is in complex and that it is not flexible. It is possible to extend the approach to account for flexibility in the subunit structure by establishing additional dimensions that represent concerted motions in the backbone.

Bibliography

[1] R. Abagyan, M. Totrov, and D. Kuznetsov. ICM - a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem*, 15(5):488–506, 2004.

[2] P. D. Adams, I. T. Arkin, D. M. Engelman, and A. T. Brünger. Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat. Struct. Biol.*, 2:154–62, 1995.

[3] P. D. Adams, D. M. Engelman, and A. T. Brünger. Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins*, 26:257–61, 1996.

[4] A. Aho, J. E. Hopcroft, and J. D. Ullman. *Data Structures and Algorithms*. Addison-Wesley, 1982.

[5] P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, and L. Serrano. Structure-based assembly of protein complexes in yeast. *Science*, 303:2026–2029, 2004.

[6] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, 332:989–998, 2003.

[7] P. Aloy and R. B. Russell. The third dimension for protein interactions and complexes. *Trends Biochem. Sci.*, 27:633–638, 2002.

[8] E. Althaus, O. Kohlbacher, H. P. Lenhof, and P. Muller. A combinatorial approach to protein docking with flexible side chains. *J. Comp. Biol.*, 9(4):587–612, 2002.

[9] G. S. Anand, D. Law, J. G. Mandell, A. N. Snead, I. Tsigelny, S. S. Taylor, L. F. T. Eyck, and E. A. Komives. Identification of the protein kinase A regulatory R-I alpha-catalytic subunit interface by amide H/2H exchange and protein docking. *PNAS*, 100:13264–13269, 2003.

[10] G. Ausiello, G. Cesareni, and M. Helmer-Citterich. ESCHER: A new docking procedure applied to the reconstruction of protein tertiary structure. *Proteins*, 28:556567, 1997.

[11] Y. Azuma, L. Renault, J. A. Garcia-Ranea, A. Valencia, T. Nishimoto, and A. Wittinghofer. Model of the RanRCC1 interaction using biochemical and docking experiments. *J. Mol. Biol.*, 289:11191130, 1999.

[12] J. W. Back, L. de Jong, A. O. Muijsers, and C. G. de Koster. Chemical cross-linking and mass spectrometry for protein structural modeling. *J. Mol. Biol.*, 331:303–313, 2003.

[13] D. J. Bacon and J. Moult. Docking by least squares fitting of molecular surface patterns. *J. Mol. Biol.*, 225:849858, 1992.

[14] W. Baumeister. Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Curr. Opin. Struct. Biol.*, 12:679–684, 2002.

[15] W. Baumeister, R. Grimm, and J. Walz. Electron tomography of molecules and cells. *Trends Cell Biol*, 9:81–85, 1999.

[16] E. Ben-Zeev and M. Eisenstein. Weighted geometric docking: incorporating external information in the rotation-translation scan. *Proteins*, 52:2427, 2003.

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[18] M. J. Bloomers, C. B. Lucasius, G. Kateman, and R. Kaptein. Conformational analysis of a di-nucleotide photodimer with the aid of genetic algorithm. *Biopolymers*, 32:4552, 1992.

[19] A. Bohm, J. Diez, K. Diederichs, W. Welte, and W. Boos. Structural model of malk, the ABC subunit of the maltose transporter of *Escherichia coli*: implications for mal gene regulation, inducer exclusion, and subunit assembly. *J. Biol. Chem.*, 277(5):3708–17, 2002.

[20] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing Inc., 1998.

[21] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4(2):187–217, 2004.

[22] A. T. Brünger. *XPLOR: A system for X-ray crystallography and NMR*. Yale University Press:New Haven, 1993.

[23] A. T. Brünger, 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.

[24] A. T. Brünger and M. Nilges. Computational challenges for macromolecular structure determination by x-ray crystallography and solution NMR spectroscopy. *Quart. Rev. Biophys*, 26:49–125, 1993.

[25] M. Caffrey. Membrane protein crystallization. *J. Struct. Biol.*, 7:697–701, 1997.

[26] D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R. Woods. The AMBER biomolecular simulation programs. *J. Comp. Chem.*, 26:1668–1688, 2005.

[27] M. H. Chen, Q. M. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer Verlag Inc., New York, 2002.

[28] R. Chen, L. Li, and Z. Weng. ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, 52:80–87, 2003.

[29] J. B. Clarage, T. Romo, B. K. Andrews, B. M. Pettitt, and G. N. Phillips Jr. A sampling problem in molecular dynamics simulations of macromolecules. *PNAS*, 92:32883292, 1995.

