5-1-2013

Geometrical and probabilistic methods for determining association models and structures of protein complexes

Himanshu Chandola
Dartmouth College

Geometrical and probabilistic methods for determining association models and structures of protein complexes
Dartmouth Computer Science Technical Report TR2013-731

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Computer Science

by

Himanshu Chandola

DARTMOUTH COLLEGE

Hanover, New Hampshire

May 2013

Examining Committee:

(chair) Chris Bailey-Kellogg

Devin Balkcom

Gevorg Grigoryan

Bruce Randall Donald

F. Jon Kull, Ph.D.
Dean of Graduate Studies

Abstract

Protein complexes play vital roles in cellular processes within living organisms. They are formed by interactions between either different proteins (hetero-oligomers) or identical proteins (homo-oligomers). In order to understand the functions of the complexes, it is important to know the manner in which they are assembled from the component subunits and their three dimensional structure. This thesis addresses both of these questions by developing geometrical and probabilistic methods for analyzing data from two complementary experiment types: Small Angle Scattering (SAS) and Nuclear Magnetic Resonance (NMR) spectroscopy. Data from an SAS experiment is a set of scattering intensities that can give the interatomic probability distributions. NMR experimental data used in this thesis is set of atom pairs and the maximum distance between them. From SAS data, this thesis determines the association model of the complex and intensities through an approach that is robust to noise and contaminants in solution. Using NMR data, this thesis computes the complex structure by using probabilistic inference and geometry of convex shapes. The structure determination methods are complete, that is they identify all consistent conformations and are data driven wherein the structures are evaluated separately for consistency to data and biophysical energy.

Acknowledgements

In the course of many years that I spent on this thesis, I have benefited from several people in both technical and non-technical aspects and I would like to thank at least some of them here.

Firstly, I would like to thank my advisor Prof. Chris Bailey-Kellogg for guiding me from working erratically to being more methodical in my research. A lot of my technical learnings outside computational biology have been due to the flexibility that Chris showed in allowing me to learn diverse areas that I got interested in. Chris's way of directing research has helped me learn how to approach solving a problem from scratch on my own.

During the course of my research I benefitted a lot from collaborations. For my work on solution scattering, Prof. Alan Friedman at Purdue helped me understand the nuances of the problem and gave valuable feedback. Prof. Bruce Craig at Purdue was a great help on anything related to Statistics on this problem. Prof. Bruce Donald at Duke was very helpful on problems related to NMR data. I especially benefitted through my conversations with Anthony Yan at Donald Lab while working on one of the NMR problems.

I would also like to thank committee members: Profs Devin Balkcom and Gevorg Grigoryan for giving valuable inputs through the process. I would also like to thank CBK lab members: Andrew, Lu, Ickwon, Tuobin and Yoonjoo for discussions on things other than proteins.

Outside lab, my friends helped in reducing the stresses of academic life. I can't thank them enough for helping me stay balanced and optimistic.

Lastly, I would like to thank my family for being supportive through this whole process.

Contents

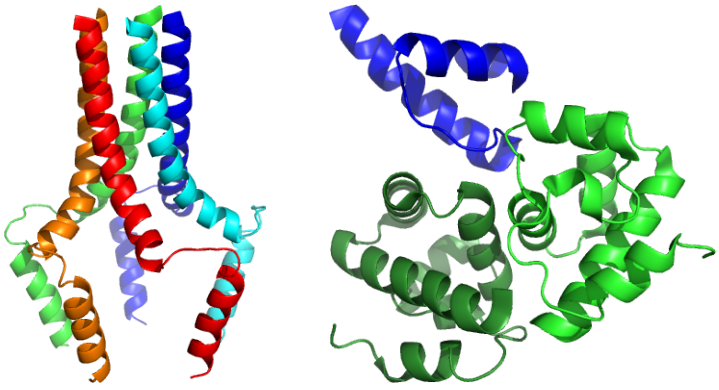
List of Tables

List of Figures

1. INTRODUCTION

Proteins are polymers of amino acids that carry out diverse functions in a living organism. The function of a protein is determined by its amino acid sequence and three dimensional structure. The sequence of amino acids in a protein is referred to as its primary structure. The sequence organizes itself into helices and sheets which are referred to as secondary structures while the three dimensional shape of a protein is called tertiary structure. Multiple proteins can interact with each other to form protein complexes (Fig. 1.1). Protein complexes are formed either out of association between identical proteins (homo-oligomers) or different proteins (hetero-oligomers). Two central problems in studies of protein complexes are determining the association models and the structure of these complexes. The association model of complex formation is given by the association pathway and the association constants of the pathway. The complex structure is given by the three dimensional atomic coordinates of the complex.

Solving these are important since the function of a protein-protein complex depends on the stoichiometry as well as the strength of association between monomers, and the structures of the monomers and the complexes. Existing approaches to determine association model, fall short of computing a structure. For example techniques like hydrogen/deuterium exchange, analytical ultracentrifugation, titration calorimetry can yield information about the nature of these associations, but are very limited in the structural information they carry. We therefore use SAS data to compute association model and low resolution structure for hetero-oligomers. In determining high resolution structure, X-ray crystallography – the most accurate structure determination method, requires the protein



Phospholamban BAF-emerin complex

Fig. 1.1: Two protein complexes in nature: Phospholamban is a symmetric pentamer while BAF-emerin complex is a hetero-oligomer consisting of a dimer (green, shades of green denote individual monomers in the dimer) and a monomer (blue).

to be crystallized which can be hard for these complexes due to weak and transient interactions. For computing protein structure of homo-oligomers, therefore, we use data from NMR experiments.

1.1 Association model

Data from an SAS experiment can be represented as a ”scattering curve” capturing the intensity of scattering by angle of x-rays or neutrons. The scattering curve contains important information about a protein, including sufficient inter-atomic distance information to allow the computation of a low resolution structure.

A scattering curve is a series of diffraction intensities measured at various angles. A solution containing a protein complex contains the subunits and the protein complex as distinct chemical forms. The scattering curve obtained from such a solution is a linear combination of the scattering curves of the individual forms. Deconvolving the measured curve can yield scattering curves of the individual forms at equilibrium, as well as their mass ratios. This in turn enables the determination of the underlying association model. In this thesis, methods are presented to find the association model and scattering curves by iterating over possible association models and scoring them on two metrics: statistical and

physical. The model that scores the best is predicted as the correct model.

Statistical metric We want to find the association model and scattering curves of individual forms for which the reconstructed oligomer scattering curves best fit the experimentally measured data. We can quantify this distance by χ^2 scores that evaluate the fit of the reconstructed data with the experimental data. The scores can then be used to discriminate between various models.

Physical metric In determining the quality of an association model, we would like to strike a balance between the fit of reconstructed data and their physically plausibility. To evaluate physical plausibility, we define a score that is computed from the scattering curves. The scattering intensity at zero angle can be computed from a scattering curve. The scattering intensity at zero angle is proportional to the molecular weight of the chemical form. We use it to to compute a physical score that can be used in conjunction with the statistical score to evaluate a model and eliminate models that score high on statistical score but are not physically feasible.

Robust to noise Experimental data contains noise and we do not want it to result in the selection of an incorrect model. The noise can either be a random noise (e.g. due to counting error) or a systematic noise (e.g. due to a contaminant).

For random noise, we use a decomposition from linear algebra that filters the data on the assumption that the measured data is a linear combination of independent curves and the noise is Gaussian. We then use the denoised data to find the association model.

We consider a contaminant to be a low quantity of either a non-participating monomer or a homo-oligomeric aggregate. To account for contaminant, we extend the model to estimate the scattering curves through an optimization procedure.

1.2 Homo-oligomeric structure from NMR

In an NMR Nuclear Overhauser Enhancement Spectroscopy (NOESY) experiment, interactions (due to the Nuclear Overhauser Effect or NOE) are obtained for pairs of interacting nuclei that are close in space. The NOE peaks are represented as distance restraints, giving lower and upper bounds on distances between the interacting nuclei. Our goal is to compute the three dimensional structure of the homo-oligomer from a set of NOE distance restraints. The determination of the structure is complicated by the fact that the restraints suffer from ambiguit and noise, and only provide distance estimates, consequently more than one structure can satisfy the data.

This thesis works with symmetrical homo-oligomers - complexes in which the subunits are arranged symmetrically around an axis. This leads to a lower dimensional representation of the protein structure but adds an ambiguity in interpretation of distance restraints. With NMR data as input, the two problems solved in this thesis are NMR structural inference and structural determination from ambiguous restraints.

NMR distance restraints suffer from noise and sparsity which results in uncertainty in the determined structures. A Bayesian structural inference approach can quantify the uncertainty in the resulting structures but computing the moments of distribution accurately is a hard problem for which at best only heuristics are at place that are unable to give guarantees. In the special case when an individual subunit's structure is available and the inter-subunit distance restraints are known, the symmetric homo-oligomer can be represented as a four dimensional variable. By using geometric bounds, bounds on probability mass and expectation are computed with guarantee of accuracy. Ultimately, this enables us to do inference with guarantees on the computed metrics.

In general, however, the determination of inter-subunit restraints for a given symmetric homo-oligomer is hard to do and can only work with a small set of homo-oligomers with a specialized set of experiments. The NMR distance restraints, in almost all cases, are not classified as inter-subunit or intra-subunit. Structure determination from such data is

hard, especially when coupled together with a physical energy function which has a rugged landscape. This thesis focuses on just the NMR distance restraints, leaving the physical energy as a possible post-processing step. The problem is still hard, given the exponential combinations of possible interpretations of the set of restraints. In the second work with NMR data, we work on computing protein structure from such data. From geometric results on convex domains of the configuration space, we design an algorithm that does not require us to compute all possible interpretations of the set of restraints. As a result, we are able to design an algorithm that gets conformations consistent with the data but is more efficient than enumeration of all possible interpretations of the data.

Complete Current methods to compute a protein structure that satisfies NOE restraints rely on heuristic searches that try to find the protein structure that has the lowest energy according to an pseudo-energy function combining physical energy and restraint terms. It is not possible to find the global minimum of this function through analytical techniques and it is hard to tell whether a given point is global minimum. If there is sufficient data available then we have a highly constrained system that often converges to the global minimum. On the other hand if the restraints do not yield a tightly constrained system, the searches may not converge or they can yield a local minimum.

In this thesis, the methods for homo-oligomeric structure determination are complete in nature. In our work with inferential structure determination, we compute the expectation of the 3D coordinates of the complex with an error guarantee. We also compute the probability mass of a given set of configurations correct up to a user defined error cutoff. In our work in structure determination from ambiguous data, we compute all conformations up to a user defined resolution.

Data driven Methods to obtain protein structure from NOE data take both data and biophysical modeling of the energy to find the best structure. The biophysical model involves selecting parameters that can bias the final structure. Also, the minimum of the energy

function, if found, is through heuristic approaches without any guarantees. By focusing on obtaining structures from data alone and leaving detailed biophysical modeling as a post processing step for the end user, we present an objective approach to finding the structures from NOE data. We incorporate basic biophysical modeling by eliminating structures that have guaranteed steric clashes.

In our inferential structure determination work, we use geometric and probabilistic approaches that are used in computing bounds on the probability mass and the expectation integral. These bounds are useful in the pruning criterion used in the algorithm.

In our work with structure determination from ambiguous NMR data, we used geometrical approaches to characterize the Minkowski sums of convex shapes and their intersections. The intersections between these shapes were used to determine the number of restraint violations, which was used as a pruning criterion in the algorithm.

2. STOICHIOMETRIES AND AFFINITIES OF INTERACTING PROTEINS FROM CONCENTRATION SERIES OF SOLUTION SCATTERING DATA: DECOMPOSITION BY LEAST SQUARES AND QUADRATIC OPTIMIZATION

Abstract

In studying interacting proteins, complementary insights are provided by analyzing the association model (the stoichiometry and affinity constants of the intermediate and final complexes) and the quaternary structure of the resulting complexes. Many current methods for analyzing protein interactions give either a binary answer to the question of association or at best provide only part of the complete picture. We present here a method to extract both types of information from x-ray or neutron scattering data for a series of solutions containing the complex components in different concentrations. Our method determines the association pathway and constants, along with the scattering curves of the individual members of the mixture, so as to best explain the scattering data for the set of mixtures. The derived curves then enable reconstruction of the intermediate and final complexes. Using a new analytic method, we also extend our approach to evaluate the association models and scattering curves in the presence of contaminants, teseting both a non-participating monomer and a large homo-oligomeric aggregate. Using simulated solution scattering data for four hetero-oligomeric complexes with different structures, molecular weights, and association

models, we demonstrate that our method accurately determines the simulated association model and monomer scattering profiles. We also demonstrate that the method is robust to both random noise and systematic noise from such contaminants, and is applicable over a large range of weak association constants typical of transient protein-protein complexes.

2.1 Introduction

In order to gain deeper understanding into the functions and mechanisms of protein-protein interactions, it is necessary to extend the binary information (interaction or not) provided by high-throughput techniques, and characterize the stoichiometries, affinities, and three-dimensional structures of protein complexes. However, experimental methods for detailed studies of protein complexes typically fall into two separate categories: some (e.g., x-ray crystallography and NMR spectroscopy) enable structure determination but do not readily reveal the association model, while others (e.g., H/D exchange [10], analytical ultracentrifugation [30], titration calorimetry [62], and composition gradient static light scattering [3, 26]) enable characterization of the stoichiometry and strength of interaction but provide no or very limited structural information.

Small-angle scattering in solution (SAS) [15] provides an alternative experimental technique that we show here to be able to provide simultaneously both structural and association information for a complex. Although available for many years, SAS has recently gained popularity in low-resolution structural studies of protein monomers and tight complexes [7, 55, 57, 58, 64], as it is applicable to proteins of practically any size under physiological conditions, and data can now be collected rapidly at new higher-flux x-ray or neutron sources. However, its applicability to studies of complexes has been limited due to the need for a homogeneous and monodisperse sample, rendering it unsuitable for important weaker-binding complexes (e.g., associated with cellular signaling, which contain mixtures of the component monomers and intermediate and final complexes). We recently

described a method for the elucidation of homo-oligomeric complexes from solution scattering data [67], which was rapidly followed by reports of similar numerical approaches applied to experimental data [5], demonstrating the value of such methods. These methods, however, were only applied to homo-oligomers and were limited in their ability to handle systematic noise in the scattering data. Here we extend our earlier method so as to characterize hetero-oligomeric complexes, and develop a new analytical approach to handle contaminants in the mixtures, thereby yielding a method with potential applicability to an even broader range of biological systems and experimental conditions.

The method presented here determines the association model (the stoichiometry and affinity constants of all the association steps) from SAS data for a set of solutions containing the components of a hetero-oligomer in varying concentrations. (These solutions may also contain a contaminant, such as a non-participating monomer or homo-oligomeric aggregate.) In addition to the association model, our method accurately reconstructs the scattering curves of all the individual molecular species. These reconstructed curves can form the basis for low-resolution structural analysis of the intermediate and final complexes.

Scattering from an equilibrium mixture of initial components is a fractional mass-weighted linear combination of the "pure" scattering from all the molecular species in solution. We first employ low-rank approximation to remove some experimental noise from the observed mixture data. We then search over possible association models (which define a set of expected fractional masses for all the species), establishing a least-squares problem for each. Solution of the least-squares problem yields reconstructions of the "pure" scattering curves. We evaluate these hypothesized reconstructions for consistency with the data and with the postulated association model, and select the best model. If no model is of sufficient quality, we can expand the search to consider association models containing a contaminant. We have investigated the situation where the contaminant that is either a non-participating monomer or a homo-oligomeric aggregate of one of the initial components, since these represent the most important practical situations where the contaminants

are less likely to be removed by biochemical means during preparation of the initial components. In these cases, the least-squares approach is no longer applicable, so at the cost of computational time, we employ a convex quadratic program to compute scattering curves that are consistent with the data and satisfy additional constraints expected of physically realistic scattering curves.

We demonstrate the effectiveness of our method on simulations of four hetero-oligomeric complexes with different association pathways, association constants, molecular weights, and three-dimensional structures. Our simulation studies further demonstrate the robustness of our method to both random noise and systematic noise due to contaminants. In all cases, we are able to infer the correct association pathway and association constants that are very close to those used in simulation, as well as scattering curves that closely approximate those of the monomers and oligomers.

2.2 Methods

When several molecules are present in a solution, the observed scattering curve is the mass fraction-weighted linear combination of the scattering intensities for the individual components. Starting with scattering intensities collected from the equilibrium mixtures of a series of different concentrations of the initial components, our goal is to infer the association model along with the underlying scattering curves of the involved molecular species, including the initial components and intermediate and final complexes. Fig. 2.1 provides an overview of our approach for an example in which initial components A and B form an AB complex, with the association constant K_{AB} establishing the fractional amount of each of these forms at equilibrium. Each molecular species has an underlying scattering curve, but the association model and underlying scattering curves are unknown (gray shaded box). At given initial concentrations of A and B, the scattering curve for the equilibrium mixture is a weighted sum of that for the pure A, pure B, and pure AB, weighted by the equilibrium

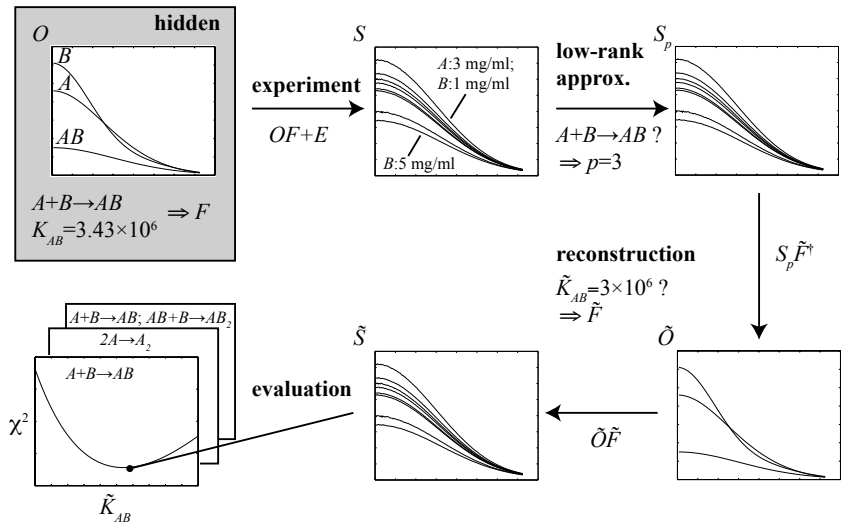


Fig. 2.1: An overview of our method, for an example one-stage system. The association model and scattering curves of the various molecular species are unknown. Scattering curves are collected over a series of different initial concentrations of the components. Each observed scattering curve is a linear combination of the unknown curves of the different species, according to the association model and initial concentrations of the components, plus noise. We systematically search over possible association models; for each, we use a corresponding low-rank approximation to de-noise the data, and we employ a least-squares formulation to reconstruct scattering curves of the different species. We evaluate agreement of each model's reconstructed curves with the experimental data, and select the best model. We extend this ideal framework to account for the most problematic possible contaminants (we have tested non-participating monomers and homo-oligomeric aggregates) by including an additional unknown scattering curve and fractional mass, and solving a quadratic optimization problem for reconstruction.

fractional masses. The experimentally-measured curve (normalized by total mass concentration of the mixture) is then composed from this weighted sum, plus experimental error. We collect a series of such curves, over a range of initial concentrations of A and B. We then search over possible association models, considering alternative pathways and values for the corresponding association constants (here only K_{AB}). When considering a possible pathway, we hypothesize an associated p, the number of molecular species that should be present for that model, and can thus extract a corresponding reduced set of scattering curves with random experimental noise partly removed, suitable for our analysis. When considering a set of association constants under this pathway, a set of fractional masses is hypothesized. Using them, we can compute a reconstruction of the underlying curves and

a corresponding reconstruction of the observed mixture curves. To determine the best association model and reconstructed curves, we perform a broad coarse-grid search followed by a narrow fine-grid search over possible association constants, scoring each for quality of fit to the observed data and agreement between the scattering curves and proposed stoichiometries of the complexes. We finally return the best pathway and constant, along with the corresponding reconstructions.

More formally, we represent our input scattering data as an $m \times n$ matrix S, with n columns for the samples at different starting concentrations of the initial A and B components, each with m rows for the scattering intensities at a fixed set of m scattering angles. Each mass concentration normalized scattering curve column in S represents a linear combination of p curves (initial components along with intermediate and product oligomers each at the standard mass concentration), weighted according to their equilibrium fractional masses. Collecting the curves into an $m \times p$ matrix O (one column per molecular species) and the fractional masses into a $p \times n$ matrix F (one row per set of initial monomer concentrations), and adding experimental noise E (one value per data point), we obtain

$$S = OF + E \; . \tag{2.1}$$

While S is the observed data, the values in the other matrices are unknown, and our goal is to infer the association model which detetermines F and set of curves O which together produce the observed S.

We now detail each of the steps in the following subsections. The presentation is generalized from that of our homo-oligomeric study [67], and refocused directly on solving the underlying least-squares problem. We initially assume that only the species in the modeled association are present in the various mixtures. We subsequently show how to modify the methods to handle potential situations where the presence of a contaminant that is a non-participating monomer or homo-oligomeric aggregate alters the ideal situation.

2.2.1 Low-rank approximation

When considering an association pathway (recall that we will search over the possibilities), we know the number p of molecular species that are present at equilibrium. Since the relationship between their mass fractions (and hence between rows of F) is non-linear, and since the number of concentrations is greater than the number of molecular species, we can extract a p-rank approximation S_p. This low-rank approximation S_p is a "de-noised" version of S (i.e., with E partially removed), containing the appropriate number p of curves with which to reconstruct the scattering curves according to the association model.

Singular value decomposition (SVD) is a popular technique for low-rank approximation, and has been employed by us [67] and others [9, 52, 53] in analysis of scattering data. SVD computes the low-rank approximation with the smallest distance to the input matrix, as measured by the Frobenius norm of the matrix difference, $\|S - S_p\|_F = \sqrt{\sum_{i,j}(S(i,j) - S_p(i,j))^2}$. The SVD of our $m \times n$ matrix S is given by $S = U\Sigma V^T$, where $m \times m$ matrix U and $n \times n$ matrix V are orthogonal matrices whose column vectors are the left and the right singular vectors, and $m \times n$ matrix Σ is a diagonal matrix whose elements are the singular values associated with the corresponding left/right singular vectors. The singular values are in order along the diagonal from largest to smallest, weighting the contributions from the most to least important singular vectors. To compute the pth low-rank approximation, we replace with zero the smallest $m - p$ singular values on the diagonal of Σ to give Σ_p, and then compute $S_p = U\Sigma_p V^T$.

2.2.2 Reconstruction

When considering a set of association constants for a pathway (recall that we will conduct a grid search over possible values for the association constants), we can apply standard association equilibria to compute the resulting equilibrium fractional mass of each of the p molecular species. We collect these fractional masses into a matrix $\tilde{F}$ (using the tilde to

indicate that it is our reconstruction of the "true", unobserved F). Combining this with the low-rank approximation S_p in a de-noised version of Eq. 2.1, we compute the least-squares solution in order to reconstruct scattering curves of the various species:

$$\tilde{O} = S_p \tilde{F}^{\dagger} \tag{2.2}$$

where $\tilde{F}^{\dagger}$ denotes the Moore-Penrose pseudoinverse. This formalization in terms of a p rank approximation is a generalization of the approach in [67], where using basis vectors from Singular Value Decomposition was an explicit part of the equations. It clarifies the role of the decomposition and allows the use of alternative approximation approaches. It is also different from [5] where PCA is used only to find the number of components in the solution.

If the least-squares solution $\tilde{O}$ has more than 10 percent negative intensity values or contains negative values in the small scattering angle range considered for Guinier analysis [12], we consider it to be non-physical, and reject the reconstruction without further analysis.

We then use $\tilde{O}$ to compute $\tilde{S}$, an approximation of the observed scattering curves of the equilibrium mixtures, by linearly combining the curves of the involved species at the appropriate fractional masses:

$$\tilde{S} = \tilde{O}\tilde{F} = S_p \tilde{F}^{\dagger} \tilde{F} \tag{2.3}$$

We thus reconstruct the scattering data from the low-rank approximation, consistent with the hypothesized association model.

2.2.3 Evaluation

To assess an association model, we evaluate how well the reconstructed scattering curves $\tilde{S}$ match the experimental ones S. We employ the two scoring approaches of our homo-

oligomeric work [67], customized for hetero-oligomers.

First a χ^2 score quantifies the differences over the entire set of scattering curves, weighted by the estimated error $\sigma(i,j)$ for each experimental data point:

$$\chi^2 = \frac{1}{m(n-p)} \sum_{j=1}^{n} \sum_{i=1}^{m} \left(\frac{S(i,j) - \tilde{S}(i,j)}{\sigma(i,j)} \right)^2 . \tag{2.4}$$

The sum of squared differences between points on the reconstructed and original curves is normalized by $m(n-p)$ degrees of freedom to yield a χ^2 score. While there are mn data points, p of the n degrees of freedom are fixed by the low-rank approximation. We show that in practice this score approximately equals 1 for the best fit to data with Gaussian simulated noise.

Second, the Mean Squared Mass Ratio Difference (MSMRD) score calculates whether the zero-angle intensities match the stoichiometry of the hetero-oligomeric forms. The scattering intensity at zero angle, estimated by Guinier analysis [12], is proportional to the molecular weight. Thus for example, we would expect $I(0)$ for species AB, $I_{AB}(0)$, to equal $I_A(0) + I_B(0)$, and thus $\frac{I_{AB}(0)}{I_A(0)+I_B(0)}$ to be 1. Thus the MSMRD score computes the average, over the various hetero-oligomeric forms, of the deviations of such ratios from the ideal value of one. Its expected value is thus zero. For a hetero-oligomer formed from A and B monomers, we compute the MSMRD as

$$\text{MSMRD} = \frac{1}{p-2} \sum_{(a,b)\in C} \left(1 - \frac{I_{A_a B_b}(0)}{a\, I_A(0) + b\, I_B(0)} \right)^2 \tag{2.5}$$

where C is a set of (a,b) pairs indicating the various $A_a B_b$ hetero-oligomeric forms, and $I_{A_a B_b}(0)$ represents their zero-angle intensity. For example, if the association model is $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$, the MSMRD score is given by:

$$MSMRD = \frac{1}{2} \left(\left(1 - \frac{I_{AB}(0)}{I_A(0) + I_B(0)} \right)^2 + \left(1 - \frac{I_{AB_2}(0)}{I_A(0) + 2I_B(0)} \right)^2 \right) \tag{2.6}$$

These two scores are complementary. The χ^2 is global, assessing the overall agreement

between the reconstruction and the data. However, two related association pathways (with an appropriate choice of association constants) can generate similar solutions and similar χ^2 values. For example, this can happen with a one-stage association pathway $A + B \rightarrow AB$ and the extended two-stage association pathway $A + B \rightarrow AB$, $A + B \rightarrow AB_2$, with a similar K_{AB} for both cases and a very weak K_{AB_2} for the second (see Results). This is because Eq. 2.3 can give similar solutions for two different matrices F, as long as the column space spanned by the fractional matrix is the same. On the other hand, the MSMRD is very local, ignoring the agreement over most of the curve and focusing on the zero angle intensity in order to assess the agreement between the independent (and not directly optimized) expected molecular weights and the stoichiometry. We have found that considering both χ^2 and MSMRD improves the determination of the correct association model (see Results).

2.2.4 Association model search

We have discussed how to reconstruct and evaluate scattering curves for a given association model defined by a pathway and corresponding set of association constants. To determine the best model, we separately reconstruct and evaluate models for a set of plausible pathways, over a grid of possible association constants.

The pathways to be considered are chosen based on the set of oligomers that could possibly be present in the equilibrium mixture. Although potentially infinite, a most likely set of oligomers can be selected, for example, from an analysis of the zero-angle scattering or by the radii of gyration of the experimental scattering curves. Then we consider all pathways that could form complexes with the allowed sets of subunits. For example if we knew that there were two monomers, A and B, and determined that the final oligomer had at most three subunits, then we would evaluate the one-stage associations $2A \rightarrow A_2$, $2B \rightarrow B_2$, and $A + B \rightarrow AB$, along with the two-stage associations that extend these

to yield A_2B and AB_2. Like other approaches, e.g., analytical ultracentriguation, where postulated association models are fit to the data, assumptions have to be made for the most likely models to be assessed.

We perform coarse and fine grid searches over possible values for the association constants. Each association constant is an independent dimension in the grid. Results presented below use grids covering the range of plausible constants: 10^{-6} to 10^{25} for a one-stage association and 10^{1} to 10^{15} for a two-stage association. An initial coarse grid is searched at integer multiples of the powers of 10 (e.g., $1 \times 10^3, 2 \times 10^3, 3 \times 10^3, \ldots, 9 \times 10^3, 1 \times 10^4, 2 \times 10^4, \ldots$). For each point (representing one or a pair of association constants), the curves are reconstructed and evaluated by χ^2 and MSMRD, as described above. The constants with the best scores establish a region for a fine grid search, plus or minus one unit in each dimension, with a spacing of 1% of that of the coarse grid. We only perform fine grid searches for the models with the best χ^2 and MSMRD values from the coarse grid search and for which the best coarse grid association constants from the χ^2 and MSMRD scores are in sufficient agreement. We finally select the model with the best fine-grid χ^2 and MSMRD scores, determining the corresponding pathway, association constants, and reconstructed curves. In cases where the fine grid search fails to yield an acceptable model, due to either a high χ^2 for the best fine grid point, or large disagreement between the best χ^2 and MSMRD fine grid points, the methods in the next section can be employed to account for contaminants.

2.2.5 Accounting for contaminants

When the scattering data contain a substantial contaminant, we have developed an extension to our methodology. Since contaminants that are unrelated to the initial components are generally readily purified by current protein separation methods, we seek to solve biochemical situations that arise most frequently. We focus on cases in which the contaminant

is either a non-participating monomer or a large homo-oligomeric aggregate of one of the components.

Let us assume that the contaminant is a non-participating monomer or homo-oligomeric aggregate of A (the methodology works the same for any component and could be generalized to multiple such contaminants). Note that in our approach, the contributions from all species in a polydisperse homo-oligomeric aggregate can be accounted for by one combined scattering curve and one total contaminant fraction. Let c be the unknown mass fraction of A that forms the contaminant. As part of our grid search, we will consider possible values for c along with those for the association constant(s). Given a hypothesized value for c and the association constant(s), we must build a fractional mass matrix $\tilde{F}$ for each, now containing $p + 1$ rows, with the extra row for the contaminant. In constructing this matrix, let a_i be the initial amount of A in sample i. Then the amount of a_i still participating in the hypothesized association (rather than in the contaminant) is $a_i(1 - c)$. We determine the equilibrium concentrations and thereby masses of the other forms from the reduced A concentration and the initial concentrations of the other initial component(s).

Unfortunately, the extended $\tilde{F}$ is no longer of full rank in the presence of contaminant, as the fractional mass vector for the contaminant is linearly dependent on A. This in turn implies that there is an infinite set of widely varying least-squares solutions $\tilde{O}$ satisfying $\tilde{O}\tilde{F} = S_p$. One of these, which we call $\tilde{O}_0$, is the solution from our earlier formula (Eq. 2.2), $\tilde{O}_0 = S_p\tilde{F}^\dagger$. If we use this $\tilde{O}_0$ to reconstruct $\tilde{S}$, as in Eq. 2.3, we obtain $S_p\tilde{F}^\dagger\tilde{F}$, which we call $S_{p,\tilde{F}}$. Each least-squares solution $\tilde{O}$ produces this same $S_{p,\tilde{F}}$ and thus cannot be distinguished by comparison to the data S or the denoised data S_p. This equivalence of solutions $\tilde{O}$ is due to the fact that the set of least-squares solutions is composed of the sum of $\tilde{O}_0$ with an infinite set of matrices of row vectors (that is, adjustments to the scattering curves) from the null space of $\tilde{F}^T$. Post-multiplication by $\tilde{F}$ then reduces the second matrix in this sum to zero, resulting in no change to $S_{p,\tilde{F}}$.

In summary, there is an infinite number of reconstructions of the pure curves $\tilde{O}$, but each

produces the same reconstructed data $S_{p,\tilde{F}}$. Since we use the reconstructed data to compute χ^2 (Eq. 2.4), we can find the best association model (best $\tilde{F}$) via coarse and fine grid searches as before, with an additional dimension of the contaminant fraction in addition to association constant(s). This approach does not, however, produce correct reconstructed pure scattering curves and thus also does not give MSMRD values. Therefore, after identifying the best χ^2 point (or a set of feasible points for consideration), we must search over the space of satisfying $\tilde{O}$ to reconstruct and evaluate pure scattering curves and identify a best one.

We have developed a quadratic optimization framework that seeks an $\tilde{O}$ that not only explains the data (which all $\tilde{O}$ do equally) but also has properties desirable of physically realistic scattering curves. In particular, we establish smoothness as our objective function, and incorporate constraints limiting the sub-optimality of χ^2, while also enforcing the expected decaying exponential trend in the Guinier region of the scattering curves as well as the expected ratios of $I(0)$ values (as also employed in our MSMRD score). We note that if the contaminant only involves form A, for example, then the row for B in the fractional mass matrix is linearly independent from the contaminant and yields a unique least-squares solution (the same in $\tilde{O}$ for any $\tilde{O}_0$). Thus after computing $\tilde{O}_0$, we remove the row for initial component B in $\tilde{F}$ and from $S_{p,\tilde{F}}$ (via its row in $\tilde{F}$ and column in $\tilde{O}_0$). For simplicity, we continue to refer to $\tilde{O}$ and $\tilde{F}$ without distinguishing the reduced-parameter versions.

Objective: smoothness. With the available freedom in $\tilde{O}$, there are curves that use wildly fluctuating values to obtain good χ^2 scores upon post-multiplication by $\tilde{F}$. Since we expect physical curves to be relatively smooth, we establish as our objective function a discrete evaluation of smoothness. We construct a finite difference matrix D that, when multiplied with $\tilde{O}$, approximates the second order derivative at each point on the curve. We then seek to minimize the total of the squared differences, i.e., the square of the Frobenius norm of

$D\tilde{O}$.

$$\min_{\tilde{O}} ||D\tilde{O}||_F^2 \tag{2.7}$$

Constraint: χ^2 deviation. We seek a reconstruction with the optimal χ^2 (as with all the $\tilde{O}$, satisfying $\tilde{O}\tilde{F} = S_{p,\tilde{F}}$), but since the data are noisy, we may sacrifice a little in χ^2 score in order to ensure a feasible optimization problem and do better in terms of smoothness and other characteristics. We thus impose a constraint that the reconstructed curves are no more than ϵ_{fit} away from the one that gives the lowest χ^2. This tolerance should be set fairly low to keep the identified curves near the optimal one; for our results, we use 10^{-3}.

$$(1 - \epsilon_{\text{fit}})S_{p,\tilde{F}} \leq \tilde{O}\tilde{F} \leq (1 + \epsilon_{\text{fit}})S_{p,\tilde{F}} \tag{2.8}$$

Constraint: non-negativity. Scattering curves are non-negative.

$$\tilde{O} \geq 0 \tag{2.9}$$

Constraint: Guinier. Scattering curves exhibit decaying exponential intensity in the Guinier region [12]. Therefore, we impose a constraint so that curves are non-increasing (within a tolerance) in the initial Guinier region. To approximate the Guinier region in the scattering curves in $\tilde{O}$ without iterating on R_g values we use $q_{\max} = 1.33/R_g$ [19] and a fixed $R_g = 40$. To allow for noise, we enforce this property only to within a tolerance $\epsilon_{\text{Guinier}}$: within the Guinier region, one intensity is no more than $(1 + \epsilon_{\text{Guinier}})$ times the intensity at the next lower scattering angle. A reasonable value for $\epsilon_{\text{Guinier}}$ can be estimated by examining some pure intensity curves that reconstructed from uncontaminated simulations with standard noise; we use $2 \cdot 10^{-2}$. Note that this value is dependent on the extent of noise and the spacing of scattering angles. We formulate this constraint with a matrix G which, when multiplied by $\tilde{O}$, gives the differences between (1 + $\epsilon_{\text{Guinier}}$) times a point and

the next point, for points in the scattering curves in $\tilde{O}$ at $q < q_{\max}$.

$$G\tilde{O} \geq 0 \tag{2.10}$$

Constraint: molecular weights. When we are considering a contaminant that is a non-participating form of A (either monomer or aggregate), we know that its native mass must be at least that of A, that is $M_X > M_A$. Thus the zero angle intensity of its scattering curve should be at least equal to that of $I_A(0)$. Since the extrapolation to obtain $I(0)$ requires an exponential fit (which would render our system non-linear), we instead use $I(q_{\min})$, the intensity at the smallest angle measured.

$$I_X(q_{\min}) - I_A(q_{\min}) \geq 0 \tag{2.11}$$

where the scattering curves I_A and I_X (for A and the contaminant X) are particular vectors of $\tilde{O}$.

Imposing this constraint on $I(q_{\min})$ instead of $I(0)$ results in negligible error, since, from the Guinier relationship we have:

$$I_X(q_{\min})/I_A(q_{\min}) = M_X/M_A \exp(-1/3\ q_{\min}^2(R_g(X)^2 - R_g(A)^2)) \tag{2.12}$$

where R_g is the radius of gyration. Given that $q_{\min}^2$ is generally quite small (on the order of 10^{-6} in experimental data), the difference in radii of gyration is not large enough to substantially impact the results.

Furthermore, since we have found a unique scattering curve for B, we can use its intensity at $q_{\min}$ to constrain the intensity at $q_{\min}$ of the scattering curve for A and other forms (excluding the contaminant). We are essentially encoding MSMRD (relative to the independent form B) as a constraint but for intensities at $q_{\min}$, instead at zero angle. As with most other constraints, we use a tolerance to allow for some noise. We have found

$\epsilon_{\text{msmrd}} = 0.1$ to work well for our tests, but for other data, this tolerance could potentially be further tightened as long as feasible solutions still result. For the scattering from A (I_A) and every other molecular species $A_k B_l$ ($I_{A_k B_l}$), we add constraints of the form:

$$\begin{aligned}(1 - \epsilon_{\text{msmrd}}) I_B(q_{\min}) M_A / M_B &\leq I_A(q_{\min}) \\ &\leq (1 + \epsilon_{\text{msmrd}}) I_B(q_{\min}) M_A / M_B\end{aligned} \tag{2.13}$$

$$\begin{aligned}(1 - \epsilon_{\text{msmrd}}) I_B(q_{\min})(k\, M_A + l\, M_B)/M_B &\leq I_{A_k B_l}(q_{\min}) \\ &\leq (1 + \epsilon_{\text{msmrd}}) I_B(q_{\min})(k\, M_A + l\, M_B)/M_B\end{aligned} \tag{2.14}$$

where again the scattering curves I are particular vectors in $\tilde{O}$.

Solving the system. While we have written the objective and constraints in terms of $\tilde{O}$ and other matrices, we can re-shape these matrices into long vectors (i.e., by stacking columns). The combination of the objective function and constraints yields a convex quadratic optimization problem that can be solved by numerous solvers. If the quadratic optimization program is infeasible for a hypothesized association model, we discard that model. If more than one feasible model were to remain, we could compute MSMRD values and select the best, but that did not happen in our simulation studies presented below.

2.2.6 Implementation

The methods have been implemented in a platform-independent Python package that is available upon request. The package calls the IBM ILOG CPLEX optimizer to solve the system of equations. Our program lets a user search over possible association models based on specifications provided via the command line or in an input file. The package contains implementations for both the contaminant-free search and the extension to handle non-participating monomers and homo-oligomeric contaminants. In addition to the methods in this paper, it also contains an implementation for homo-oligomeric association models from our previous work [67].

To obtain the results presented below, coarse and fine grid searches for a one-stage model took less than a minute, while searches for a two-stage model took a few minutes on a single core Intel Xeon 2.50 GHz processor. The three stage searches took a few hours. The time for contaminant searches was similar to the contaminant-free search being increased by another dimension.The quadratic program solver usually took less than a minute.

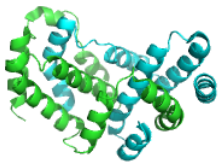
2.3 Results

In order to evaluate the effectiveness of our method in a range of scenarios, we performed an extensive set of simulation studies, with different association pathways and association constants, and varying levels of random noise, data resolution, and monomer size. Fig. 2.2 summarizes the complexes used in these studies, and illustrates their crystal structures and the simulated scattering curves of the monomers and intermediate and final oligomers at a constant mass concentration. The complex structures were taken from the PDB [4] (pdb ids indicated), and monomer and intermediate complex structures extracted. The association models for simulation were not taken from experimental data; instead, we chose them to challenge the ability of our method to determine the correct model even in the presence of alternatives that have intermediate and final complexes of similar mass (note the similarity of initial component masses in the Bovine IFN-gamma and the human growth hormone-receptor cases). We chose association constants in the middle of a feasible range; however, we explicitly assessed the impact of the constants in one set of simulations.

We have found that as few as eight different initial concentrations provides a sufficient set of different scattering curves for subsequent reconstruction, and the results shown are based on eight for all test cases. The initial concentrations used (Supplementary Tab. 2.5 and Tab. 2.6) are all in the 0.5-5.0 mg/ml range where SAS data is easily collected. They were chosen so as to yield a diverse set of row vectors (fractional masses) in the fractional mass matrix F, adequately sampling the space and ensuring that important vectors (scat-

Bovine IFN-gamma
$A + B \rightarrow AB$
$K_{AB} = 3.43 \times 10^6$
A: 14.2 kDa; B: 13.3 kDa

1D9G

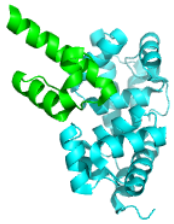


Human calcineurin
$A + B \rightarrow AB$
$K_{AB} = 4.24 \times 10^4$
A: 43.6 kDa; B: 18.8 kDa

1AUI

BAF-emerin complex
$A + B \rightarrow AB$, $AB + B \rightarrow AB_2$
$K_{AB} = 3.21 \times 10^5$, $K_{AB_2} = 4.23 \times 10^5$
A: 5.7 kDa; B: 10.1 kDa

2ODG



Human growth hormone-receptor complex
$A + B \rightarrow AB$, $AB + B \rightarrow AB_2$
$K_{AB} = 8.43 \times 10^5$, $K_{AB_2} = 6.26 \times 10^4$
A: 21.0 kDa; B: 22.5 kDa

3HHR

Fig. 2.2: Case studies.

tering from intermediate and final complexes) are included in the low-rank approximation. Even so, the equilibrium mixtures are rarely more than 70% of one form. In practice, of course, F cannot be assessed initially, but we still recommend ensuring that there is a diverse set of initial concentrations, with different combinations of low and high monomer concentrations. In the absence of approximate knowledge of the association constants that determine F, a first-round analysis can be used to identify a definitive set of initial concentrations for which to collect data. We do include pure monomer solutions (only A, only B) as initial components so as to better characterize them and account for their contributions to the mixtures. Of course, pure monomers may not be biochemically available, but the method is not dependent on this and any available components could be used.