[30] H. Claussen, C. Buning, M. Rarey, and T. Lengauer. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.*., 308(2):377–395, 2001.

[31] G. M. Clore and C. D. Schwieters. Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1H-15N chemical shift mapping and backbone 15N-1H residual dipolar couplings using conjoined rigid body torsion angle dynamics. *JACS*, 125:29022912, 2003.

[32] S. R. Comeau and C. J. Camacho. Predicting oligomeric assemblies: N-mers a primer. *J. Struct. Biol.*, 150(3):233–44, 2005.

[33] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *JACS*, 117:5179–5197, 1995.

[34] P. B. Crowley, G. Otting, B. G. Schlarb-Ridley, G. W. Canters, and M. Ubbink. Hydrophobic interactions in a cyanobacterial plastocyanin-cytochrome f complex. *JACS*, 123:1044410453, 2001.

[35] B. C. Cunningham, P. Jhurani, P. Ng, and J. A. Wells. Receptor and antibody epitopes in human growth hormone identified by homolog-scanning mutagenesis. *Science*, 243:1330–1336, 1989.

[36] M. de Berg, M. van Krevald, M. Overmars, and O. Schwarzkopf. *Computational Geometry*. Springer-Verlag, 2000.

[37] J. Desmet, M. DeMaeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[38] A. Dobrodumov and A. M. Gronenborn. Filtering and selection of structural models: combining docking and NMR. *Proteins*, 53:18–32, 2003.

[39] C. Dominguez, R. Boelens, and A. M. Bonvin. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *JACS*, 125:1731–1737, 2003.

[40] D. Duhovny, Y. Inbar, V. Polak, M. Shatsky, I. Halperin, A. Benyamini, A. Barzilai, O. Dror, N. Haspel, and R. Nussinov. Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins*, 52:107–112, 2003.

[41] D. Duhovny, R. Nussinov, and H. J. Wolfson. Efficient unbound docking of rigid molecules. In Gusfield et al., editor, *Proceedings of the 2'nd Workshop on Algorithms in Bioinformatics (WABI)*, Lecture Notes in Computer Science 2452, pages 185–200. Springer Verlag, 2002.

[42] H. Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer-Verlag, 1987.

[43] S. Elliott, T. Lorenzini, D. Chang, J. Barzilay, E. Delorme, J. Giffin, and Hesterberg L. Fine-structure epitope mapping of antierythropoietin monoclonal antibodies reveals a model of recombinant human erythropoietin structure. *Blood*, 87(7):2702–13, 1996.

[44] C. Ericson. *Real-time collision detection*. Morgan Kaufmann, 2005.

[45] A. Fahmy and G. Wagner. TreeDock: a tool for protein docking based on minimizing van der Waals energies. *JACS*, 124:12411250, 2002.

[46] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, 45 (S5):157–162, 2002.

[47] H. Fasold, J. Klappenberger, C. Meyer, and H. Remold. Bifunctional reagents for the crosslinking of proteins. *Angew Chem Int*, 10:795–801, 1971.

[48] J. Fernandez-Recio, M. Totrov, and R. Abagyan. Soft protein-protein docking in internal coordinates. *Protein Science*, 11:280–291, 2002.

[49] F. Fiorito, S. Hiller, G. Wider, and K. Wüthrich. Automated resonance assignment of proteins: 6D APSY-NMR. *J. Biomol. NMR*, 35:27–37, 2006.

[50] J. Flaux, E. B. Bertelsen, A. L. Horwich, and K. Wüthrich. NMR analysis of a 900k GroEL-GroES complex. *Nature*, 418:207–211, July 2002.

[51] R. H. A. Folmer, M. Nilges, C. H. M. Papavoine, B. J. M. Harmsen, R. N. H. Konings, and C. W. Hilbers. Refined structure, DNA binding studies, and dynamics of

the bacteriophage Pf3 encoded single-stranded DNA binding protein. *Biochemistry*, 36:91209135, 1997.

[52] H. A. Gabb, R. M. Jackson, and M. J. Sternberg. Modeling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, 272:106–120, 1997.

[53] C. Gaboriaud, J. Juanhuix, A. Gruez, M. Lacroix, C. Darnault, D. Pignol, D. Verger, J. C. Fontecilla-Camps, and G. J. Arlaud. The crystal structure of the globular head of complement protein C1q provides a basis for its versatile recognition properties. *J. Biol. Chem.*, 278:4697446982, 2003.

[54] H. Gao, J. Sengupta, M. Valle, A. Korostelev, N. Eswar, S. M. Stagg, P. V. Roey, R. K. Agrawal, S. C. Harvey, and A. Sali A et al. Study of the structural dynamics of the *Escherichia coli* 70S ribosome using real-space refinement. *Cell*, 113:789–801, 2003.

[55] E. J. Gardiner, P. Willett, and P. J. Artymiuk. Graph-theoretic techniques for macromolecular docking. *J. Chem. Inform. Comput. Sci.*, 40:273279, 2000.

[56] R. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66:13351340, 1994.

[57] D. S. Goodsell, G. M. Morris, and A. J. Olson. Automated docking of flexible ligands: applications of AUTODOCK. *J. Mol. Recog.*, 9:15, 1996.

[58] K. E. Gottschalk, M. Soskine, S. Schuldiner, and H. Kessler. A structural model of EmrE, a multi-drug transporter from *Escherichia coli*. *Biophys. J.*, 86:33353348, 2004.

[59] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, 331:281–299, 2003.

[60] J. J. Gray, S. E. Moughon, T. Kortemme, O. Schueler-Furman, K. M. Misura, A.V. Morozov, and D. Baker. Protein-protein docking predictions for the CAPRI experiment. *Proteins*, 52:118–122, 2003.

[61] K. Grunewald, P. Desai, D. C. Winkler, J. B. Heymann, D. M. Belnap, W. Baumeister, and A. C. Steven. Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science*, 302:1396–1398, 2003.

[62] P. Güntert. Automated NMR protein structure calculation with CYANA. *Meth. Mol. Biol.*, 278:353–378, 2004.

[63] P. Güntert, W. Braun, and K. Wüthrich. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J. Mol. Biol.*, 217:517–530, 1991.

[64] P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.*, 273:283–298, 1997.

[65] M. Habeck, M. Nilges, and W. Rieping. Replica-exchange monte carlo schemce for bayesian data analysis. *Phys. Rev.*, 94:01805, 2005.

[66] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, 47:409–443, 2002.

[67] J. Han and K. Micheline. *Data Mining Concepts and Techniques*. Morgan Kaufmann, 2001.

[68] K. K. Han, C. Richard, and A. Delacourte. Chemical crosslinks of proteins by using bifunctional reagents. *Int. J. Biochem.*, 16:129–145, 1984.

[69] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans on SSC*, 4:100–114, 1968.

[70] D. R. Haudenschild, M. M. Tondravi, U. Hofer, Q. Chen, and P. F. Goetinck. The role of coiled-coil helices and disulfide bonds in the assembly and stabilization of cartilage matrix protein subunits. A mutational analysis. *J. Biol. Chem.*, 270:23150–23154, 1995.

[71] T. Herrmann, P. Güntert, and K. Wüthrich. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.*, 319:209–227, 2002.

[72] K. C. Holmes, I. Angert, F. J. Kull, W. Jahn, and R. R. Schroder. Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature*, 425:423–427, 2003.

[73] Y. J. Huang, G. V. Swapna, P. K. Rajan, H. Ke, B. Xia, K. Shukla, M. Inouye, and G. T. Montelione. Solution NMR structure of ribosome-binding factor A (RbfA), a cold-shock adaptation protein from *Escherichia coli*. *J. Mol. Biol.*, 327:521536, 2003.

[74] R. M. Jackson, H. A. Gabb, and M. J. Sternberg. Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J. Mol. Biol.*, 276:265285, 1998.

[75] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. CAPRI: a critical assessment of predicted interactions. *Proteins*, 52:2–9, 2003.

[76] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620, 1957.

[77] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge Univ. Press, Cambridge, 2003.

[78] H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proc. Roy. Soc. A*, 186, pages 453–461, 1946.

[79] W. Jiang, Z. Li, Z. Zhang, M. L. Baker, P. E. Prevelige Jr, and W. Chiu. Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nat. Struct. Biol.*, 10:131–135, 2003.

[80] G. Jones, P. Willet, R. Glen, A. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727748, 1997.

[81] R. L. Dunbrack Jr. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.*, 12(4):431–40, 2002.

[82] R. L. Dunbrack Jr. and M. Karplus. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[83] K. Juszewski, C. D. S. Schwieters, D. S. Garrett, R. A. Byrd, N. Tjandra, and G. M. Clore. Completely automated, highly error-tolerant macromolecular structure determination from multi-dimensional nuclear Overhauser enhancement spectra and chemical shift assignments. *JACS*, 126:62586273, 2004.