The program CRYSOL [56] was used at the default settings to simulate noiseless scattering intensities O from the 3D structures of each initial component and complex. The noiseless equilibrium mixture intensities were then simply calculated as OF. Noise E was then added, following the method employed by Williamson *et al.* [67] to simulate realistic angle-dependent Gaussian noise based on noise levels observed in experimental samples. Ten datasets were generated for each example, with different random noise added for each dataset.

While we studied two one-stage associations and two two-stage associations, we present detailed results for only one of each and summarize the second, since results were similar in each category. We first show that our method yields the correct association model on our initial simulated data, for both one-stage and two-stage examples. We then demonstrate the robustness of our method to noise, and investigate the range of association constants for which the method is applicable. Finally, we consider test cases with simulated contamination and present results from our expanded method that accounts for the contaminant.

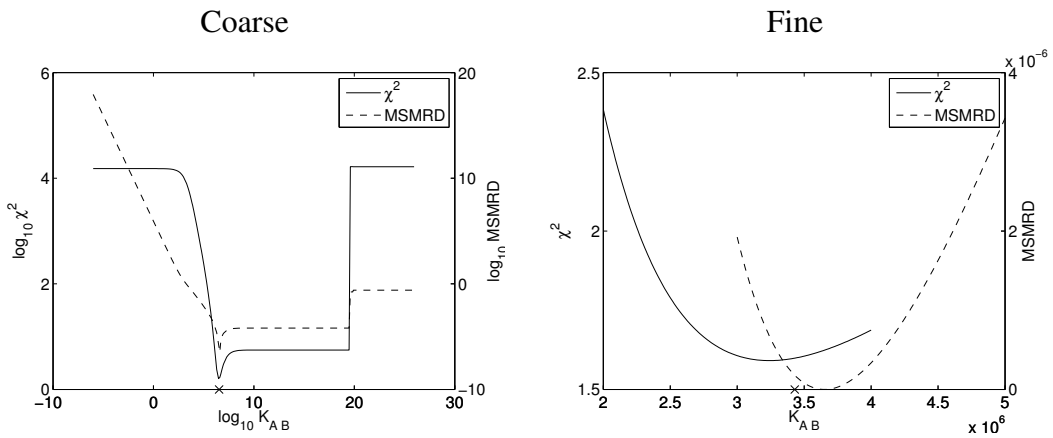


Fig. 2.3: Association constant searches for one Bovine IFN-gamma dataset, for the correct $A + B \rightarrow AB$ pathway. The '$\times$' mark on the x-axis indicates the simulated association constant (3.43×10^6).

2.3.1 Baseline simulations

Bovine IFN-gamma (one stage). We first examine the results for one of the 10 simulated datasets (i.e., one Gaussian noise matrix E), with the correct pathway $A + B \rightarrow AB$ and varying the association constants on a coarse grid (Fig. 2.3, left) and fine grid (Fig. 2.3, right). Both plots show a steep decline in χ^2 and MSMRD scores around the simulated association constant value (3.43×10^6), with a minimal χ^2 of 1.59 at 3.34×10^6, and minimal MSMRD of 1.67×10^{-11} at 3.65×10^6. The close agreement of these association constants and high quality of the scores under these complementary metrics gives us confidence in this solution.

While in an experimental setting we would not have access to the "true" scattering curves of the various molecular species (O), here we do (from the CRYSOL calculation on the model components and complexes), and can evaluate how well the reconstructed curves agree with them ($\tilde{O}$, computed by Eq. 2.2). Fig. 2.4 shows the approximately random residuals between the reconstructed and simulated curves, at the association constant $K_{AB} = 3.34 \times 10^6$ which yields the best χ^2 score. (The apparent deviation from random residuals seen at higher resolution for component B (Fig. 2.4, middle) was not explained by deviation between simulated and best χ^2 association constant.) To quantify the extent of

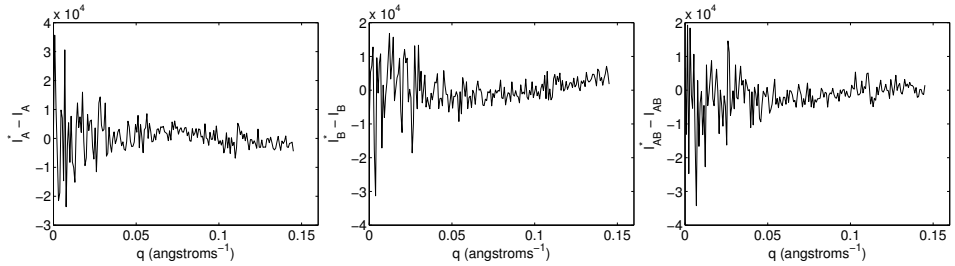


Fig. 2.4: Residuals between pure simulated scattering intensities and reconstructed ones for Bovine IFN-gamma χ^2-optimal association models.

agreement, we compute the *median of the absolute relative deviation (MARD)*, as a percentage deviation of the reconstructed curve from the simulated one; a MARD value close to zero indicates that the reconstructed curve is very close to the original noiseless CRYSOL curve. MARD scores confirm the agreement illustrated in the figure: A has a MARD of 0.24%, B has 0.16%, and AB has 0.22%, averaged across the ten datasets with different simulated noise. Tab. 2.1 summarizes results over all 10 simulated noisy datasets, comparing the correct pathway with alternatives. The $A + B \rightarrow AB$ pathway was always chosen and the average association constant was close to the simulated one with only small variation between data sets. Only the related two-stage pathways $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ and $A + B \rightarrow AB$, $AB + A \rightarrow A_2B$ obtained coarse-grid χ^2 scores (averaging 1.62 and 1.55, resp.) competitive with that of the correct model (1.53); the rest were much worse. Both alternative models extend the correct model with an additional association of weak affinity, keeping the $A + B \rightarrow AB$ association as the primary one. Any additional association hurts the MSMRD scores (1.19×10^{-3} and 8.21×10^{-4}, vs. 3.74×10^{-7} for the correct model), as the low angle data do not support an oligomer with molecular weight corresponding to AB_2 or A_2B. In addition, while the optimal association constants for χ^2 and MSMRD are very similar for the correct model, the best association constants by these two metrics are quite different for the alternative ones. Furthermore, there is not a choice of constants that scores moderately well under both metrics, and the association constants giving the best χ^2 score yield a poor MSMRD score and *vice versa*. For pathway $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$, the MSMRD for the association constant

with the best χ^2 score averages 9.53×10^{-2} across the ten datasets, versus an average best MSMRD of 1.19×10^{-3}. On the other hand, the χ^2 score for the association constants with the best MSMRD score is 33.45 on average. These values are more than an order of magnitude worse than the best χ^2 and MSMRD scores for the correct pathway. We find similar results for the second alternative pathway. Even though the χ^2 scores are not good discriminators, the substantial deterioration in the MSMRD and the disagreement between MSMRD and χ^2 metrics for the alternative models point to the correct $A + B \rightarrow AB$ pathway.

BAF-Emerin complex (two stage). Fig. 2.5 shows both χ^2 and MSMRD scores on the coarse and fine grids for the correct $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ pathway, for one example noisy dataset. As in the one-stage case, there are well-defined minima, with the best association constants yielding much better χ^2 and MSMRD scores than nearby alternatives, at both coarse and fine resolutions. We again see good agreement as to the best association constants under the two scores: χ^2 gives $K_{AB} = 3.16 \times 10^5$, $K_{AB_2} = 4.16 \times 10^5$, and MSMRD gives $K_{AB} = 3.27 \times 10^5$, $K_{AB_2} = 4.35 \times 10^5$, with the simulated constants being $K_{AB} = 3.21 \times 10^5$, $K_{AB_2} = 4.23 \times 10^5$. Interestingly, under both metrics, the best association constants lie on a diagonal line in which K_{AB} and K_{AB_2} are increasing at a similar rate, ensuring that if more AB is produced than the data dictate it is also converted to AB_2. While this keeps the fraction of AB relatively constant, the resulting excessive depletion of A and excessive formation of AB_2 yield worse scores at points along the diagonal line other than the minimum. The reconstructed intensities at the best association constants are quite similar to the original simulated noiseless ones as illustrated in the residuals (not shown) and quantified by average MARD values for A: 0.08%, B: 0.08%, AB: 0.15%, and AB_2: 0.06%.

Tab. 2.2 summarizes the results across ten noisy datasets for a set of possible association pathways. The best χ^2 score, averaging 1.17, is obtained by the correct pathway ($A + B \rightarrow$

Tab. 2.1: Bovine IFN-gamma association model searches, over 10 sets of simulated noise. Pathways are abbreviated by the reaction products; e.g., (AB) means $A + B \rightarrow AB$, while (AB, AB_2) means $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$. The bolded entries obtain the best scores, and are for the correct pathway. The simulated association constant was $K_1 = 3.43e6$.

	χ^2			MSMRD		
Pathway	$K_1 \pm$SD	$K_2 \pm$SD	Score$\pm$SD	$K_1 \pm$SD	$K_2 \pm$SD	Score$\pm$SD
(AB)	**3.30e6$\pm$4.8e5**	-	**1.53$\pm$0.11**	**3.70e6$\pm$4.8e5**	-	**3.74e-7$\pm$3.0e-7**
(AB) **Fine**	**3.40e6$\pm$7.1e4**	-	**1.49$\pm$0.12**	**3.64e6$\pm$5.2e5**	-	**1.82e-11$\pm$1.8e-11**
(AB_2)	8.00e7$\pm$0.0e0	-	881$\pm$2.1	8.00e7$\pm$0.0e0	-	1.97e-5$\pm$1.0e-5
(A_2B)	8.00e22$\pm$1.8e7	-	1820$\pm$3.2	3.00e5$\pm$0.0e0	-	2.69e-3$\pm$3.6e-4
(A_2)	1.00e19$\pm$0.0e0	-	7320$\pm$6.4	1.00e4$\pm$0.0e0	-	1.07e-4$\pm$1.3e-5
(B_2)	1.00e-6$\pm$2.2e-22	-	2860$\pm$2.9	3.00e14$\pm$0.0e0	-	1.36e-2$\pm$2.4e-4
(AB, AB_2)	1.38e11$\pm$1.7e11	3.30e8$\pm$4.0e8	1.62$\pm$0.12	1.07e8$\pm$3.1e8	9.01e4$\pm$2.5e5	1.19e-3$\pm$1.2e-3
(B_2, AB_2)	5.60e6$\pm$5.2e5	8.30e14$\pm$1.2e14	105$\pm$10	5.00e4$\pm$0.0e0	2.00e10$\pm$0.0e0	4.41e-4$\pm$4.1e-5
(AB, A_2B)	7.50e6$\pm$4.8e6	7.29e3$\pm$8.8e3	1.55$\pm$0.20	1.20e9$\pm$1.3e9	1.40e1$\pm$5.2e0	8.21e-4$\pm$1.8e-3
(A_2, A_2B)	9.00e6$\pm$0.0e0	7.00e14$\pm$0.0e0	394$\pm$1.6	1.20e9$\pm$1.3e9	1.40e1$\pm$5.2e0	8.21e-4$\pm$0.0e0

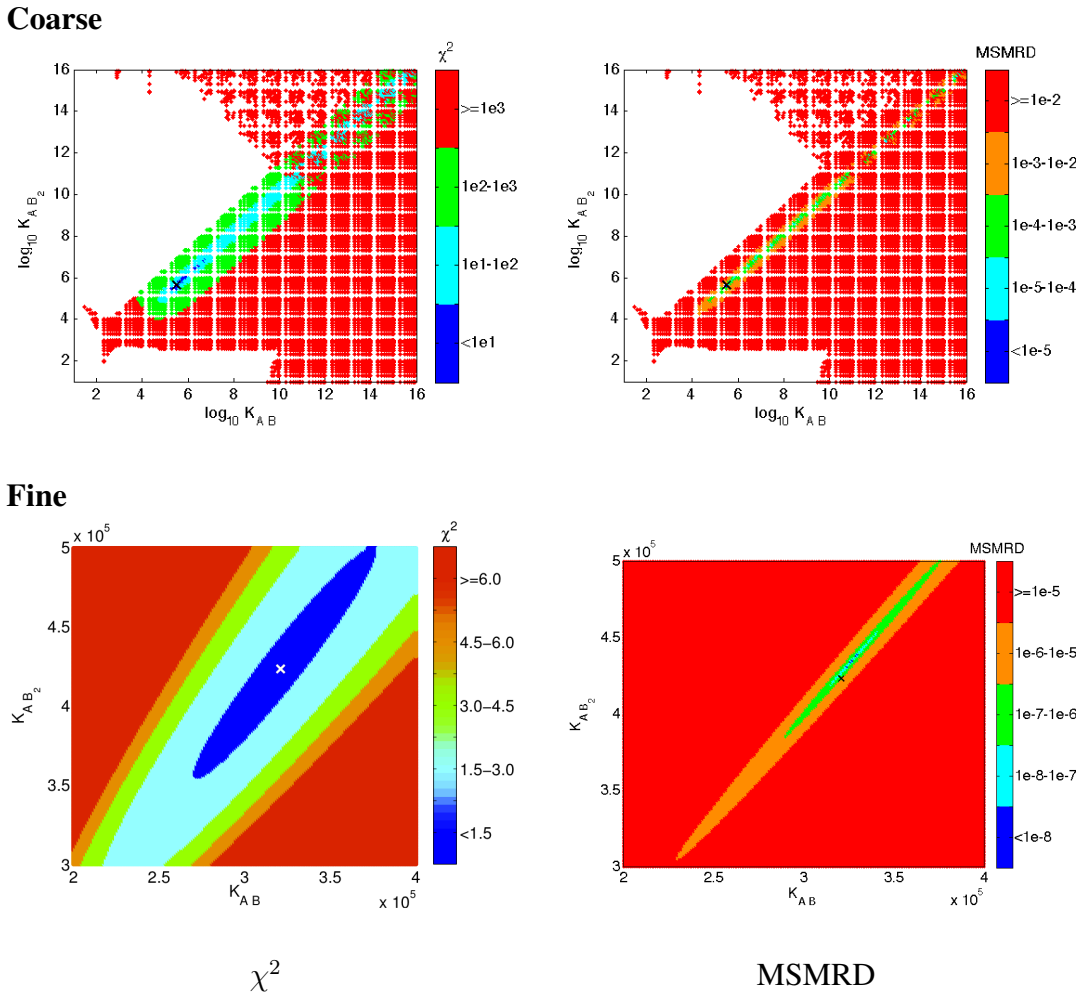


Fig. 2.5: Association constant searches for one BAF-Emerin complex dataset, for the correct $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ pathway. The '$\times$' marks indicate the simulated association constant ($K_{AB} = 3.21 \times 10^5$, $K_{AB_2} = 4.23 \times 10^5$). The white regions in the coarse grid plots indicate the constants yielding nonphysical scattering curves (those with substantial negative intensities).

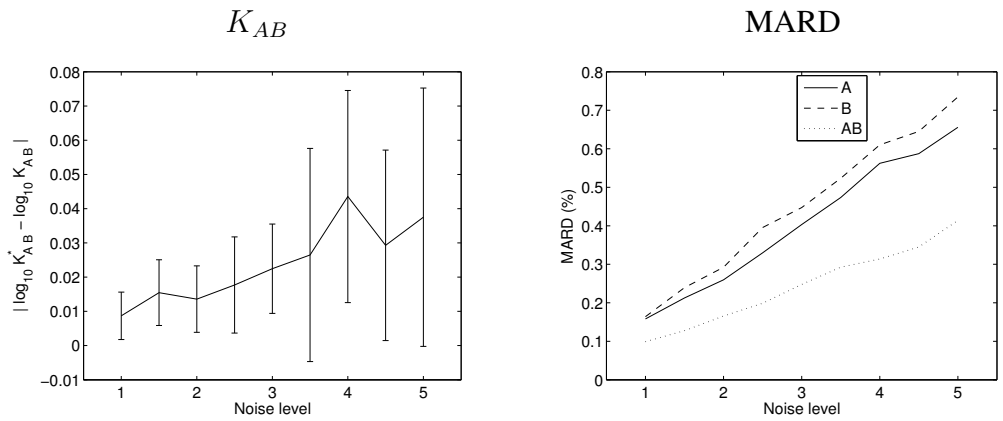


Fig. 2.6: Effect of noise level on error in association constant (left), assessed by absolute difference in $\log_{10} K_{AB}$; and on reconstructed scattering curves (right), assessed by MARD. Values are for the Bovine IFN-gamma best χ^2 fine grid point, averaged over ten datasets at each noise level. The association constant plot shows the means and standard deviations for the 10 datasets at each noise level; only means are shown in the MARD plot, for clarity.

AB, $AB + B \rightarrow AB_2$). The next best χ^2 scores, averaging 1.28 and 4.49, are obtained by alternative three-stage pathways that add weak association reactions $AB_2 + B \rightarrow AB_3$ or $AB_2 + A \rightarrow A_2B_2$ to the correct pathway. As before larger changes in the MSMRD scores are seen. The first alternative (adding AB_3) has an MSMRD score that is almost 10^4 times worse than the best MSMRD score. The second alternative (adding A_2B_2) has an MSMRD score that is more than 40-fold higher than the best MSMRD score (6.09×10^{-5} compared to 1.33×10^{-6} of the correct pathway). Furthermore, comparing the best χ^2 association constants against the best MSMRD constants in these alternative pathways reveals that they differ by approximately 10^2 in K_1, 10^3 in K_2, and 10^4 in K_3. Furthermore, as before, neither alternative pathway has a set of constants that score well under both metrics. Thus by using χ^2 and MSMRD scores together, we can determine the correct pathway.

2.3.2 Robustness to noise

Our simulated datasets include a realistic estimate to Gaussian noise found in experimental datasets at third generation synchrotron sources [67], but our simulation framework enables

Tab. 2.2: BAF-Emerin complex association model searches, over 10 sets of simulated noise. Pathways are abbreviated by the reaction products; e.g., (AB, AB_2) means $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$. The bolded entries obtain the best scores, and are for the correct pathway. The simulated association constants were $K_1 = 3.21e5$, $K_2 = 4.23e5$.

	χ^2				MSMRD			
Pathway	K_1±SD	K_2±SD	K_3±SD	Score±SD	K_1±SD	K_2±SD	K_3±SD	Score±SD
(AB)	1.48e10±3.0e10	-	-	39800±130	1.00e19±0.0e0	-	-	6.23e-4±1.9e-5
(AB_2)	3.00e18±0.0e0	-	-	8200±5.9	1.00e9±0.0e0	-	-	1.09e-5±3.1e-6
(A_2B)	5.00e18±0.0e0	-	-	55600±12	9.00e4±0.0e0	-	-	8.75e-5±1.1e-5
(A_2)	3.00e14±0.0e0	-	-	43700±6	1.00e2±0.0e0	-	-	1.20e-1±5.2e-4
(B_2)	4.00e3±0.0e0	-	-	5250±6.2	7.00e3±0.0e0	-	-	4.62e-6±2.0e-6
(AB, AB_2)	**3.00e5±0.0e0**	**4.00e5±0.0e0**	-	**1.17±0.24**	**3.00e5±0.0e0**	**4.00e5±0.0e0**	-	**1.33e-6±6.8e-7**
(AB, AB_2) **Fine**	**3.21e5±4.5e3**	**4.24e5±6.9e3**	-	**1.12±0.24**	**3.24e5±1.3e4**	**4.29e5±1.9e4**	-	**1.03e-9±6.6e-10**
(B_2, AB_2)	1.00e7±0.0e0	6.00e14±0.0e0	-	4880±7.4	9.00e3±0.0e0	4.00e10±0.0e0	-	4.53e-4±2.7e-5
(AB, A_2B)	6.00e15±0.0e0	3.00e13±0.0e0	-	1390±3	5.00e4±0.0e0	2.00e15±0.0e0	-	1.42e-2±2.0e-5
(A_2, A_2B)	4.00e7±0.0e0	9.00e15±0.0e0	-	22500±6	6.00e1±0.0e0	4.00e5±0.0e0	-	3.40e-3±2.4e-4
(A_2, B_2, A_2B_2)	9.00e1±0.0e0	9.00e10±0.0e0	2.23e8±6.2e8	2620±5	8.00e1±0.0e0	9.00e10±0.0e0	2.00e5±0.0e0	1.27e-1±3.7e-4
(AB, AB_2, A_2B_2)	2.00e5±0.0e0	3.00e5±0.0e0	1.00e1±0.0e0	4.49±0.39	8.00e2±0.0e0	8.00e2±0.0e0	5.00e1±0.0e0	2.50e-3±2.5e-6
(AB, A_2B, A_2B_2)	4.50e10±1.4e10	1.09e10±5.0e9	8.90e3±7.4e2	2180±38	6.20e10±1.8e10	7.50e8±1.4e8	7.90e6±9.9e5	4.85e-2±5.2e-5
(AB, AB_2, AB_3)	3.80e5±4.2e4	4.80e4±4.2e4	1.60e3±8.4e2	1.28±0.38	1.07e7±1.4e7	1.60e7±2.1e7	3.10e1±1.7e1	6.09e-5±2.7e-5

us to easily assess how robust our method is to much noisier data. As one example, we generated ten noisy datasets for the one-stage Bovine IFN-gamma with the resolution-dependent Gaussian noise scaled up by a factor of two. The correct $A + B \rightarrow AB$ pathway was still the clear winner in all the datasets. It achieved a very good fine-grid χ^2 score (an average of 1.23 across 10 datasets, compared to 1.00 with the standard noise) at a nearly-correct association constant (3.35×10^6, the same as with the standard noise, and near the simulated value of 3.43×10^6). It also achieved a good fine-grid MSMRD score (3.41×10^{-11}, compared to 1.46×10^{-14}), with a good association constant (3.86×10^6).

We then tested the performance of our method over a range of noise levels, increasing the Gaussian width up to 5-fold, generating 10 datasets for each noise level. We assessed the results in terms of identification of the association constant as well as reconstruction of the underlying scattering curves of the monomers and oligomers. For association constants, we assess the error with the absolute difference between the base 10 logs of the correct K^*_{AB} and the inferred K_{AB}, i.e., $|\log_{10} K^*_{AB} - \log_{10} K_{AB}|$. For scattering curves, our evaluation is the median absolute relative deviation (MARD) discussed above. Fig. 2.6 illustrates these error measures with respect to increasing noise (averaged over the ten datasets for each level). The figure shows that as the noise increases our best fine grid points and reconstructions gradually become further away from the correct ones. Even at 5x noise, the errors in asssociation constants remain acceptable, approaching 10% (averaged across ten datasets), while the MARD values remain under 1% (0.6% for A, 0.7% for B, and 0.3% for AB averaged across ten datasets). Thus we conclude that the method is indeed robust to such random noise. Robustness to some aspects of systematic noise (contamination with non-participating molecules) is discussed below.

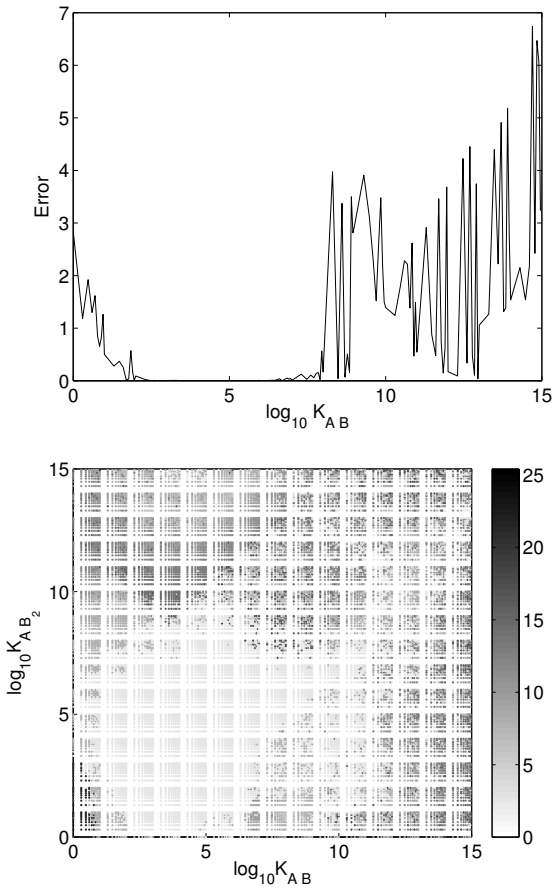


Fig. 2.7: Error in inferring simulated association constant for (left) one-stage Bovine IFN-gamma and (right) two-stage BAF-Emerin complex. The error for an association constant is the absolute log difference between simulated and inferred association constant; for the two-stage case, the overall error is the square root of the sum of the squared errors.

2.3.3 Robustness across ranges of association constants

The ability of our method to recover the contribution from a particular species depends on that species making a non-negligible contribution to the mixture scattering data. That in turn depends on the association constants. Our simulations used physiologically reasonable constants, selected to ensure non-negligible quantities of each molecular species at equilibrium. However, since there is a wide range of reasonable values for weak association, we conducted a set of one- and two-stage simulations with varying association constant pairs to assess the range of values suitable for our method. For each association constant of pair of association consetants, we compared the simulated value with the best χ^2 constants (results with MSMRD are similar and not shown). Absolute log differences were used to assess the differences between simulated and inferred values. For two association constants, we evaluated the Euclidean distance:

$$\sqrt{(\log_{10} K^*_{AB} - \log_{10} K_{AB})^2 + (\log_{10} K^*_{AB_2} - \log_{10} K_{AB_2})^2}$$

Fig. 2.7 shows the error over the range of association constant(s). For the one-stage Bovine IFN-gamme, our method works best for values of K_{AB} between 10^2 to 10^8. For the two-stage BAF-Emerin complex, our method works best for most combinations over a broad range of K_{AB} values between 10^1 and 10^{11} and K_{AB_2} values between 10^2 and 10^9. Poor scores for the one-stage association at low and high K_{AB} values can be attributed to near-zero fractional masses of initial or final components at those extremes. Likewise, for the two-stage association, poor scores for low K_{AB} values can be attributed to the near zero fractional mass of AB (and hence AB_2) in such cases. The error is also large with high K_{AB_2} values due to the very small amount of AB remaining at equilibrium.

2.3.4 Robustness to monomers and complex size and shape

We also studied the performance of our method on two other complexes that are quite different in molecular weight and structure from the two that have been discussed so far. While the main one-stage study, Bovine IFN-gamma, has monomers that are relatively small and close in molecular weight (14.2 and 13.3 kDa), our additional study, human calcineurin, has monomers that are larger and have very different molecular weights (43.6 and 18.8 kDa) and shapes. The main two stage study, BAF-Emerin complex, has monomers with weights 5.7 and 10.1 kDa while the additional HGH-receptor complex had monomers with weights 21.0 and 22.5 kDa and different shapes.

In both cases, our method inferred the correct pathway and association constants and reconstructed scattering curves that are very similar to the simulated ones. For the one-stage human calcineurin (Supplementary Tab. 2.7), the χ^2 value averaged 1.10 over ten simulated datasets, with association constants averaging 4.24×10^4 (which was the simulated value). The resulting MARDs for the best χ^2 association constant averaged 0.24% for A, 0.16% for B, and 0.22% for AB. As in our initial one-stage study, an alternative two-stage model yielding both AB and AB_2 scored well by χ^2 (1.28), but poorly by MSMRD (1.06×10^{-3}), with substantial disagreement on best association constants ($K_{AB} = 4.30 \times 10^4$, $K_{AB_2} = 5.35 \times 10^2$ for χ^2 and $K_{AB} = 3.00 \times 10^4$, $K_{AB_2} = 5.10 \times 10^2$ for MSMRD). The χ^2 score at the best MSMRD point and the MSMRD score at the best χ^2 point were also worse. Several other pathways scored moderately well by χ^2, but all of these could be eliminated by evaluating the MSMRD scores and the disagreement between best association constants.

Similarly good results were seen for the HGH-receptor complex (Supplementary Tab. 2.8). The lowest χ^2 was on average 0.98 at association constants averaging $K_{AB} = 8.43 \times 10^5$, $K_{AB_2} = 6.26 \times 10^4$ (which were the simulated values). The average MARDs across ten datasets at the lowest χ^2 points were 0.08% for A, 0.09% for B, 0.10% for AB, and 0.10% for AB_2. Alternative models that extend the correct two-stage pathway with $AB_2 + A \rightarrow A_2B_2$ or $AB_2 + B \rightarrow AB_3$ third stages also have low χ^2 scores (1.38 and

1.33, respectively), but poorer MSMRD scores and large disagreement on best association constants.

2.3.5 Contaminated data

A frequent problem in the analysis of associating systems is the presence of "incompetent protein" contaminants, either monomer protein which behaves similarly to ideal material during purification but which does not participate in associations, or oligomers that do not dissociate (irreversible aggregate) [69]. In both cases the protein appears in the initial concentrations but not in any complex. For example, we found in our previous work on homo-oligomers that the addition of 2% of another oligomeric form would lead to large χ^2 values and incorrect association constants and reconstructions [67].

To test the robustness of our method to such contaminants, we used a non-participating fraction of monomer A as a contaminant in our one-stage Bovine IFN-gamma. We also used a non-participating A_{13} aggregate in our two-stage BAF-Emerin complex, using a single aggregated form to represent the total possible contribution from multiple aggregated forms. To construct an A_{13} structure for this simulation, we repeatedly docked copies of A together with GRAMM-X [59]. Scattering curves from all forms were again simulated with CRYSOL. We simulated data with off-grid values of $.0047$, $.0113$, and $.0231$ contaminant mass fraction in the initial mass of A, using the same association constants as before. Ten datasets were generated for each case with different random Gaussian noise.

We first performed our regular coarse- and fine-grid searches on the simulated data with contaminants, assuming as in previous sections the absence of any contaminant (Supplementary Tab. 2.9 and 2.10). All of the alternative (incorrect) association pathways were immediately eliminated due to high χ^2 or inconsistency between best χ^2 and best MSMRD (not shown).

Using the correct association pathway for Bovine IFN-gamma, the χ^2 values increase

monotonically with contaminant fraction. As expected, in the presence of a nonparticipating monomer, the apparent association constants also shift towards smaller values. When the contaminant fraction increases to $.0231$ the χ^2 score has more than doubled and indicates a clear problem in the analysis. The MSMRD scores have also increased significantly (although these scores do not have a standard baseline to reference).

The behavior of the BAF-emerin complex is similar. χ^2 scores also increase monotonically with contaminant fractions. The behavior of the MSMRD score is more variable, perhaps because the A_{13} contaminant used here has an outsized effect on the $I(0)$ values. For the $.0231$ contaminant fraction, even the coarse grid search is unable to identify the nearest grid point. Here again, a significantly increased χ^2 and disagreement between best χ^2 and best MSMRD association constants indicates problems for the $.0113$ and $.0231$ contaminant fractions. Here the increasing presence of the A_{13} contaminant shifts the association constants to larger values forming more of the larger complexes.

In both cases the presence (or suspicion) of an incorrect analysis (particularly the disagreement between best χ^2 and best MSMRD values) would signal the need for a more sophisticated analysis. We have developed a convex quadratic optimization method specifically to deal with problems arising from non-participating contaminants.

We performed grid searches extended to include contaminant fraction for all cases. The coarse contaminant fraction grid dimension ranged from 0 to 0.1 by steps of 0.01. Fine grid searches (including contaminant fraction) were then performed for all pathways with a χ^2 value for the extended coarse grid search within 1.0 of the best χ^2 pathway (note that the MSMRD cannot be used to assess the quality of these searches because scattering curves are only generated upon applying the quadratic optimization). The fine contaminant grid then ranged from the point below the identified coarse grid contaminant fraction to that above it, with a step size of 0.001. The grid searches were performed considering either an A or B homo-oligomeric contaminant (but not both). Optimized scattering intensities were then computed for the best χ^2 fine grid association constants by solving the quadratic

program with constraints and parameter values as presented in Methods.

Tab. 2.3 summarizes fine grid contaminant search results. For the one-stage Bovine IFN-gamma contaminanted with non-participating A, three pathways passed the χ^2 cut off: the correct model and the same two alternatives that were found in the baseline studies. While it is hard to distinguish the three based solely on χ^2, the intensity reconstruction optimization procedure found no feasible solution for the alternative models, but successfully yielded scattering curves for the correct model, in all 10 datasets. For the two-stage BAF-Emerin complex contaminated with the A_{13} aggregate, only the correct model passed the χ^2 filter, and its intensity reconstruction optimization was successful. For both cases and at all contaminant levels, the identified fine grid association constants and contaminant fractions are close to the simulated values (Bovine IFN-gamma $K_1 = 3.43 \times 10^6$; BAF-Emerin complex: $K_1 = 3.21 \times 10^5$ and $K_2 = 4.23 \times 10^5$) and, for the higher contaminant fractions, notably closer than the values obtained in the contaminant-free searches.

Scattering intensities optimized using the quadratic program (labeled OPT) were compared with simulated intensities (labeled TRUE) and those computed by least squares (labeled LSQ) visually (Fig. 2.8) and by calculating MARD (Tab. 2.4). Here the quadratic program is consistently successful. MARD scores are substantially improved for the optimized reconstructions, with the greatest improvement at the higher contaminant fractions, although even the lower ones benefitted, presumably as a result of the added constraints. Examining the scattering curve reveals that the greatest deviations from simulated and the greatest improvement come at small q values. We note that I_B is an independent vector in the intensity matrix, and thus MARDs are the same for the two methods. We found that the reconstructed scattering curve for the contaminating molecule (not shown) was not a close approximation to the true curve, probably due to the extremely small fraction of the contaminant in the solution.

As a final test, we did contaminant grid searches on uncontaminated data ($.0000$ entries in Tables 2.3 and 2.4). This approach did not perform as well as the contaminant-free

Tab. 2.3: Fine grid χ^2 results for contaminated simulations with coarse-grid χ^2 within 1.0 of the lowest scoring model.

Contam	K_1	K_2	χ^2	c_A^*	c_B^*
		Bovine IFN-gamma			
		$A + B \rightarrow AB$			
.0000	4.80e6±6.3e5	-	1.61±0.14	4.1e-3±6.0e-4 (9)	2.00e-3±0.0 (1)
.0047	3.59e6±3.6e5	-	1.51±0.1	5.30e-3±9.5e-4	n/a
.0113	3.41e6±1.3e5	-	1.43±0.1	1.13e-2±6.7e-4	n/a
.0231	3.39e6±1.7e5	-	1.46±0.1	2.34e-2±5.2e-4	n/a
		$A + B \rightarrow AB,\ AB + B \rightarrow AB_2$			
.0000	3.94e13±9.6e13	1.01e11±2.0e11	1.51±0.1	5.18e-2±4.2e-2 (6)	1.08e-2±6.8e-3 (4)
.0047	8.81e14±2.5e15	2.70e12±7.6e12	1.45±0.2	5.39e-2±4.2e-2	n/a
.0113	5.39e12±1.1e13	1.55e9±3.2e9	1.42±0.2	1.37e-2±8.6e-3	n/a
.0231	1.26e14±2.6e14	1.94e11±4.1e11	1.46±0.1	3.11e-2±1.6e-2	n/a
		$A + B \rightarrow AB,\ AB + A \rightarrow A_2B$			
.0000	1.02e10±3.2e10	1.33e6±3.1e6	1.60±0.1	n/a	1.70e-2±2.2e-2
.0047	3.91e12±1.2e13	6.63e9±2.1e10	1.55±0.3	n/a	1.66e-2±2.3e-2
.0113	7.25e6±5.1e6	6.65e3±9.1e3	1.40±0.0	1.04e-3±8.4e-4	n/a
.0231	4.59e10±1.5e11	8.75e5±2.7e6	1.19±0.4	2.24e-2±7.3e-4 (9)	8.00e-2±0.0 (1)
		BAF-Emerin complex			
		$A + B \rightarrow AB,\ AB + B \rightarrow AB_2$			
.0000	5.44e5±4.3e3	8e5±0.0	1.78±0.10	n/a	6.60e-3±5.2e-4
.0047	6.87e5±9.2e5	1.04e6±1.6e6	1.74±0.0	1.00e-2±0.0 (8)	8.5-3±7.1e-4 (2)
.0113	3.13e5±1.2e4	4.08e5±2.2e4	1.49±0.1	1.18e-2±9.2e-4	n/a
.0231	3.60e5±1.7e4	4.92e5±3.0e4	1.56±0.1	2.09e-2±8.8e-4	n/a

* The search considers only A or B contaminant; rows with values for both c_A^* and c_B^* are due to different identified contaminants for different simulations (number of times in parentheses).

Tab. 2.4: MARDs (%) for contaminated reconstructions.

	Contam	method	I_A	I_B	I_{AB}	I_{AB_2}
Bovine IFN-gamma	.0000	LSQ	0.83±0.2	0.28±0.2	0.41±0.1	-
		OPT	0.27±0.1	0.24±0.1	0.22±0.1	-
	.0047	LSQ	0.53±0.1	0.17±0.0	0.17±0.0	-
		OPT	0.20±0.0		0.09±0.0	-
	.0113	LSQ	1.13±0.1	0.16±0.0	0.34±0.0	-
		OPT	0.42±0.0		0.14±0.0	-
	.0231	LSQ	2.28±0.1	0.17±0.0	0.69±0.0	-
		OPT	0.83±0.0		0.29±0.0	-
BAF-Emerin complex	.0000	LSQ	0.08±0.0	0.64±0.1	0.20±0.0	0.20±0.0
		OPT	Unfeasible			
	.0047	LSQ	2.32±0.0	0.08±0.0	0.85±0.0	0.56±0.0
		OPT	1.91±0.0		0.77±0.1	0.53±0.0
	.0113	LSQ	4.45±0.1	0.08±0.0	0.90±0.1	0.42±0.1
		OPT	2.41±0.2		0.56±0.1	0.27±0.1
	.0231	LSQ	8.72±0.1	0.08±0.0	1.41±0.1	0.60±0.1
		OPT	1.31±0.1		0.31±0.1	0.21±0.0

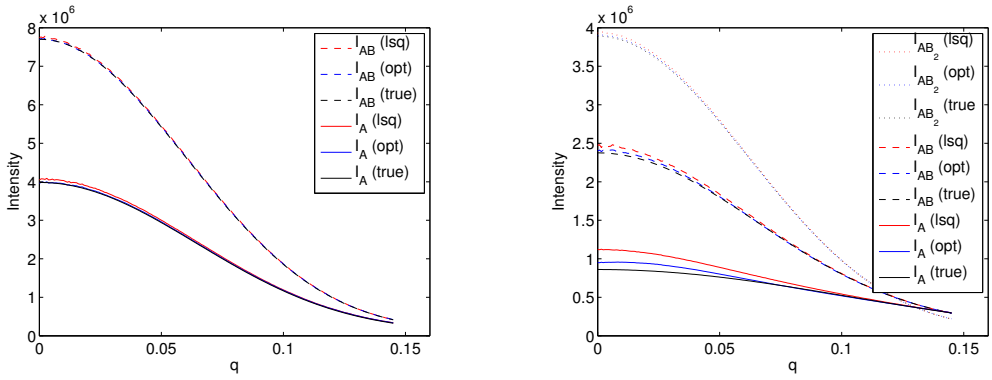


Fig. 2.8: Simulated intensities compared with reconstructed ones computed by the quadratic program (opt) and the initial least squares $\tilde{O}_0$ (lsq), for one .0231 contaminant fraction dataset of Bovine IFN-gamma (left) and BAF-Emerin complex (right). The I_B reconstruction, which is independent of contaminant, is not shown.

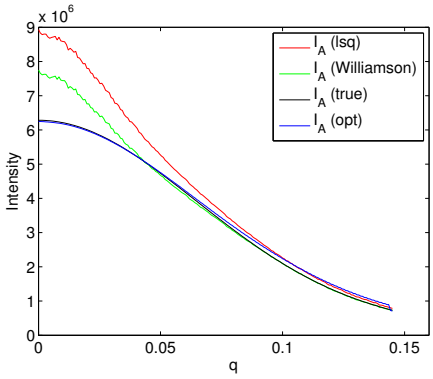


Fig. 2.9: Reconstructed pure monomer intensity from a monomer-tetramer-octamer association contaminated with 16-mer. I_A (Williamson) is computed using the original grid search without contaminant fraction and subsequent intensity reconstruction as in [67].

search on uncontaminated data. As expected, fitting the additional contaminant parameter has driven the association constants somewhat away from their best values.

2.3.6 Application of contaminant methods to homo-oligomers

Contamination with aggregates proved to be a problem for our earlier method for characterizing homo-oligomers [67]. Thus we performed our contaminant search and reconstruction on the case studied there: octameric purE from *E. coli* (PDB id 1QCZ) [38], under a monomer-tetramer-octamer association with a 2% mass fraction of a 16-mer as a contaminant. The best association constants resulting from our contaminant search were $K_{12} = 4.00 \times 10^{12}$, $K_{23} = 1.25 \times 10^{1}$, close to the simulated association constants $K_{12} = 2.87 \times 10^{12}$, $K_{23} = 1.29 \times 10^{1}$; although the identified contaminant fraction was higher than simulated, at 6.6%. The association model found by the previous method [67] was $K_{12} = 3.46 \times 10^{12}$, $K_{23} = 1.00 \times 10^{1}$, also close to the simulated association constants. However, our reconstructed monomer scattering curve is much better than the previous one, whose χ^2 is four times worse. The optimized monomer intensity curve is much closer to the simulated curve than that computed by least squares (after a contaminant search), and that found without contaminant search (as in [67]), especially at low q (Fig. 2.9). As we

can see, contaminant search plus the quadratic program reconstruction produce a curve that closely approximates the true one, while the contaminant-free and least squares reconstructions introduce substantial error. We note again that the least squares curve is just one of the infinitely many satisfying solutions, and thus it is not too surprising that it is actually much worse. The curves for tetramer and octamer are not plotted since for both methods they are extremely similar to the true curves. These results demonstrate that our method can also be profitably applied to homo-oligomers in the presence of contaminants.

2.4 Discussion

We have presented a method to infer an association model (pathway and association constants), along with the underlying scattering curves of the initial components and intermediate and final complexes, from solution scattering data for a set of equilibrium mixtures undergoing hetero-association with different initial component concentrations. Our method searches over possible association models and contaminant fractions, solving for the reconstructions of the underlying scattering curves by a least squares method in the absense of "incompetent protein" contaminants or by a convex quadratic program in their presence. The model and scattering curves are evaluated in terms of how well they can then reconstruct de-noised input data. We use two complementary scores, a χ^2 to assess the overall fit between the data and the association model combined with reconstructed scattering, and the MSMRD to assess the consistency between the association model stoichiometry and the reconstructed scattering. The convex quadratic program provides an optimization-based method for the difficult problem of reconstructing the underlying scattering curves in the presence of either non-participating monomers or irreversible aggregates.