[84] A. Kariakin, D. Davydov, J. A. Peterson, and C. Jung. A new approach to the study of protein-protein interaction by FTIR: complex formation between cytochrome P450BM-3 heme domain and FMN reductase domain. *Biochemistry*, 41:13514–13525, 2002.

[85] M. Karplus and G. A. Petsko. Molecular dynamics simulations in biology. *Nature*, 347:631–639, 1990.

[86] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *PNAS*, 89:2195–2199, 1992.

[87] L. E. Kay, G. M. Clore, A. Bax, and A. M. Gronenborn. Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1 beta in solution. *Science*, 249:364–5, 1990.

[88] S. D. Khare, K. C. Wilcox, P. Gong, and N. V. Dokholyan. Sequence and structural determinants of Cu, Zn superoxide dismutase aggregation. *Proteins*, 61:617–632, 2005.

[89] G. F. King, Y. L. Shih, M. W. Maciejewski, N. P. Bains, B. Pan, S. L. Rowland, G. P. Mullen, and L. I. Rothfield. Structural basis for the topological specificity function of MinE. *Nat. Struct. Biol.*, 7:1013–1017, 2000.

[90] R. Knegtel, I. Kuntz, and C. Oshiro. Molecular docking to ensembles of protein structures. *J. Mol. Biol*, 266:424440, 1997.

[91] B. Kramer, G. Metz, M. Rarey, and T. Lengauer. Ligand docking and screening with FlexX. *Med. Chem. Res.*, 7/8:463478, 1999.

[92] I. Kwaw, J. Sun, and H. R. Kaback. Thiol cross-linking of cytoplasmic loops in lactose premease of *Escherichia coli*. *Biochemistry*, 39:3134–3140, 2000.

[93] D. B. Lacy, M. Mourez, A. Fouassier, and R. J. Collier. Mapping the anthrax protective antigen binding site on the lethal and edema factors. *J. Biol. Chem.*, 277(4):3006–10, 2002.

[94] J. Lanman, T. T. Lam, S. Barnes, M. Sakalian, M. R. Emmett, A. G. Marshall, and P. E. Prevelige Jr. Identification of novel interactions in HIV-1 capsid protein assembly by high-resolution mass spectrometry. *J. Mol. Biol.*, 325:759–772, 2003.

[95] A. R. Leach. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, 235:345356, 1994.

[96] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–39, 1998.

[97] O. S. Lequin, D. Staunton, B. Mulloy, M. J. Forster, K. Yoshida, and I. D. Campbell. Mapping the heparin-binding site on the (1314), F3 fragment of fibronectin. *J. Biol. Chem.*, 277:5062950635, 2002.

[98] R. H. Lilien, C. Bailey-Kellogg, A. A. Anderson, and B. R. Donald. A subgroup algorithm to identify cross-rotation peaks consistent with non-crystallographic symmetry. *Acta Crystallographica D: Biological Crystallography*, pages D60:1057–1067, 2004.

[99] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *JACS*, 104(17):4546 – 4559, 1982.

[100] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.

[101] K.R. MacKenzie, J.H. Prestegard, and D.M. Engelman. Leucine side-chain rotamers in a glycophorin A transmembrane peptide as revealed by three-bond carbon-carbon couplings and 13C chemical shifts. *J. Biomol. NMR*, 7:256260, 1996.

[102] A. D. Mackerell. Empirical force fields for biological macromolecules: overview and issues. *J. Comp. Chem.*, 25:15841604, 2004.

[103] S. Macura and R. R. Ernst. Elucidation of cross relaxation in liquids by two-dimensional NMR spectroscopy. *Mol. Phys.*, 41:95–117, 1980.

[104] J. G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and L. F. Ten Eyck. Protein docking using continuum electrostatics and geometric fit. *Protein Eng.*, 14:105–113, 2001.

[105] J. A. Marquez, C. I. Smith, M. V. Petoukhov, P. Lo Surdo, P. T. Mattsson, M. Knekt, A. Westlund, K. Scheffzek, M. Saraste, and D. I. Svergun. Conformation of full-length bruton tyrosine kinase (Btk) from synchrotron X-ray solution scattering. *EMBO J.*, 22:4616–4624, 2003.

[106] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annual Rev. Biophys. Biomol. Struct.*, 29:291–325, 2000.

[107] T. Matsuda, T. Ikegami, N. Nakajima, T. Yamazaki, and H. Nakamura. Model building of a protein-protein complexed structure using saturation transfer and residual dipolar coupling without paired intermolecular NOE. *J. Biomol. NMR*, 29:325338, 2004.

[108] M. Y. Mizutani, N. Tomioka N, and A. Itai. Rational automatic search method for stable docking models of protein and ligand. *J. Mol. Biol.*, 243:310326, 1994.

[109] X. J. Morelli, P. N. Palma, F. Guerlesquin F, and A. C. Rigby. A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data. *Protein Science*, 10:21312137, 2001.

[110] M. Morillas, P. Gomez-Puertas, B. Rubi, J. Clotet, J. Arino, A. Valencia, F. G. Hegardt, D. Serra, and G. Asins. Structural model of a malonyl- CoA-binding site of carnitine octanoyltransferase and carnitine palmitoyltransferase I: mutational analysis of a malonyl-CoA affinity domain. *J. Biol. Chem.*, 277:11473–11480, 2002.

[111] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.*, 19:1639–1662, 1998.

[112] M. Nilges. A calculation strategy for the structure determination of symmetric dimers by 1H NMR. *Proteins*, 17(3):297–309, 1993.

[113] M. Nilges, M. Macais, S. Odonoghue, and H. Oschkinat. Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *J. Mol. Biol.*, 269:408–422, 1997.

[114] E. Nogales, S. G. Wolf, and K. H. Downing. Structure of the alpha beta tubulin dimer by electron crystallography. *Nature*, 391:199–203, 1998.

[115] R. Norel, D. Petrey, H. J. Wolfson, and R. Nussinov. Examination of shape complementarity in docking of unbound proteins. *Proteins*, 36:307–317, 1999.

[116] S. I. O'Donoghue, X. Chang, R. Abseher, M. Nilges, and J. J. Led. Unraveling the symmetry ambiguity in a hexamer: calculation of the R6 human insulin structure. *J. Biomol. NMR*, 16(2):93–108, 2000.

[117] S. I. O'Donoghue, F. K. Junius, and G. F. King. Determination of the structure of symmetric coiled-coil proteins from NMR data: application of the leucine zipper proteins Jun and GCN4. *Protein Eng.*, 6(6):557–564, 1993.

[118] C. Ostermeier and H. Michel. Crystallization of membrane proteins. *Curr. Opin. Struct. Biol.*, 142:108–132, 2003.

[119] K. Oxenoid and J. J. Chou. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *PNAS*, 102:10870–10875, 2005.

[120] P. N. Palma, L. Krippahl, J. E. Wampler, and J. J. G. Moura. BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins*, 39:372384, 2000.

[121] B. Pierce and Z. Weng. M-ZDOCK: A grid-based approach for C_n symmetric multimer docking. *Bioinformatics*, 21(8):1472–1476, 2005.

[122] Introduction to protein interactions. http://www.piercenet.com.

[123] J. W. Ponders and F. M. Richards. Tertiary templates for proteins. *J. Mol. Biol.*, 193:775–791, 1987.

[124] 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 and van der Waals packing. *Proteins*, 65(1):203–219, 2006.

[125] M. Rarey, B. Kramer, and T. Lengauer. Docking of hydrophobic ligands with interaction-based matching algorithms. *Bioinformatics*, 15:243250, 1999.

[126] M. Revington, A. Semesi, A. Yee, and G. S. Shaw. Solution structure of the *Escherichia Coli* protein ydhR: A putative mono-oxygenase. *Protein Science*, 14:3115–3120, 2005.

[127] W. Rieping, M. Habeck, and M. Nilges. Modeling errors in NOE data with a log-normal distribution improves the quality of NMR structures. *A. Am. Chem. Soc.*, 127:16026–16027, 2005.

[128] D. W. Ritchie and G. J. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins*, 39:178–194, 2000.

[129] L. C. Roisman, J. Piehler, J. Y. Trosset, H. A. Scheraga HA, and G. Schreiber. Structure of the interferon-receptor complex determined by distance constraints from double- mutant cycles and flexible docking. *PNAS*, 98:1323113236, 2001.

[130] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.*, 14:313–324, 2004.

[131] B. R. Seavey, E. A. Farr, W. M. Westler, and J. Markley. A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, 1:217236, 1991.

[132] G. Serino, H. Su, Z. Peng, T. Tsuge, N. Wei, H. Gu, and X. W. Deng. Characterization of the last subunit of the arabidopsis COP9 signalosome: implications for the overall structure and origin of the complex. *Plant Cell*, 15:719–731, 2003.