In a variety of simulated test cases covering one- and two-stage association pathways, our approach correctly determined the pathway, accurately estimated the association constants with generally less than 2% error, and accurately reconstructed the scattering curves

to within an average deviation of less than 0.25%. While such accuracy cannot be expected for all experimental scattering data, the potential for such accurate evaluation exists in the most favorable cases. The good accuracy for reconstructing the scattering curve bodes well for the application of 3D structural modeling based on the reconstructed scattering curves. We found that the χ^2 and MSMRD were effective as complementary metrics. Cases where an alternative model with an extra association step obtained a fairly good χ^2 value could be ruled out by a greater MSMRD and inconsistency between the best scoring association constants under one metric vs. the other. We also found our method to be amenable to a range of association constants, Gaussian noise levels, different complex sizes and shapes, and contaminants.

The range of association constants that were found acceptable for our method (Fig. 2.7) compares well with the range of 10^4 to 10^9 routinely available from analytical ultracentrifugation [31] while also revealing the molecular weight of each complex (via $I(0)$ calculations) calibrated by the molecular weights of the initial components. At the same time the SAS method provides complex scattering curves that can serve as the basis for 3D reconstruction. In addition, this range of affinities is explored with the same fixed set of initial concentrations used in the earlier simulation. The initial concentrations could also be adjusted upwards to explore weaker interactions (limited by the solubility of the proteins) and downwards to explore stronger ones (limited by the strength of observed scattering). We note that the strongest beam lines at third generation sources can generate an accurate scattering profile at concentrations as low as 0.05 mg/ml (Williamson and Friedman, unpublished results).

At realistic contaminant levels, our method was able to reconstruct the scattering curves quite accurately, a result not possible by previous methods that assumed an absence of contaminants. While by no means perfect, the objective and set of constraints we have chosen yield good solutions in practice. Smoothness is taken as the primary objective, and the potential for over-smoothing is mitigated by a counterbalancing constraint from the χ^2

constraint. Other constraints could potentially be incorporated in order to encode shape characteristics and relationships between the different forms. We are not able to adequately determine the exact contaminant fraction or its scattering curve, but the incorporation of additional constraints could help. Extensions to other forms of contamination and systematic noise may be amenable to analogous techniques.

In our test cases, we included pure A and pure B as two of our samples. This suggests an alternative strategy to use the intensity curves of these pure samples to reduce the number of unknowns (removing known intensity column vectors for A and B in $\tilde{O}$) from our computations. However, when contaminants are present, there may be no such thing as a "pure" sample. Likewise, our approach works with a self-associating system which does not contain pure monomers at the lowest concentration. We have shown that even without pure A and pure B in the input, we can obtain the correct model as long as the samples are diverse enough. We were able to do this successfully for two of the test cases where additional samples at other concentrations replaced the pure samples.

Acknowledgement

This work was supported in part by National Science Foundation grant IIS-0502801 to C.B.K., A.M.F., and B.A.C. and CCF-0915388 to C.B.K., along with National Institutes of Health grant R01 GM-65982 to Bruce Randall Donald (Duke University).

2.5 Supplementary Material for "Stoichiometries and affinities of interacting proteins from concentration series of solution scattering data: Decomposition by least squares and quadratic optimization"

The supplementary material includes tables for the initial concentrations (one-stage in Tab. 2.5 and two-stage in Tab. 2.6), the complete results for the additional test cases (Human calcineurin in Tab. 2.7 and HGH-receptor complex in Tab. 2.8), and the results of running our contaminant-free searches on data containing a simulated contaminant (Tab. 2.9 and Tab. 2.10).

Tab. 2.5: Initial concentrations and fractional masses for one-stage simulations ($A + B \rightarrow AB$).

Init Conc A (mg/ml)	Init Conc B (mg/ml)	Fractional Masses					
		1AUI			1D9G		
		A	B	AB	A	B	AB
0.5	1.0	0.108	0.610	0.282	0.001	0.357	0.642
1.0	1.0	0.182	0.396	0.422	0.010	0.043	0.947
2.0	1.0	0.281	0.207	0.512	0.310	0.001	0.689
3.0	1.0	0.360	0.123	0.517	0.483	0.000	0.517
4.0	1.0	0.429	0.079	0.493	0.586	0.000	0.414
4.5	1.0	0.458	0.065	0.477	0.624	0.000	0.376
0.5	0.0	1.000	0.000	0.000	1.000	0.000	0.000
0.0	1.0	0.000	1.000	0.000	0.000	1.000	0.000

Tab. 2.6: Initial concentrations and fractional masses for two-stage simulations ($A + B \rightarrow AB$, $AB + B \rightarrow AB_2$).

Init Conc A (mg/ml)	Init Conc B (mg/ml)	Fractional Masses							
		3HHR				2ODG			
		A	B	AB	AB_2	A	B	AB	AB_2
1.0	1.0	0.115	0.017	0.661	0.206	0.294	0.004	0.359	0.343
3.0	1.0	0.525	0.001	0.453	0.021	0.629	0.001	0.276	0.095
1.0	3.0	0.000	0.253	0.073	0.674	0.016	0.012	0.132	0.839
1.0	5.0	0.000	0.486	0.019	0.496	0.000	0.249	0.004	0.747
5.0	1.0	0.681	0.001	0.311	0.008	0.749	0.000	0.207	0.045
5.0	5.0	0.108	0.004	0.666	0.223	0.293	0.001	0.360	0.346
5.0	0.0	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
0.0	5.0	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000

Tab. 2.7: Human Calcineurin association model searches, over 10 sets of simulated noise. Pathways are abbreviated by the reaction products; e.g., (AB) means $A + B \rightarrow AB$, while (AB, AB_2) means $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$. The bolded entries obtain the best scores, and are for the correct pathway. The simulated association constant was $K_1 = 4.24e4$.

	χ^2			MSMRD		
Pathway	K_1±SD	K_2±SD	Score±SD	K_1±SD	K_2±SD	Score±SD
(AB)	**4.00e4±0.0e0**	-	**1.10±0.05**	**4.00e4±0.0e0**	-	**5.03e-5±2.4e-5**
(AB) **Fine**	**4.24e4±9.7e2**	-	**1.08±0.05**	**4.24e4±5.9e2**	-	**6.96e-9±7.1e-9**
(AB_2)	6.00e7±0.0e0	-	2.91±0.10	3.00e8±0.0e0	-	3.55e-4±8.4e-5
(A_2B)	3.20e18±4.2e17	-	120±0.90	3.00e7±0.0e0	-	1.00e-3±1.6e-4
(A_2)	3.00e14±0.0e0	-	327±2.8	3.00e3±0.0e0	-	1.88e-4±5.1e-5
(B_2)	7.00e3±0.0e0	-	771±3.6	6.70e17±1.2e18	-	4.57e-2±1.3e-3
(AB, AB_2)	4.30e4±4.8e3	5.35e2±7.8e2	1.28±0.08	3.00e4±0.0e0	5.10e2±7.4e1	1.06e-3±3.3e-4
(B_2, AB_2)	6.30e2±4.8e1	1.00e8±0.0e0	1.92±0.09	3.00e3±0.0e0	1.00e9±0.0e0	4.52e-3±3.4e-4
(AB, A_2B)	7.30e4±4.7e4	9.41e3±1.5e4	1.29±0.10	5.60e4±7.0e3	7.5e1±1.8e1	8.15e-4±2.9e-4
(A_2, A_2B)	8.20e5±1.9e5	5.27e12±1.7e12	1.31±0.10	5.80e3±1.9e3	2.24e8±1.8e8	2.79e-2±3.1e-3

Tab. 2.8: HGH-receptor complex association model searches, over 10 sets of simulated noise. Pathways are abbreviated by the reaction products; e.g., (AB, AB_2) means $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$. The bolded entries obtain the best scores, and are for the correct pathway. The simulated association constants were $K_1 = 8.43e5$ and $K_2 = 6.26e4$.

	χ^2				MSMRD			
Pathway	K_1±SD	K_2±SD	K_3±SD	Score±SD	K_1±SD	K_2±SD	K_3±SD	Score±SD
(AB)	6.00e4±0.0e0	-	-	11900±9.2	2.20e6±4.2e5	-	-	5.25e-6±1.9e-6
(AB_2)	9.00e9±0.0e0	-	-	1010±3.3	2.00e8±0.0e0	-	-	1.57e-5±3.9e-6
(A_2B)	5.00e17±0.0e0	-	-	26200±15	1.00e6±0.0e0	-	-	4.60e-3±8.3e-5
(A_2)	6.00e2±0.0e0	-	-	10000±2.9	1.00e-6±2.2e-22	-	-	2.50e-1±0.0e0
(B_2)	1.00e19±0.0e0	-	-	45600±10	1.00e19±0.0e0	-	-	2.97e-2±1.1e-4
(AB, AB_2)	**8.00e5±0.0e0**	**6.00e4±0.0e0**	-	**1.14±0.00**	**8.00e5±0.0e0**	**6.00e4±0.0e0**	-	**3.22e-6±1.7e-6**
(AB, AB_2) Fine	**8.43e5±9.1e3**	**6.26e4±3.1e2**	-	**0.98±0.15**	**8.45e5±2.2e4**	**6.27e4±8.4e2**	-	**3.37e-10±3.4e-10**
(B_2, AB_2)	4.00e6±0.0e0	9.00e15±0.0e0	-	440±2.6	9.10e2±1.5e2	4.70e8±4.8e7	-	7.14e-06±3.3e-06
(AB, A_2B)	2.00e4±0.0e0	9.00e4±0.0e0	-	2190±4.5	1.00e11±0.0e0	3.00e2±0.0e0	-	1.91e-2±7.4e-5
(A_2, A_2B)	3.00e6±0.0e0	6.00e13±0.0e0	-	11100±7.2	5.00e2±0.0e0	9.00e6±0.0e0	-	8.92e-4±8.5e-5
(A_2, B_2, A_2B_2)	9.00e10±0.0e0	4.00e4±0.0e0	2.00e6±0.0e0	2620±4.80	9.00e10±0.0e0	1.60e5±5.2e4	6.00e2±0.0e0	7.47e-2±2.3e-4
(AB, AB_2, A_2B_2)	8.00e5±0.0e0	6.00e4±0.0e0	1.10e1±3.2e00	1.38±0.19	2.30e4±4.8e3	1.00e3±0.0e0	1.00e1±0.0e0	1.27e-2±3.6e-5
(AB, A_2B, A_2B_2)	9.70e3±4.8e2	8.00e3±1.3e3	2.00e3±0.0e0	2340±14	9.00e10±0.0e0	9.00e6±0.0e0	5.00e7±0.0e0	2.39e-1±4.4e-4
(AB, AB_2, AB_3)	8.80e5±4.2e4	5.00e4±1.63e4	6.81e4±1.5e5	1.33±0.30	8.3e5±1.1e5	7.20e4±4.2e3	3.40e1±1.5e1	7.71e-5±5.1e-5

Tab. 2.9: Contaminant-free search results for varying levels of contaminant.

	χ^2			MSMRD		
Contam %	K_1±SD	K_2±SD	Score±SD	K_1±SD	K_2±SD	Score±SD
Bovine IFN-gamma, $A + B \rightarrow AB$						
0.00	3.40e6±7.1e4	-	1.49±0.12	3.64e6±5.2e5	-	1.82e-11±1.8e-11
0.47	2.81e6±6.2e4	-	1.48±0.09	2.36e6±2.4e5	-	1.65e-10±1.4e-10
1.13	2.22e6±3.4e4	-	1.93±0.07	1.57e6±1.8e5	-	4.81e-10±5.9e-10
2.31	1.52e6±2.6e4	-	3.56±0.35	9.19e5±7.4e4	-	1.98e-10±4.6e-10
BAF-Emerin complex, $A + B \rightarrow AB,\ AB + B \rightarrow AB_2$						
0.00	3.21e5±4.5e3	4.24e5±6.9e3	1.12±0.24	3.24e5±1.3e4	4.29e5±1.9e4	1.03e-9±6.6e-10
0.47	4.23e5±6.0e3	5.98e5±8.2e3	1.60±0.08	3.14e5±2.7e4	3.90e5±3.1e4	4.07e-5±1.1e-5
1.13	6.53e5±2.4e4	1.02e6±4.8e4	2.10±0.09	1.00e5±0.0e0	1.10e5±1.5e3	1.09e-6±7.3e-7
2.31	3.57e15±7.2e14	6.23e15±1.2e15	2.68±0.16	6.35e4±1.1e3	5.75e4±8.1e2	1.64e-9±9.1e-10

Tab. 2.10: Mean MARDs (%) for the best fine grid points resulting from contaminant-free searches for varying levels of contaminant.

Contam %	I_A	I_B	I_{AB}	I_{AB_2}
Bovine IFN-gamma, $A + B \rightarrow AB$				
0.47	0.23	8.10	3.77	-
1.13	0.45	8.10	3.77	-
2.31	0.94	7.90	3.74	-
BAF-Emerin complex, $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$				
0.47	1.33	33.13	23.94	27.93
1.13	3.25	33.15	23.92	28.02
2.31	6.57	33.13	23.83	27.98

3. NMR STRUCTURAL INFERENCE OF SYMMETRIC HOMO-OLIGOMERS

H. Chandola, A. K. Yan, S. Potluri, B. R. Donald and C. Bailey-Kellogg. NMR structural inference of symmetric homo-oligomers. *J. Comp. Biol.*, 12:1757–1775, 2011.

Abstract

Symmetric homo-oligomers represent a majority of proteins, and determining their structures helps elucidate important biological processes including ion transport, signal transduction, and transcriptional regulation. In order to account for the noise and sparsity in the distance restraints used in NMR structure determination of cyclic (C_n) symmetric homo-oligomers, and the resulting uncertainty in the determined structures, we develop a Bayesian structural inference approach. In contrast to traditional NMR structure determination methods, which identify a small set of low-energy conformations, the inferential approach characterizes the entire posterior distribution of conformations. Unfortunately, traditional stochastic techniques for inference may undersample the rugged landscape of the posterior, missing important contributions from high-quality individual conformations and not accounting for the possible aggregate effects on inferred quantities from numerous unsampled conformations. However, by exploiting the geometry of symmetric homo-oligomers, we develop an algorithm that provides provable guarantees for the posterior distribution and the inferred mean atomic coordinates. Using experimental restraints for three proteins, we demonstrate that our approach is able to objectively characterize the structural diversity supported by the data. By simulating spurious and missing restraints, we further demonstrate that our approach is robust, degrading smoothly with noise and sparsity.

3.1 Introduction *

Protein structure determination by nuclear magnetic resonance (NMR) spectroscopy provides insights into functional mechanisms, dynamics, and interactions of proteins in solution. Traditionally, NMR structure determination has been formulated as an optimization problem [6,20,21], seeking a minimum-energy structure according to a potential that evaluates both agreement with experimental data (e.g., distance restraints) and biophysical quality according to an empirical molecular mechanics energy function. Because traditional methods typically employ heuristic optimization methods, they are subject to the problem of only finding local minima. As a result, traditional methods are repeated many times in the hope that the global optimum is captured in the ensemble of generated structures. Identification of an optimum is especially difficult in cases where the data are noisy, sparse, and/or ambiguous. While the computed ensemble illustrates structural variability, it does not provide an objective measure of the uncertainty in atomic coordinates, because different members of the ensemble may have different likelihoods. In addition, the traditional NMR ensemble does not provide guarantees that all plausible solutions have been discovered.

In contrast to optimization-based approaches, Nilges and co-workers [48] cast protein structure determination by NMR as a statistical inference problem, *inferential structure determination* (or *structural inference*), in which the goal is to compute the posterior distribution of plausible structures. The posterior captures both the satisfaction of restraints (as a likelihood) and biophysical modeling terms (as a prior). The inferential approach provides an objective measure of confidence and is not focused on trying to find a single "optimal" solution (or ensemble of solutions that are optimal in different runs). Nilges and co-workers developed a sampling-based method to perform structural inference for monomers, and applied it to characterize the posterior distribution of the 59-residue Fyn SH3 domain, given 154 NOE restraints [48].

*Abbreviations used: NMR: Nuclear Magnetic Resonance, NOE: nuclear Overhauser effect, RMSD: root-mean-square deviation, SCS: symmetry configuration space, SO(3): Special Orthogonal Group, vdW: van der Waals

We develop here an algorithm that performs structural inference for symmetric homo-oligomers—protein complexes comprised of identical subunits (monomer proteins) arranged symmetrically. Symmetric homo-oligomers are a valuable target since they make up *a majority of proteins* [18]; they play pivotal roles in important biological processes including ion transport and regulation, signal transduction, and transcriptional regulation. Experimentally, it is possible to distinguish intra-subunit restraints (distance restraints between atoms within a subunit) from inter-subunit ones (distance restraints between atoms in different subunits) by isotopic labeling strategies and X-filtered NOESY experiments [24,32,63,71]. As a result of this, the complex structure determination can proceed by first determining the subunit structure *as it exists in complex*, and then computing the oligomeric assembly [43,50,65]. Our problem (Fig. 3.1) is thus to compute the posterior distribution of homo-oligomeric complex structures given the subunit structure, by evaluating their consistency with experimental data and their packing quality. Having computed the posterior distribution over complex structures, we also infer other quantities of interest, namely the means and variances of atomic coordinates.

Our inference algorithm characterizes the *entire* posterior distribution of a homo-oligomeric complex structure, to within user-specified thresholds on allowed error in computing the posterior over structures and the mean atomic coordinates. *Error guarantees* are possible due to our focus on symmetric homo-oligomers, whose complex structures can be specified in terms of their symmetry axes, enabling us to employ a four degree-of-freedom representation which we call the *symmetry configuration space* (SCS). We build upon our earlier work on searching symmetry configuration spaces [44, 46], but this paper represents a significant extension in order to support inference and compute error bounds, which account for experimental noise and uncertainty. Our algorithmic approach, hierarchical subdivision with error guarantees, stands in contrast to sampling techniques, such as the replica-exchange MCMC algorithm employed by Nilges and co-workers [22, 48], which may under-sample the high-dimensional and very rugged posterior distribution of a

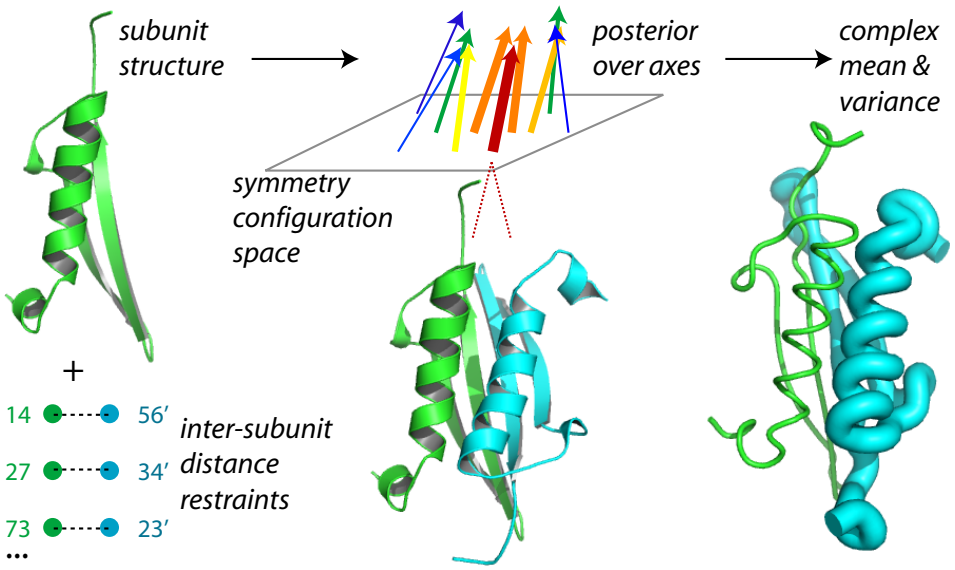


Fig. 3.1: Structural inference of symmetric homo-oligomers. Given a subunit structure and set of inter-subunit distance restraints, we compute the posterior distribution over all possible complex structures, represented in terms of a configuration space of symmetry axes. The posterior distribution evaluates the quality (depicted via color and thickness) of the satisfaction of the restraints and the packing of the subunits. By integrating over the posterior distribution over axes (and thereby structures) we obtain means and variances for atomic coordinates, depicted as a sausage plot (thicker implying greater variance). In the homo-dimer shown here, we fix one subunit and evaluate possible axes and thereby positions of the other subunit.

monomer, and does not characterize (or place bounds on) the error in inferred quantities. Unlike sampling methods, we account for both the individual and aggregate effects of leaving out possible conformations. That is, *by applying provable bounds on the error of the posterior (including the underlying normalization constant), we ensure that we have not missed any high quality conformations or a large number of lower quality conformations, either of which could result in incorrect inferences.*

3.2 Methods

As mentioned in the Introduction, the subunit structure as it exists in complex can be obtained prior to complex structure determination [43, 50, 65]. Thus, we are given the subunit structure (Euclidean coordinates $\mathbf{p}_i$ for each atom i) and a set of n distance restraints

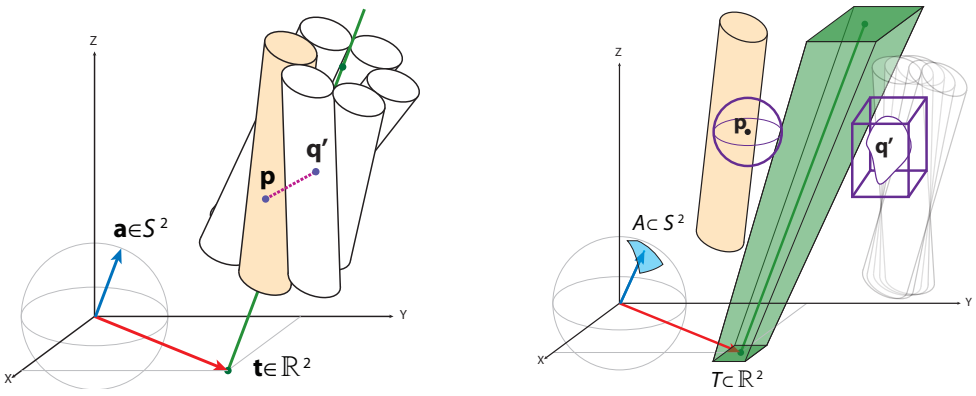


Fig. 3.2: Symmetry configuration space (SCS). (a) Each structure is defined by a point $(\mathbf{a}, \mathbf{t}) \in \mathbb{S}^2 \times \mathbb{R}^2$ in the configuration space of symmetry axes. Each subunit is depicted by a cylinder; the structure is obtained by rotating the fixed subunit (shaded cylinder) by the angle of symmetry around the symmetry axis (line). A distance restraint is shown between an atom at position $\mathbf{p}$ on the fixed subunit and one at position $\mathbf{q}'$ on the adjacent subunit. (b) An SCS cell $C \subset \mathbb{S}^2 \times \mathbb{R}^2$ defines a set of symmetry axes (green region) and thereby a corresponding set of structures. We can bound the possible positions $\mathbf{q}'$ over these structures.

$R = \{r_1, \ldots, r_n\}$ (each specifying an atomic pair and allowed distances). We assume here that the oligomeric number has also been previously determined (e.g., from ultracentrifugation), but see Potluri *et al.* [44] for a discussion of how to score possible oligomeric states based on how well the restraints fit as well as empirical energy functions. If we fix the position of the initial subunit structure, then the homo-oligomeric complex structure is completely specified by the symmetry axis (Fig. 3.2(a)). We focus on cyclic symmetry C_n, in which we position at the origin one *fixed* subunit, and obtain the complex structure by rotating the fixed subunit structure around the symmetry axis c to generate the other subunit(s). Thus the symmetry axis c, can be used to parametrize all possible oligomer structures.

We compute the posterior distribution $p(c \mid R)$ over oligomer structures in terms of the symmetry axis c. Given the posterior, we also infer the expectation $\mathrm{E}(\mathbf{q}_{ij} \mid R)$ and variance $\mathrm{var}(\mathbf{q}_{ij} \mid R)$ of the atomic coordinates $\mathbf{q}_{ij}$ for each atom i in each rotated subunit j. (See again Fig. 3.1.) Unfortunately, the posterior distribution is difficult to compute

and to integrate analytically, and in cases of sparse and noisy data, sampling methods may get trapped in local minima and may miss important contributions to the posterior, either individually or in aggregate. In contrast, we approximate the integral with a discrete sum over *cells* defining contiguous sets of axes at a resolution that is sufficiently fine to consider the axes as making a uniform contribution. While there are too many such cells to simply enumerate all of them, we recognize that many have a sufficiently small posterior that they can be safely ignored without impacting our inferences. Thus we develop a hierarchical subdivision algorithm (Fig. 3.3) to find the high-quality cells and provide guarantees on the resulting error introduced due to eliminating other cells. The algorithm also obeys restrictions on the allowed error in expected atomic coordinates inferred from the cells it returns.

We first summarize our earlier work on representing and computing with a configuration space representation of symmetry axes (Sec. 3.2.1). We then present our inferential framework based on this representation (Sec. 3.2.2), our error bounds (Sec. 3.2.3), and our hierarchical subdivision algorithm for computing the posterior and performing the inference (Sec. 3.2.4).

3.2.1 Symmetry configuration space

For cyclic symmetry, C_n, the symmetry is completely specified by a line representing its axis. The line representing the symmetry axis can be specified by the position where it intersects the xy plane at (x, y), relative to the fixed subunit at the origin, and its orientation (θ, ϕ), relative to the major axis of the fixed subunit which we orient along the z-axis. Thus all possible axes belong to a *symmetry configuration space* (SCS), $\mathbb{S}^2 \times \mathbb{R}^2$ [44], with orientations from the two-sphere $\mathbb{S}^2$ and translations from the xy plane $\mathbb{R}^2$. See Fig. 3.2(a).

Given a symmetry axis $c = (\mathbf{a}, \mathbf{t}) \in \mathbb{S}^2 \times \mathbb{R}^2$ and an angle of rotation $\alpha = 2\pi j/m$ for subunit $j \in \{1, \ldots, (m-1)\}$ (treating the 0^{th} subunit as the fixed one), we compute the

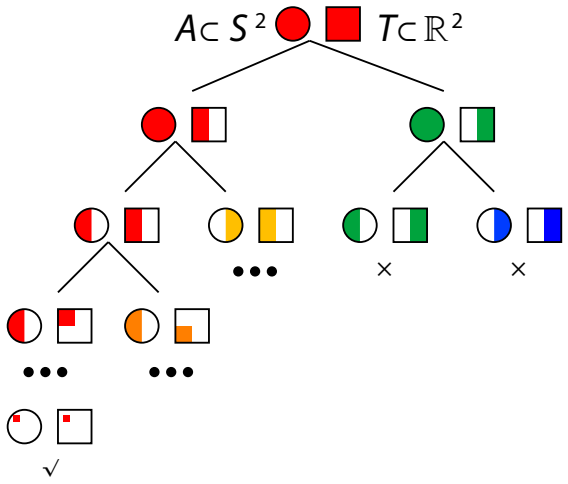


Fig. 3.3: Hierarchical subdivision of SCS. The 4-dimensional SCS is depicted as two 2D regions, a sphere representing the orientation space $\mathbb{S}^2$ and a rectangle representing the translation space $\mathbb{R}^2$. We compute a bound for the best posterior of a configuration in the cell (shaded red for high posterior to blue for low posterior), and recursively subdivide cells. Ultimately (bottom left of the tree) we find cells representative of structures with high posterior, and can eliminate cells (right side of the tree) guaranteed to have a total probability mass less than a user-specified cutoff.

coordinates $\mathbf{q}'$ for an atom in a rotated subunit from corresponding coordinates $\mathbf{q}$ for the same atom in the fixed subunit:

$$\mathbf{q}' = \mathcal{T}(c, \mathbf{q}, \alpha) = \mathcal{R}_{\mathbf{a}}(\alpha)(\mathbf{q} - \mathbf{t}) + \mathbf{t} \tag{3.1}$$

where $\mathcal{R}_{\mathbf{a}}(\alpha) \in SO(3)$ is a rotation by α radians around the unit vector $\mathbf{a}$.

In our algorithm for computing the posterior, we consider simultaneously a set of axes in a *cell* of the SCS (Fig. 3.2(b)). An SCS cell is given by the Cartesian product of the individual lengths in each of the four dimensions $[x_l, x_h] \times [y_l, y_h] \times [\theta_l, \theta_h] \times [\phi_l, \phi_h]$. Note that the SCS cell represents a continuously infinite set of structures. We previously derived a geometric bound, using convex hulls and/or axis-aligned bounding boxes, for the possible coordinates $\mathbf{q}'$ under rotation by α around an axis c in a cell C [44].

$$\mathbf{q}' \in \mathcal{B}(C, \mathbf{q}, \alpha) \tag{3.2}$$

We can use this geometric bound on the rotated positions to evaluate feasibility of a distance restraint within a cell. Consider distance restraint r_i on the distances between positions $\mathbf{p}_i$ and $\mathbf{q}'_i$, where the first atom is in the fixed subunit and the second in the neighboring subunit in the cyclic symmetry, rotated by $2\pi/m$ for oligomeric number m. We geometrically bound the minimum $l(C)$ and maximum $u(C)$ distances between these positions under rotations around axes $c \in C$:

$$l_i(C) \leq \min_{c \in C} \|\mathbf{p}_i - \mathcal{T}(c, \mathbf{q}_i, \alpha)\| \geq u_i(C) \tag{3.3}$$

For the geometric computations giving these bounds, we refer the reader to Potluri *et al.* [44], in particular the subsection "Bound from SCS" in the Methods. In Sec. 3.2.3, we apply these bounds to derive upper and lower bounds on the posterior $p(c \mid R)$.

3.2.2 Inferential framework

We develop here a Bayesian model for the posterior distribution over axes $p(c \mid R)$, along with expectations and variances of atomic coordinates. Our basic framework is like that of Nilges and co-workers [40]. However, our formulation exploits the symmetry in the problem and thus expresses the distribution in terms of the four-dimensional symmetry configuration space.

To compute posterior $p(c \mid R)$, we apply Bayes' rule and integrate out a nuisance parameter σ that is independent of c and encodes the error in the system including both experimental noise and systematic effects such as internal dynamics [33] and spin diffusion [36].

$$p(c \mid R) = \int p(c, \sigma \mid R) \, \mathrm{d}\sigma \tag{3.4}$$

$$\propto \int p(R \mid c, \sigma) \, p(c) \, p(\sigma) \, \mathrm{d}\sigma \tag{3.5}$$

In the following sections we individually examine the various factors: likelihood $p(R \mid c, \sigma)$

and priors $p(c)$ and $p(\sigma)$. We then consider how to properly integrate over the configuration space and infer quantities in the conformation space.

Restraints likelihood $p(R \mid c, \sigma)$

The distance restraints R are conditionally independent given the structure (defined by c):

$$p(R \mid c, \sigma) = \prod_{i=1}^{n} p(r_i \mid c, \sigma) \tag{3.6}$$

To evaluate a single restraint r_i, we adopt the log-normal distribution advocated by Nilges and co-workers [23, 47] as a better representation of the errors in NOE distances and NMR data than the traditional flat-bottom harmonic well (FBHW). The FBHW suffers from problems including subjectiveness associated with fixing the bounds for the well [41]; the log-normal more gracefully degrades and we integrate out its variance parameter σ. Furthermore, the log-normal is non-negative and multiplicative.

Thus given a symmetry axis c and variance σ, the inter-subunit NOE restraint r_i has a log-normal likelihood over the observed distances between atoms in the restraint:

$$p(r_i \mid c, \sigma) = \frac{1}{\sqrt{2\pi}\sigma d_i} \exp\left(-\frac{1}{2\sigma^2} \log^2 g_i(c)\right) , \tag{3.7}$$

where, to abbreviate subsequent equations, we define $g_i(c)$ for cell c and restraint i as the ratio between the desired and actual distances for the restrained pair of atoms:

$$g_i(c) = \frac{d_i}{\|\mathbf{p}_i - \mathcal{T}(c, \mathbf{q}_i, 2\pi/m)\|} . \tag{3.8}$$

The position $\mathbf{p}_i$ is on the fixed subunit and $\mathbf{q}'_i$ is taken on the neighboring subunit (obtained by rotating position $\mathbf{q_i}$ in the fixed subunit by $2\pi/m$).

Prior $p(\sigma)$

The log-normal variance σ is a classical example of a nuisance parameter. Thus its prior is derived through Jeffrey's method of maximizing Fisher's information index [25]:

$$p(\sigma) = 1/\sigma \ . \tag{3.9}$$

Prior $p(c)$

Laplace postulated [27] that in the absence of sufficient reason, each point in the parameter space should be assigned a uniform prior. We follow the same rule and assign equal probability to those symmetry axes that yield structures without steric clashes. In order to produce a data-driven inferential approach, we currently use a weak prior, only distinguishing whether or not a structure has steric clashes.

$$p(c) = \begin{cases} 0, & \text{if complex structure has steric clash} \\ 1, & \text{otherwise.} \end{cases} \tag{3.10}$$

If desired, a stronger prior, incorporating energy evaluation based on a molecular mechanics force field, could instead be employed.

Sec. 3.2.2 below details how to appropriately define (and integrate) such a probability distribution over the SCS parameterization.

Marginalizing over σ

Since σ denotes the experimental and systematic error, we can integrate over all possible values of σ to eliminate it:

$$\begin{aligned} p(c \mid R) &= \int_0^\infty p(c, \sigma \mid R)\, \mathrm{d}\sigma \\ &\propto p(c) \int_0^\infty \sigma^{-(n+1)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \log^2 g_i(c)\right)\, \mathrm{d}\sigma \\ &\propto p(c) \left(\sum_{i=1}^{n} \log^2 g_i(c)\right)^{-n/2} . \end{aligned} \tag{3.11}$$

The integral is obtained by substituting for $\frac{1}{2\sigma^2} \sum_{i=1}^{n} \log^2 g_i(c)$ and multiplying the numerator and denominator by $2^{n/2-1} \,/\, \sum \log^2 g_i(c)$. This yields a gamma function expressed as an integral. Since the gamma function is in terms of n, a constant, we drop it in the proportionality of Eq. 3.11.

Probability distributions in SCS

While our symmetry configuration space representation greatly reduces the degrees of freedom and thus leads to a better characterization of the posterior (including error bounds), it also complicates the inferential process, since the quantities of interest are in the conformation space. Our posterior probabilities are integrals over the SCS, which we have parameterized as described in Sec. 3.2.1. If one does not use an appropriate volume element, then integrating over this parameterization is likely to introduce bias due to the non-uniform sampling density of the parameterization.

Intuitively, the problem is similar to defining a uniform distribution on a sphere. Simply taking uniform intervals in θ and ϕ of spherical coordinates does not work, since it overrepresents the poles. This overrepresentation of poles is an arbitrary bias introduced by the particular parameterization used (i.e., spherical coordinates). To remove the bias of the parameterization, we need to define a mathematical area element. Likewise, to integrate

over a sphere, a Jacobian is employed to account for the coordinate transformation. In the case of the sphere, the surface area is the invariant volume defining the uniform distribution.

Returning to symmetry axes and building upon this analogy, integrating with respect to SCS volume ($\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}\theta\,\mathrm{d}\phi$) would result in a different probability measure upon translation/rotation of the same portion of the space. To perform probabilistic inference, the probability density must be integrated with respect to a volume that is invariant to these Euclidean transformations. It has been shown that this invariant volume is well-defined and is completely determined (up to a constant factor) by requiring integrals of probability density to be invariant under change of coordinate frames. Such an invariant infinitesimal volume is defined in many classical texts on the subject of stochastic and geometrical probability (e.g., Moran and Kendall [39, p. 20], [49]). Applying that approach with SCS parameters gives an infinitesimal invariant volume $\mathrm{d}\mu$:

$$\mathrm{d}\mu = \mid \cos\theta \mid \sin\theta \; \mathrm{d}\theta \; \mathrm{d}\phi \; \mathrm{d}x \; \mathrm{d}y \tag{3.12}$$

where $\mathrm{d}\mu$ is a function of c which is specified by (θ, ϕ, x, y). Thus to integrate over the SCS, we do so with respect to $\mathrm{d}\mu$ instead of the four SCS parameters, thereby correctly distributing the probability density over the axes.

Posterior $p(c \mid R)$

Finally, to define the posterior probability, we divide Eq. 3.11 by normalization factor Z:

$$p(c \mid R) = \frac{1}{Z}\, p(c) \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \tag{3.13}$$

where

$$Z = \int_{\Omega} p(c) \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \mathrm{d}\mu \tag{3.14}$$

Ω denotes the SCS and, as discussed above, the integration is with respect to the invariant volume $\mathrm{d}\mu$ (Eq. 3.12).

Inference using posterior

Given the posterior (Eq. 3.13), we compute the mean atomic coordinates and their variances, integrating over the posterior density for the symmetry axis according to the transformation yielding rotated subunits (Eq. 3.1):

$$\mathrm{E}(\mathbf{q}' \mid R) = \int_{\Omega} \mathbf{q}' \; p(c \mid R) \; \mathrm{d}\mu \tag{3.15}$$

$$\mathrm{var}(\mathbf{q}' \mid R) = \int_{\Omega} (\mathbf{q}' - \mathrm{E}(\mathbf{q}'))^T (\mathbf{q}' - \mathrm{E}(\mathbf{q}')) \; p(c \mid R) \; \mathrm{d}\mu \; , \tag{3.16}$$

where $\mathbf{q}' = \mathcal{T}(c, \mathbf{q}, \alpha)$, Ω is again the SCS and $\mathrm{d}\mu$ the invariant volume. var is the covariance matrix since q' is a vector but we are only interested in the variance of the components of the vector with themselves and hence we only compute the diagonal elements of the matrix and we use the term 'variances' in Sec. 3.3 for the sum of the diagonal elements in this matrix.

3.2.3 Error bounds

The previous section gave a statistical framework for the posterior distribution over axes in the SCS (and thereby, complex structures), along with expected atomic coordinates and variances in them. The following section will develop a hierarchical subdivision algorithm to compute the distribution and integrate over it. This section establishes error guarantees that will be used by that algorithm, taking advantage of the structure of the configuration space to go beyond sampling-based methods in providing such guarantees. We leverage geometric bounds [44] to bound the individual factors of the posterior distribution in Eq. 3.13. This lets us compute upper and lower bounds to the unnormalized probability density in-

side a cell. Since the normalization factor Z in Eq. 3.14 is the sum of the unnormalized density over entire space, we obtain upper and lower bounds on Z from bounds on the unnormalized density. The upper and lower bounds on the posterior density, when used with non-negativity of the integrand, give upper and lower bounds for the total posterior probability integral within a cell. These bounds on cells are then used in conjunction with the triangle inequality to obtain bounds on the error in inferred mean atomic coordinates if these cells are eliminated.

SCS Cell Volume

The invariant volume for a cell is given by the integral of the infinitesimal invariant volume (Eq. 3.12). If $\cos\theta$ is positive in the range $[\theta_l, \theta_h]$, then we have:

$$\begin{aligned}\int_C \mathrm{d}\mu &= \int_C \mid \cos\theta \mid \sin\theta \; \mathrm{d}\theta \; \mathrm{d}\phi \; \mathrm{d}x \; \mathrm{d}y \\ &= [-\frac{1}{2}\cos^2\theta]_{\theta_l,\phi_l,x_l,y_l}^{\theta_h,\phi_h,x_h,y_h} \\ &= \frac{1}{2}(\cos^2\theta_l - \cos^2\theta_h)(\phi_h - \phi_l)(x_h - x_l)(y_h - y_l) \end{aligned} \tag{3.17}$$

If $\cos\theta$ is negative in the range $[\theta_l, \theta_h]$, then there is a negative sign in front of the integral in Eq. 3.17. If $\cos\theta$ changes signs in this range, we split the integral accordingly and evaluate each part.

Upper bound on the posterior within a cell

Let us first compute an upper bound on the value of $p(c \mid R)$ (Eq. 3.13) for an axis c in an SCS cell C (a contiguous set of axes; see again Fig. 3.2(b)). To do so, we compute upper bounds on the terms in the numerator and a lower bound on the normalization factor in the denominator. The normalization factor is the integral of the numerator and can be expressed (Eq. 3.19) as sum of the probability masses in SCS cells by breaking the integral. Thus, to compute the lower bound on the normalization factor, we also have to compute the

lower bound on the probability mass in each SCS cell, which is the term in the numerator.

$$\forall c \in C:\ p(c \mid R) \leq \frac{1}{Z_l} \max_{c \in C} p(c) \max_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \tag{3.18}$$

$$Z \geq Z_l = \sum_{C \in \Omega} \min_{c \in C} p(c) \min_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \int_C \mathrm{d}\mu \tag{3.19}$$

To compute these, we need both lower and upper bounds on $p(c)$ as well as the restraint likelihood sum $\left(\sum_{i=1}^{n} \log^2 g_i(c)\right)^{-n/2}$.

Recall that our structural prior $p(c)$ captures whether or not there is a steric clash. Thus the upper and lower bound on $p(c)$ within a cell C are set to 0 if the geometric bound $\mathcal{B}(C, \mathbf{q}, \alpha)$ (Eq. 3.2) for the position $\mathbf{q}'$ of at least one rotated atom falls within the van der Waals envelope of the fixed subunit, guaranteeing a steric clash. Likewise, both of the bounds are 1 if the bound on $\mathbf{q}'$ is outside the vdW envelope for all rotated atoms, so that no axis will cause any steric clash. If neither of these two cases hold, then the lower bound for $p(c)$ for $c \in C$ is 0 and the upper bound is 1.

The upper and lower bounds on the restraint likelihood sum can be written in terms of the lower and upper bounds respectively of the individual $\log$ terms.

$$\max_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \leq \left(\sum_{i=1}^{n} \min_{c \in C} \log^2 g_i(c) \right)^{-n/2} \tag{3.20}$$

$$\min_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \geq \left(\sum_{i=1}^{n} \max_{c \in C} \log^2 g_i(c) \right)^{-n/2} \tag{3.21}$$

Since $\log^2$ is a convex function with a global minimum at 1, $\log^2 g_i(c)$ increases on both sides of $g_i(c) = 1$. From the definition of g_i (Eq. 3.8), this happens when $\|\mathbf{p}_i - \mathcal{T}(c, \mathbf{q}_i, \alpha)\| = d_i$. Employing the lower bounds $l_i(C)$ and upper bounds $u_i(C)$ on $\|\mathbf{p}_i -$

$\mathcal{T}(c, \mathbf{q}_i, \alpha)\|$ over $c \in C$ and Eq. 3.3, we can write the bounds for the $\log$ terms as:

$$\min_{c \in C} \log^2 g_i(c) = \begin{cases} 0, & \text{if } \; l_i(C) \leq \|\mathbf{p}_i - \; \mathcal{T}(c, \mathbf{q}_i, \alpha)\| \leq u_i(C) \\ \min\left(\log^2 \frac{d_i}{l_i(C)}, \log^2 \frac{d_i}{u_i(C)}\right), & \text{otherwise} \end{cases} \tag{3.22}$$

$$\max_{c \in C} \log^2 g_i(c) = \max_{c \in C} \left(\log^2 \frac{d_i}{l_i(C)}, \log^2 \frac{d_i}{u_i(C)}\right) \tag{3.23}$$

Note that we have computed lower bounds on the individual probability terms in Eq. 3.18. This enables us to define the lower bound on the unnormalized probability density ρ that we use in Sec. 3.2.4. The lower bound can be written as:

$$\forall c \in C : \; \rho(c \mid R) \geq \min_{c \in C} p(c) \min_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c)\right)^{-n/2} \tag{3.24}$$

A sum of this bound over cells is the lower bound on the normalization factor.