[133] G. R. Smith and M. J. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, 12:28–35, 2002.

[134] D. I. Svergun, I. Aldag, T. Sieck, K. Altendorf, M. H. Koch, D. J. Kane, M. B. Kozin, and G. Gruber. A model of the quaternary structure of the *Escherichia coli* F1 ATPase from X-ray solution scattering and evidence for structural changes in the delta subunit during ATP hydrolysis. *Biophys. J.*, 75:2212–2219, 1998.

[135] J. B. Swaney. Use of cross-linking reagents to study lipoprotein structure. *Methods Enzymol.*, 128:613–626, 1986.

[136] A. Tovchigrechko, C. A. Wells, and I. A. Vakser. Docking of protein models. *Protein Science*, 11:1888–1896, 2002.

[137] M. Trester-Zedlitz, K. Kamada, S. K. Burley, D. Fenyo, B. T. Chait, and T. W. Muir. A modular cross-linking approach for exploring protein interactions. *JACS*, 125:2416–2425, 2003.

[138] K. Truong and M. Ikura. The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo. *Curr. Opin. Struct. Biol.*, 11:573–578, 2001.

[139] M. Ubbink, M. Ejdeback, B. G. Karlsson, and D. S. Bendall. The structure of the complex of plastocyanin and cytochrome *f*, determined by paramagnetic NMR and restrained rigid-body molecular dynamics. *Structure*, 6:323335, 1998.

[140] I. A. Vakser. Protein docking for low-resolution structures. *Protein Eng.*, 8:371–377, 1995.

[141] I. A. Vakser, O. G. Matar, and C. F. Lam. A systematic study of low-resolution recognition in protein-protein complexes. *PNAS*, 96:84778482, 1999.

[142] K. J. Walters, H. Matsuo, and G. Wagner. A simple method to distinguish inter-monomer nuclear Overhauser effects in homodimeric proteins with C_2 symmetry. *JACS*, 119:5958–5959, 1997.

[143] C. Wang, O. Schueler-Furman, and D. Baker. Improved side-chain modeling for protein-protein docking. *Protein Science*, 14(5):1328–39, 2005.

[144] 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.

[145] L. Wang and B. R. Donald. An efficient and accurate algorithm for assigning nuclear Overhauser effect restraints using a rotamer library ensemble and residual dipolar couplings. In *Proc. IEEE Comput. Syst. Bioinform. Conf. (CSB2005)*, pages 189–202, August 2005.

[146] Z. R. Wasserman and C. N. Hodge. Fitting an inhibitor into the active site of thermolysin: a molecular dynamics study. *Proteins*, 24:227237, 1996.

[147] Wikipedia. Molecular mechanics. http://en.wikipedia.org/wiki/Molecular_mechanics.

[148] R. Wiltscheck, R. A. Kammerer, S. A. Dames, T. Schulthess, M. J. Blommers, J. Engel, and A. T. Alexandrescu. Heteronuclear NMR assignments and secondary structure of the coiled coil trimerization domain from cartilage matrix protein in oxidized and reduced forms. *Protein Science*, 6:1734–1745, 1997.

[149] R. Wolfgang, H. Michael, and M. Nilges. Inferential structure determination. *Science*, 309:303–306, 2005.

[150] K. Wüthrich. *NMR of Proteins and Nucleic Acids*. Wiley, 1986.

[151] Y. Xu, J. Wu, D. Gorenstein, and W. Braun. Automated 3D assignment and structure calculation of crambin (S22/I25) with the self-correcting distance geometry based NOAH/DIAMOD programs. *J. Magn. Reson.*, 136:76–85, 1999.

[152] Y. Yan and G. Marriott. Analysis of protein interactions using fluorescence technologies. *Curr. Opin. Chem. Biol.*, 7:635–640, 2003.

[153] M. Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science*, 12:12711282, 2003.

[154] M. I. Zavodszky and L. A. Kuhn. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Science*, 14(4):1104–1114, 2005.

[155] W. Zhang, P. R. Chipman, J. Corver, P. R. Johnson, Y. Zhang, S. Mukhopadhyay, T. S. Baker, J. H. Strauss, M. G. Rossmann, and R. J. Kuhn. Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. *Nat. Struct. Biol.*, 10:907–912, 2003.

[156] C. Zwahlen, P. Legault, S. J. F. Vincent, J. Greenblatt, R. Konrat, and L. E. Kay. Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: Application to a bacteriophage λ N-Peptide/*boxB* RNA complex. *JACS*, 119:6711–6721, 1997.