Error bound on eliminated probability mass

We can derive an upper bound on the probability mass of an eliminated cell by using the upper bound on the posterior that was derived in Eq. 3.18. The upper bound on the posterior can be written as:

$$\begin{aligned} P(C \mid R) &= \int_C p(c \mid R) \, \mathrm{d}\mu \\ &\leq \left(\max_{c \in C} p(c \mid R)\right) \int_C \mathrm{d}\mu \end{aligned} \tag{3.25}$$

Error bounds on expected structure

When we omit a portion of the SCS in computing expected atomic coordinates, we introduce error into our characterization of the structure. We define the structural error as the average of the errors in the individual backbone atom positions. Thus to bound the error from omitting part of the SCS, we must compute the effect on the expected coordinates of

each atom (Eq. 3.15).

In the derivations that follow we represent the unnormalized conditional probability density by ρ that we introduced in Sec. 3.2.3. Thus:

$$\rho(c \mid R) = p(c) \left(\sum_{i=1}^{n} \log^2 g(c, \mathbf{q}_i, \alpha) \right)^{-n/2} \tag{3.26}$$

Suppose we leave out cell C in the computation of the expectation. We define the resulting error for a single atomic position $\mathbf{q}$ as:

$$\delta(C, \mathbf{q}) = \left\| \frac{\int_{\Omega} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu}{\int_{\Omega} \rho(c \mid R)\ \mathrm{d}\mu} - \frac{\int_{\Omega \setminus C} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu}{\int_{\Omega \setminus C} \rho(c \mid R)\ \mathrm{d}\mu} \right\| \tag{3.27}$$

We can write the integrals in the first term of Eq. 3.27 as sums of integrals over C and the rest of the SCS. Through simple algebra, we can cancel a few terms. Then by applying the triangle inequality and using non-negativity of the integrand, we can derive the following inequality:

$$\delta(C, \mathbf{q}) \leq \frac{\|\mathrm{E}(\ \mathcal{T}(c, \mathbf{q}, \alpha) \mid R)\| \max_{c \in C} \rho(c \mid R) \int_{C} \mathrm{d}\mu\ + \max_{c \in C} \|\ \mathcal{T}(c, \mathbf{q}, \alpha)\| \max_{c \in C}\ \rho(c \mid R)\ \int_{C} \mathrm{d}\mu}{\int_{\Omega \setminus C}\ \rho(c \mid R)\ \mathrm{d}\mu} \tag{3.28}$$

The algorithm we present in the next section will compute the denominator. The geometric bounds (Eq. 3.2) give the maximum atomic coordinates for $\mathbf{q}' = \mathcal{T}(c, \mathbf{q}, \alpha)$, and we have already derived bounds for all the probabilistic terms except the expectation $\|\mathrm{E}(\ \mathcal{T}(c, \mathbf{q}, \alpha) \mid R)\|$, which we can write as:

$$\|\mathrm{E}(\ \mathcal{T}(c, \mathbf{q}, \alpha) \mid R)\| = \frac{\|\int_{\Omega}\ \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu\|}{Z} \tag{3.29}$$

In this equation, we already have a lower bound on Z. The integral over Ω can be broken into the integral over C and that over $\Omega \setminus C$. Applying the triangle inequality on this

sum, along with the inequality on the norm of an integral for a non-negative integrand, we can derive:

$$\begin{aligned}\left\| \int_{\Omega} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\| \leq \max_{c \in C} \| \mathcal{T}(c, \mathbf{q}, \alpha) \| \max_{c \in C} \rho(c \mid R) \int_{C} \mathrm{d}\mu \\ + \left\| \int_{\Omega \backslash C} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\|\end{aligned} \tag{3.30}$$

Our algorithm will provide the integral over $\Omega \backslash C$, and we have already discussed bounds for the other terms. Combining these bounds and equations, and substituting into Eq. 3.28 gives us the final inequality for the error in expectation:

$$\begin{aligned}\delta(C, \mathbf{q}) \leq \frac{1}{\int_{\Omega \backslash C}\ \rho(c \mid R)\ \mathrm{d}\mu} \Bigg(\max_{c \in C} \| \mathcal{T}(c, \mathbf{q}, \alpha) \| \max_{c \in C}\ \rho(c \mid R)\ \int_{C} \mathrm{d}\mu + \\ \frac{\max_{c \in C} \rho(c \mid R) \int_{C}\ \mathrm{d}\mu}{Z_l}\ \cdot \Big(\max_{c \in C} \| \mathcal{T}(c, \mathbf{q}, \alpha) \| \max_{c \in C} \rho(c \mid R) \int_{C} \mathrm{d}\mu + \\ \left\| \int_{\Omega \backslash C} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\| \Big) \Bigg)\end{aligned} \tag{3.31}$$

From the bounds for a cell C we can derive bounds for a set $\mathcal{C}$ of cells, replacing the integral over C with a sum of integrals over $C \in \mathcal{C}$:

$$\begin{aligned}\left\| \int_{\mathcal{C}} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\| &= \left\| \sum_{C \in \mathcal{C}} \int_{C} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\| \\ &\leq \sum_{C \in \mathcal{C}} \left\| \int_{C} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\|\end{aligned} \tag{3.32}$$

The upper bound in Eq. 3.28 can be rewritten in terms of the individual cells by using Eq. 3.32.

3.2.4 Hierarchical subdivision algorithm

To compute the posterior distribution, along with expectations and variances in atomic coordinates, we develop a hierarchical subdivision algorithm. The algorithm is illustrated in

Algorithm 1 Hierarchical subdivision algorithm

Input: $\mathcal{C}_0$: initial set of cells from feasible region of $\mathbb{S}^2 \times \mathbb{R}^2$
Input: R: set of distance restraints
Input: ζ_0: maximum pruned probability mass
Output: P: posterior distribution, a set of (cell, posterior) pairs

```
P ← ∅
C ← C_0    // cells for the next level
ζ ← ζ_0    // remaining allowed error
Z_l ← lower bound on Z for C_0    // Eq. 3.19
while C is not empty do    // expand the next level
  V ← Σ_{C∈C} ∫_C dμ    // invariant volume, Eq. 3.17
  C' ← ∅    // cells for the next level
  for C ∈ C do
    u ← upper bound on P(C | R) using current Z_l    // Eq. 3.25,Eq. 3.18
    if u < (ζ/V) ∫_C dμ then    // prune cell
      ζ ← ζ − u
    else if C is small enough then    // accept cell
      p ← ρ(c | R) ∫_C dμ for the centroid c of C    // unnormalized Eq. 3.13
      add to P the pair (C, p)
    else
      subdivide C into C_1 and C_2
      C' ← C' ∪ {C_1, C_2}
      // Update Z_l for subdivision, using ρ_min from Eq. 3.24
      Z^l ← Z^l − ρ_min(C | R) ∫_C dμ + ρ_min(C_1 | R) ∫_{C_1} dμ + ρ_min(C_2 | R) ∫_{C_2} dμ
    end if
  end for
end while
```

Fig. 3.3, and pseudocode is provided in Alg. 1. While we also used hierarchical subdivision in our earlier approach [44], the algorithm here is structured so as to support structural inference with error guarantees.

We start with a set $\mathcal{C}_0$ of cells covering the region of interest in the SCS. While the entire SCS is the Cartesian product of the state space of the four random variables: $x \in [-\infty, \infty]$, $y \in [-\infty, \infty]$, $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi]$, we can truncate the probability density to zero beyond a finite range of x and y values [44]. We select such x and y boundaries of the finite range so that a homo-oligomer that has a symmetry axis with x, y beyond this xy patch would be biophysically unfeasible for most θ, ϕ. This results from our choice for the z axis as the principal axis of the fixed subunit, along with the fact that protein complexes

are packed together, rather than floating loosely in space. If we encounter homo-oligomers that have axes that are nearly parallel to the xy-plane, and hence have x, y outside our finite range, then we can change our translation parameters to either $\{y, z\}$ or $\{x, z\}$ by considering each and choosing the one that does not have this problem.

The algorithm proceeds level-by-level through a hierarchical subdivision of the input cells. At a given level, each cell is considered independently of the rest. There are three possibilities for a cell under consideration: it can be safely pruned according to our error bounds, it is small enough to be considered a leaf (it is "accepted"), or it is partitioned into two smaller cells for the next level. The process continues until reaching a level at which no cell needs to be subdivided.

We prune cells when our error bounds allow us to determine that ignoring them will have a "small enough" effect on the results. To make this determination, we maintain two global quantities. One quantity is Z_l, the lower bound on the normalization constant, by which we evaluate the relative amount of posterior mass in a cell vs. other cells (used in upper-bounding the cell's contribution). We start with the value for the initial cells, from Eq. 3.19, and each time we split a cell, we subtract out the parent cell's contribution and add in the children's contributions to Z_l. The other quantity, ζ, is the remaining amount of probability mass we can still prune. We start with a user-specified maximum value ζ_0, and each time we prune a cell, we reduce the remaining prunable mass by the upper bound on the probability mass in that cell. Given the current values of these quantities and the bound on the probability mass contribution of a cell (Eq. 3.25), we can safely prune that cell if its contribution is no more than ζ multiplied by the fraction of the total invariant volume (Eq. 3.17) that it occupies.

We consider a cell to be *accepted* (a leaf node) when the structures it represents are very similar. We employ our previous approach of evaluating this by computing average backbone RMSD among the structures represented by the corners of the cell, and terminating when that average is within a threshold $\mathtt{t}0$ (e.g., 1 Å) [44].

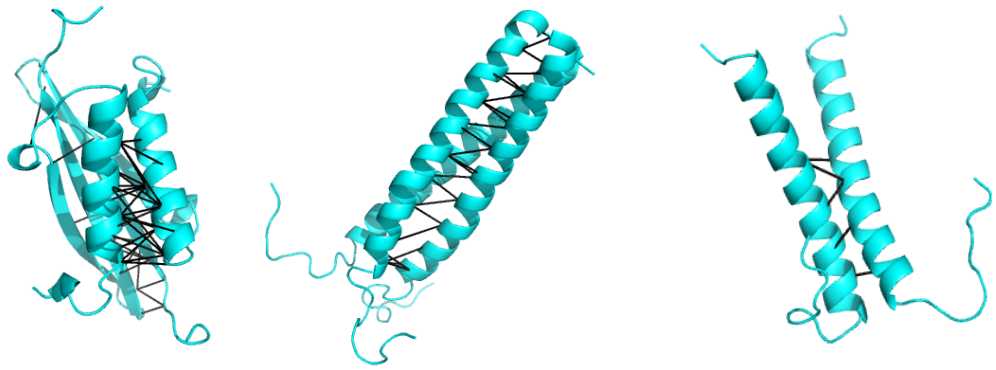
To subdivide a cell, we split one of the dimensions in half, employing the heuristic from our earlier method [44] ("Branching" in Methods). Intuitively, the goal is for restraint violations to be concentrated in one of the children, resulting in a low (potentially prunable) posterior.

Our pruning focuses on ensuring that we have sufficient probability mass represented in the posterior. In addition, we also want to ensure that we limit the error in expected atomic coordinates. We check this after the search is complete. We compute the error in expectation due to the pruned cells (Sec. 3.2.3). If this error is guaranteed to be less than a user-specified threshold ε on the allowed error, the algorithm is finished. Otherwise we must run it with a tighter ζ so that we eliminate less probability mass. In practice, we have not needed to do that; the ζ restriction is strong enough to ensure small enough error in expected atomic coordinates.

The breadth-first structure of this algorithm allows us to implement the algorithm in parallel on a cluster. To fully use the capacity of the compute cluster and to start with tighter bounds, we initialize $\mathcal{C}_0$ to be a uniformly sampled grid of 2^{17} cells. Our implementation uses Apache Hadoop (`http://hadoop.apache.org`), an open source implementation of Map/Reduce [11], which provides a framework for parallelizing the code, taking care of machine failure, scheduling jobs, and partitioning the data.

3.3 Results

We tested our approach on three protein complexes for which intra-subunit and inter-subunit NOEs had been separated and subunit structures determined from the intra-subunit NOEs. The homo-dimeric topological specificity domain of *E. coli* MinE [28] has 50 residues per subunit with 183 inter-subunit NOE restraints, the homo-trimeric coiled-coil domain of chicken cartilage matrix protein (CCMP) [68] has 47 residues per subunit and 49 inter-subunit NOE restraints and a transmembrane peptide of Glycophorin A (GpA trans-



MinE CCMP GpA transmembrane peptide

Fig. 3.4: Reference structures (cyan) and inter-subunit distance restraints (black) for MinE (183 restraints), CCMP (49) and GpA transmembrane peptide (6). For CCMP, restraints are only shown between chains A and B, to avoid clutter.

membrane peptide) [35] has 40 residues per subunit with 6 inter-subunit NOE restraints. We obtained *reference ensembles* (20 members each) of structures deposited in the protein databank (PDB) [4]—MinE: pdb id 1EV0; CCMP: pdb id 1AQ5; GpA transmembrane peptide: pdb id 1AFO. We took as the *reference structures* the member of each ensemble identified by the authors to be the best representative, and used for the subunit structure the first chain of the reference structure. We obtained the inter-subunit NOEs and assigned chemical shifts from the BioMagResBank (BMRB) [51]. The restraints are fairly well-dispersed in the structures (Fig. 3.4), except for the GpA transmembrane peptide, which has only six restraints, all between the lower halves of its two helices.

We set our expectation error threshold ε to 0.3 Å, maximum pruned probability mass threshold ζ to 0.1, and our acceptable cell threshold τ to 1 Å. The hierarchical decomposition algorithm took 10–36 hours on a 30 node cluster, with the slowest time for CCMP when using only 16 of the original 49 restraints.

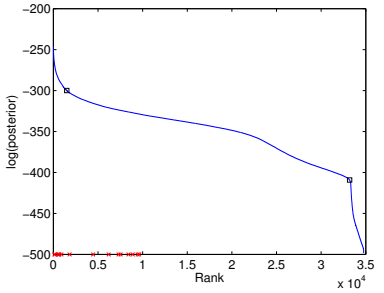
3.3.1 Posterior

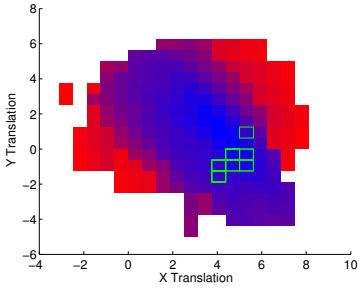
The hierarchical decomposition for MinE produced a set of 35,000 accepted cells, with a total volume of 1.81 Å^2-rad^2 out of the original 1257 Å^2-rad^2. Note that these and all

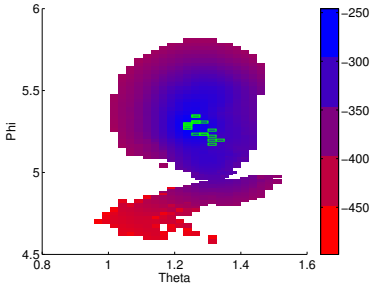
subsequent SCS volumes are with respect to our invariant volume $\mathrm{d}\mu$, and thus independent of the coordinate frame. The top row of Fig. 3.5(top row) plots both the log posterior probabilities of these cells (top left), in decreasing order and the translation and orientation components of the accepted cells, colored by log posterior (top middle/right). We simply show the "raw" unnormalized log posteriors, though our bound on the normalization constant in fact permits us to normalize them to within an error bound. The maximum *a posteriori* (MAP) cell has an unnormalized log posterior of -246. The probabilities drop steeply after the MAP up to the 1500th cell (first black square on the plot), which has a posterior of -300 and a backbone RMSD to the MAP of 0.8 Å. In general, these first cells span a small portion of the configuration space (0.1 Å^2-rad^2) and represent similar structures (0.9 Å average backbone RMSD from the MAP, over ten samples drawn from this region). After that there is a steady decrease in the posterior for the next 32,000 cells (between the two black squares in the plot), when we reach a posterior of -410 before another sharp drop-off leading down to cells that were pruned. Compared to the highest-posterior cells, the middle-range ones (posteriors between -300 and -410) are spread out in the SCS (1.5 Å^2-rad^2) and have greater structural diversity (2.2 Å average backbone RMSD from the MAP over ten samples in the region).

The posterior has a fairly sharp peak, and the high-posterior axes are aggregated in terms of translation and orientation. We compared these results against the structures in the reference ensemble. The MAP structure is very similar to the reference ensemble (Fig. 3.6(left)), with a backbone RMSD of 0.5 Å from the closest member of the ensemble. The members of the reference ensemble are highlighted in Fig. 3.5: marks on the x-axis of the posterior distribution and outlines for containing cells in the translation/orientation plots. All the reference axes are also found by our inference algorithm. The reference axes have fairly high posteriors, though clearly there are numerous solutions determined by the inference algorithm that are similar or better. Of course, the actual posterior value depends on the scoring system; the point is that a 20-member ensemble greatly underestimates the

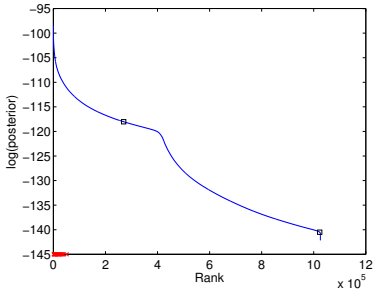
MinE

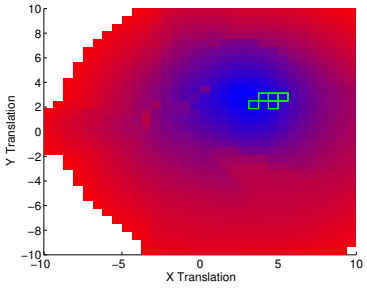

log(posterior)
Rank
x 10^4

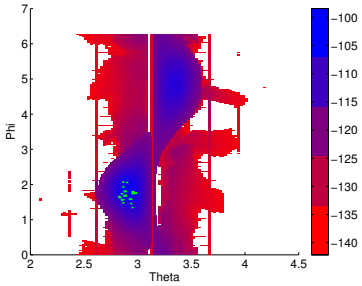

Y Translation
X Translation


Phi
Theta

CCMP


log(posterior)
Rank
x 10^5


Y Translation
X Translation


Phi
Theta

GpA transmembrane peptide

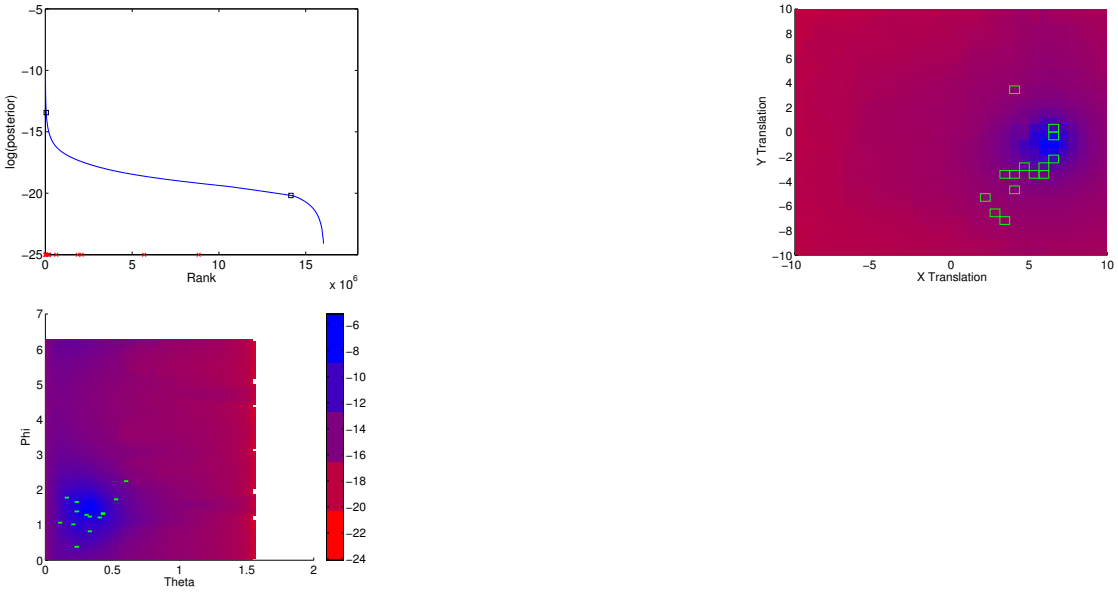
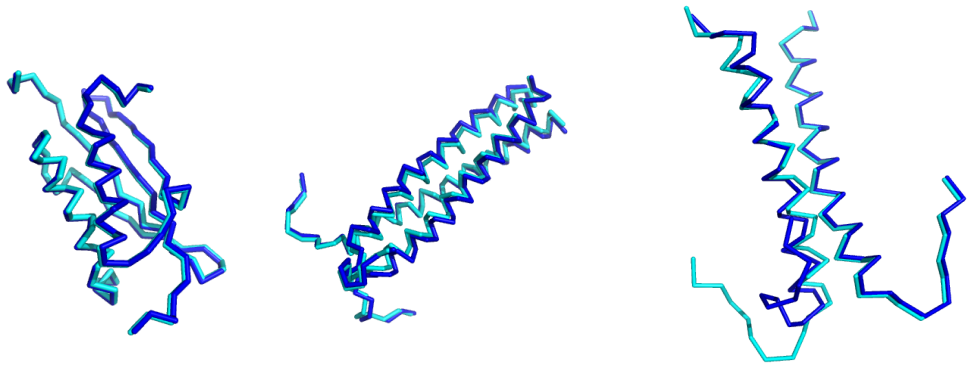


Fig. 3.5: Posterior distributions. (left) Unnormalized log posterior for accepted cells. Red points on the x-axis indicate posteriors computed for members of the reference ensembles. (right) Projections of SCS onto translation and orientation components, colored by posterior (different scales for different proteins). Cells containing reference structures are outlined in green. Since many cells can share their translations or orientations with other cells, the color of a translation or orientation is shown colored according to the highest posterior cell in that region. For CCMP, π was added to θ for display purposes (to bring together equivalent cells).



MinE CCMP GpA transmembrane peptide

Fig. 3.6: MAP structures (cyan) superimposed with closest member of reference ensembles (blue).

generally acceptable variation in conformations (as represented by axes).

We also compared our results against those obtained from our earlier "binary" approach [44], which checks only whether or not each restraint is satisfied. Again, the new method identifies all axes found by the earlier algorithm, along with many more. The binary approach is sensitive to restraint violation, and does not adequately represent the space when allowing for that. For example, in MinE the cell centered at $(2.19, 1.56, 1.29, 5.29)$ is rejected by the binary approach since 22 restraints out of 183 are violated. However, this cell is kept by the inference approach since its posterior is still sufficiently high, as 18 of the violations are all less than 1 Å and the other 4 are less than 1.5 Å. In fact, the cell containing this axis has a log posterior of -311.7 and is in the top ten percent of the accepted cells according to its posterior.

For CCMP, our algorithm accepted $\approx 10^6$ cells, with a total volume of 43.8 Å^2-rad^2 out of 2513 Å^2-rad^2. Fig. 3.5(middle row) shows the posterior and the translation/orientation components for the cells. The log posterior decreases fairly smoothly from the MAP (-98.4) for 2.7×10^5 cells, to an inflection point (first black square on the plot) at a posterior of -118, and then again for another 7.5×10^5 cells before dropping sharply (second black square) for the final 2.7×10^3 accepted cells. Unlike with MinE, the high-posterior cells are fairly dispersed in the SCS and in conformation space. The volume occupied by

cells from the MAP to the inflection point is 4.4 Å^2-rad^2 with an average backbone RMSD to the MAP of 4.9 Å over ten random samples drawn from this region, while the cells after the inflection point comprise the majority of the volume (39.3 Å^2-rad^2) with an average backbone RMSD of 8.9 Å over ten random samples in the region.

In comparison to the closest member in the reference ensemble, the MAP structure has a backbone RMSD of 1.5 Å (Fig. 3.6 (middle)). The maximum backbone RMSD is between the backbone C$^\alpha$s at the base of the helices of the two structures. As with MinE, our method identifies with a high posterior all structures in the reference ensemble (highlighted in the figure). The orientation components of high posterior cells in CCMP are grouped into two clusters. These two groups contain axes that are similar but point in opposite directions. Hence, the structures with highest probability in both groups are very close, but chain B superimposes on chain C of the other and vice versa to yield a backbone backbone RMSD of 2.5 Å. Like in earlier cases, our method also finds those axes identified by the binary algorithm.

For GpA transmembrane peptide, the algorithm accepted $\approx 10^7$ cells which had a total volume of 954.9 Å^2-rad^2 out of 1257 Å^2-rad^2. This was the least pruning of all the three proteins and it can be attributed to its having only six inter-subunit NOE restraints. Fig. 3.5 (bottom row) plots the posterior, again with the reference ensemble highlighted. The form of the posterior curve is very similar to what we saw for the other proteins: a small set of cells with a high posterior (from -5.7 for the MAP), followed by a significant drop in the posterior (down to -13.4 at the first black square after 5.0×10^4 cells, and a smooth degradation (-20.2 at the second black square after 1.4×10^7 cells). The volume occupied by the cells from the MAP to the first square is 3.4 Å^2-rad^2 while the cells between the first and second black square constituted the majority of the volume (885.5 Å^2-rad^2). The cells constituting the drop off after the second cell occupy a volume of 66.0 Å^2-rad^2. The ten random samples drawn from the volume occupied by cells from the MAP to the first square have an average backbone RMSD of 2.4 Å from the closest member of the reference

ensemble. The rest of the volume is occupied by cells with high backbone RMSDs (average of 9.8 Å in ten random samples from the region). While the translation and orientation projections in Fig. 3.5(bottom row) display a trend like those for the other proteins, the small number of restraints leads to a relatively small amount of pruning and a large number of low posterior cells.

3.3.2 Inferred means and variances

The means and variances obtained by the inferential approach are directly reflective of the ensemble that fits the data. This is in contrast to the means and variances that one may compute from the top twenty structures obtained from SA/MD methods which are only within the discrete set of top structures. Note that a "centroid" of an ensemble is different from the actual mean, in that the mean allows for differentially weighted contributions. Furthermore, the mean must be with respect to the entire space and not just a selected set; it is "unbiased" in that sense. In our method, the mean is computed to within a bound on the possible error from the "true" mean structure.

Fig. 3.7 shows "sausage plot" representations of the means and standard deviations (square roots of computed variances) of atomic coordinates inferred by our method. For MinE, the mean structure has backbone RMSD 0.45 Å from the reference structure and 0.03 Å from our MAP estimate. Since there are 183 restraints, the structure is quite constrained, and standard deviations range only from 0 to 0.37 Å along the backbone, with an average of 0.11 Å. For CCMP, the mean structure has a backbone RMSD of 1.7 Å from the closest member of the reference ensemble and 0.5 Å from the MAP. The standard deviations of backbone C^{α} atoms range from 0 to 4.62 with a mean of 1.44 Å. The "loosest" parts are at the tips of the helices. While there is a restraint that reaches there (Fig. 3.4(middle)), the structural uncertainty results from an interplay among all the restraints, and there is apparently not sufficient reinforcement to fully pin down the structure

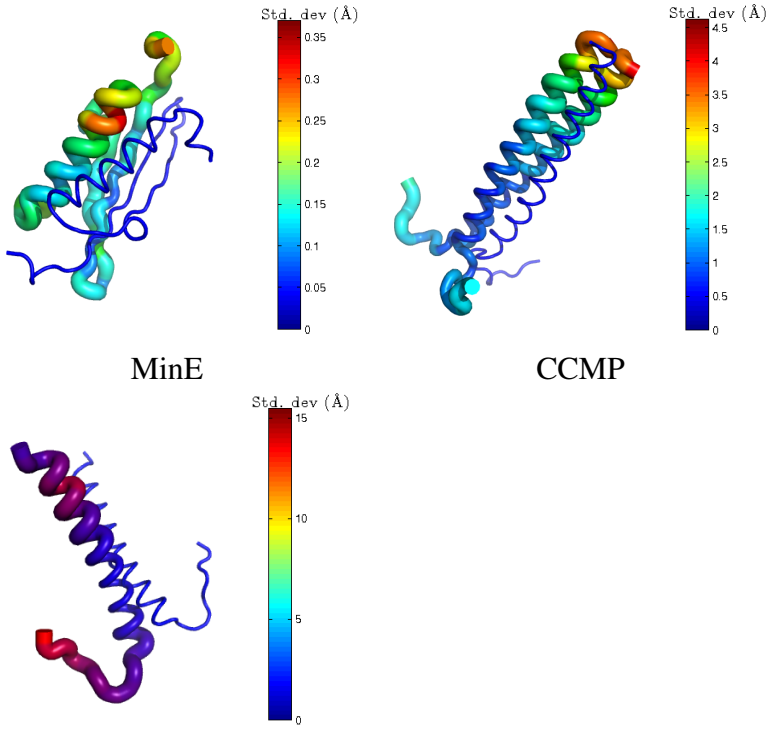


Fig. 3.7: Inferred means and standard deviations in atomic coordinates, represented as sausage plots. The fixed subunit is shown in blue with a zero standard deviation. The color and thickness of the adjacent subunit represent the standard deviation in the positions of the backbone atoms. Note that the standard deviations are on different scales for different proteins.

there. Finally, Fig. 3.7(right) shows the sausage plot for GpA transmembrane peptide. The standard deviations of backbone C^{α} atoms range from 0 to 15.45 Å with a mean of 4.33 Å. The MAP structure for GpA transmembrane peptide has a backbone RMSD of 0.83 Å with the closest member of the reference ensemble (Fig. 3.6(right)). The backbone RMSD of the computed mean structure with the closest member of the reference ensemble is 1.7 Å. The lower half of the helices in GpA transmembrane peptide are more tightly restrained through the six NOE restraints shown in Fig. 3.4(right). Therefore this part of the helix in the second subunit shows the least variance.

3.3.3 Robustness to missing restraints

We studied the robustness of our method to missing data. For MinE, we selected restraint subsets of sizes 91, 49, and 35 from the original 183 ($\approx$ 50, 25 and 20%), randomly choosing the restraints from the entire set. We generated 5 such datasets for each number of restraints. Similarly, for CCMP we generated random subsets of sizes 27 and 16 from the original 49 ($\approx$ 50 and 30%). We did not perform this test for GpA transmembrane peptide since it already has only 6 restraints. For each subsampled dataset, we first evaluated the volume of the pruned portion of the SCS, to see how many more conformations would be consistent with the reduced restraint set. We then compared the mean and MAP structures for the reduced set with those for the original, to evaluate the effects on these representative structures. Finally, we compared the variances in the atomic coordinates, to assess the increase in structural uncertainty.

Tab. 3.1 summarizes the trends over the different restraint sets. For MinE, even with only 35 of the 183 restraints, almost 99% of the volume is still pruned, suggesting that the posterior distribution is close to zero for most of the SCS. The various backbone RMSDs are also relatively small, as sufficient constraint remains to yield structures much like those with the full set of restraints. Since, the reference ensembles contains the structures that have the highest likelihood of occurrence, the backbone RMSDs of these structures to the MAP are in general smaller than those to the mean. For CCMP, the amount of pruning falls off more sharply, and the backbone RMSD values increase more. This is largely due to the fact that the absolute number of restraints is much smaller. To compute the expectation within the error tolerance we must include a larger number of cells.

With fewer restraints, more cells contribute a significant probability mass. Fig. 3.8 illustrates the expansion in accepted SCS with fewer cells; a similar trend is observed for CCMP. The volume of 1.81 Å^2-rad^2 with 183 restraints expands to 2.44 Å^2-rad^2 with 91, 4.33 Å^2-rad^2 with 48, and 9.23 Å^2-rad^2 with 35 (means taken across 10 datasets). Due to algorithmic pruning choices, some (low posterior) cells accepted with more restraints may

Protein	Restraints	Pruned%	$RMSD_1$	$RMSD_2$	$RMSD_3$	$RMSD_4$
MinE	183	99.8	0.0	0.5	0.0	0.5
	91	99.8±0.01	0.2±0.16	0.5±0.03	0.2±0.10	0.5±0.09
	49	99.6±0.08	0.4±0.09	0.4±0.09	0.5±0.20	0.5±0.28
	35	99.3±0.18	0.8±0.42	0.9±0.50	0.7±0.38	0.9±0.47
CCMP	49	97.8	0.00	1.5	0.0	1.8
	27	91.4±1.32	0.3±0.24	1.4±0.20	1.3±0.90	2.7±0.86
	16	68.6±4.42	0.6±0.28	1.6±0.36	1.3±0.51	2.6±0.64

Tab. 3.1: Effects of missing restraints on inference. Pruned%: percentage of SCS volume pruned; $RMSD_1$: reduced-restraint MAP vs. full-restaint MAP; $RMSD_2$: reduced-restraint MAP vs. reference ensemble; $RMSD_3$: reduced-restraint mean vs. full-restraint mean; $RMSD_4$: reduced-restraint mean vs. reference ensemble. All RMSDs are computed with backbone atoms. The RMSD to closest structure in reference ensemble is shown.

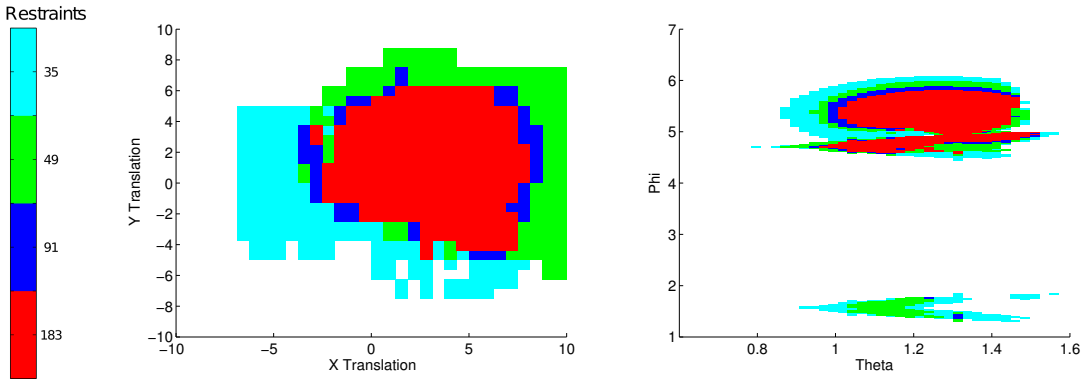
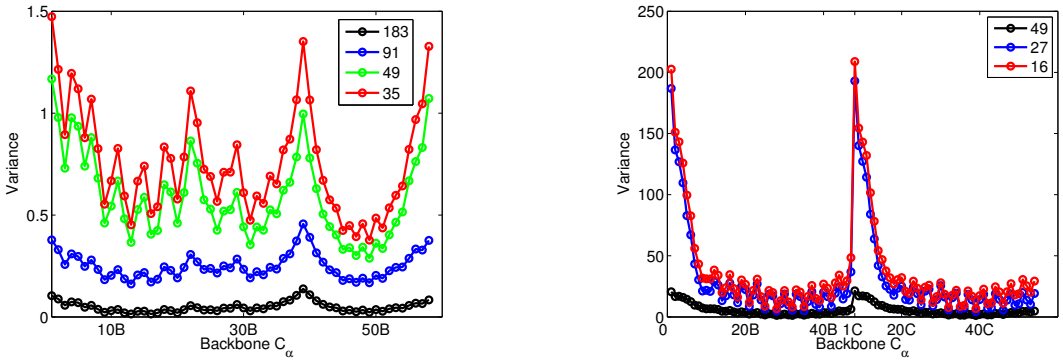


Fig. 3.8: Translation and orientation parameters of accepted MinE cells for different subsets of restraints (one dataset for each number of restraints). Colored cells are those eliminated with more restraints but not with fewer restraints.

actually be rejected with fewer restraints, though we found very few cells with a volume less than 0.25 $Å^2$-rad^2 to have this opposite trend. Fig. 3.9 plots the mean atomic variances for the C^α atoms, under the different random sets of restraints. While creating the random sets of restraints, we did not ensure that the sets with smaller number of restraints are subsets of those with larger number of restraints (except for the full restraint set). However, the trends in the plots in Fig. 3.9 show that the atoms with large variances essentially remain the same across different sets of restraints. By taking out restraints, many low posterior axes no longer have a negligible posterior and therefore the variance increases. The highest variance in CCMP is at the tips of the helices, as shown in red in its sausage plot (Fig. 3.7



MinE CCMP

Fig. 3.9: Mean variance (y-axis, Å^2) in C$^\alpha$ atom positions (x-axis) for datasets with different numbers of restraints (different lines). Each atom is identified by its residue number and a letter denoting the chain.

(middle)).

These results suggest that our approach degrades smoothly with data sparsity, appropriately representing and evaluating the increasing uncertainty in the resulting conformations.

3.3.4 Robustness to noise

We evaluated the robustness of the inference approach to experimental noise, including both uncertainty in the distance (exceeding the specified bounds) and spurious restraints. We call both scenarios "noisy" restraints, recognizing that while experimental restraints already include some padding to allow for uncertainty in distance estimation, algorithms must also be able to handle violations. We simulated noise in a manner that reflects realistic systematic structural variation and uncertainty, instead of simply adding random noise. In addition to the representative structure whose first chain we used as our input subunit, the deposited NMR ensemble contains a number of other structures. We simulated restraints (identifying pairs of protons within 6 Å) from another member of the ensemble, and identified those that were violated in the representative structure. With respect to the reference structure, some of these "noisy" restraints have small violations and some are significantly violated. For the MinE dimer, model 9 was the most different (3.7 Å backbone RMSD)

from the reference structure. It yielded 24 noisy restraints, 16 violated by more than 1 Å and two by as much as 19 Å. For the CCMP trimer, model 4 was the most different (6.6 Å), and yielded 8 noisy restraints (1 violated by more than 30 Å). We formed sets of augmented restraints by combining our experimental restraints with these noisy restraints.

We re-ran our inference algorithm with these noisy datasets. For MinE, it accepted 128,000 cells covering 4.4 Å^2-rad^2, compared to 1.8 Å^2-rad^2 without the noise. Even with the noisy data, the accepted cells still include those representing the reference ensemble and 40% of the original cells. This then yields increased uncertainty in conformation space; the MAP has a backbone RMSD of 2.3 Å from the original and 2.5 Å from the closest structure in reference ensemble, and a mean 2.2 Å from the original and 2.4 Å from the closest structure in the reference ensemble.

For CCMP, our algorithm accepted 7×10^5 cells covering 24.2 Å^2-rad^2, compared to 43.8 Å^2-rad^2 in the original. These solutions include the reference ensemble and 60% of the original cells. The MAP remain essentially the same, with an backbone RMSD of 0.0 Å from the original and 1.5 from the reference, and similarly the mean has a backbone RMSD of only 0.3 Å from the original and 2.0 Å from the reference.

All noisy restraints for MinE are concentrated in the upper and lower loops of the dimer where there are no existing non-noisy restraints. Therefore, the addition of noisy restraints results in higher posteriors for the axes representing structures with the noisy restraints satisfied in those loop regions. On the other hand, for CCMP the added noisy restraints are all in the middle of the helices where there are non-noisy restraints. The structures in which the noisy restraints are satisfied tend to violate these non-noisy restraints, which outnumber them. Hence the noisy restraints do not impact the eventual posterior distribution in CCMP to the extent observed for MinE.

Our original binary algorithm [44] would fail with this set of noisy restraints since they are inconsistent and the algorithm eliminates a cell if even one NOE is violated. Therefore, we had extended that approach, in the context of NOE assignment, to handle a fixed

maximum number of violations (denoted by δ) [46]. We tested the extended approach on our current datasets. We found that for the augmented MinE dataset, no solutions were obtained when δ was set at less than 15. As we increased δ from 15 to 20, the average backbone RMSD to the reference structure increased from 0.59 Å to 0.75 Å and the non-overlapping volume increased from 0 Å^2-rad^2 to 0.021 Å^2-rad^2 (compared to 0.70 Å and 0.0004 Å^2-rad^2 by the inference approach). With CCMP, we needed $\delta \geq 4$ to find any solutions; as δ increased from 4 to 9, the average backbone RMSD increased 0.92 Å to 0.99 Å and the non-overlapping volume increased from 0.0001 Å^2-rad^2 to 0.0081 Å^2-rad^2 (compared to 0.98 Å and 0.0001 Å^2-rad^2 by inference).

Our inference approach is robust to noise: there is no need for a maximum number of restraint violations; it degrades smoothly. It also appropriately accounts for the influence of noisy data on the resulting structures, via the weighted integration.

3.4 Conclusion

We have developed an approach that performs structural inference for symmetric homo-oligomers. By working with a configuration space representation and employing a hierarchical subdivision algorithm, our approach gives error guarantees on the resulting posterior and inferred expectations in atomic coordinates. The method provides a probability measure for sets of conformations, allowing for an objective assessment of the information content in the data and the resulting constraint on the plausible structures. It can then evaluate the resulting uncertainty in atomic coordinates.

In our case study applications, we have found that in addition to all the structures found by previous methods, our method also identifies other diverse structures with high posterior probabilities. That is, our probabilistic restraint evaluation and complete characterization of the posterior distribution enables identification of structures that are missed when employing either binary restraint violation testing or stochastic sampling of low-energy conforma-

tions. In particular, the set of twenty reference structures deposited in the PDB suffers from the problem of under-sampling the conformation space. Furthermore, the inferred atomic means provide a more accurate characterization of the structural uncertainty than a simple superposition of an ensemble of low-energy representatives.

As NOESY experiments are subject to noisy and missing data, the input set of distance restraints may include some distance restraints that are violated to a small extent (even after padding) or completely spurious, and may not include some correct distance restraints. Our approach takes into account such sources of uncertainty and degrades smoothly. With simulated missing data, most of the originally accepted cells were still accepted, and consequently the MAP and mean structures were not very different from those obtained with the full set of restraints. We simulated noisy restraints, we found similar robustness, with results similar to those obtained from the original set of restraints.

Our approach currently evaluates structural quality only in terms of steric clash, rather than in terms of finer-grained molecular mechanics modeling. The posterior is driven by restraint satisfaction, and the prior only prunes structures that display serious steric clashes. This leads to a "data-driven" search for and evaluation of structures, with conclusions regarding structural uncertainty based mainly on the experimental data. We were able to use a binary structural prior since, as in our previous work [44], we assumed that the subunit structure was fixed (solved as it exists in complex, from the intra-subunit subset of the NMR data). While we previously performed energy minimization on the side-chains as a post-processing step, that is not as appropriate here, since that would affect the probabilities and error cutoffs, and thus our inference moments would no longer be provably accurate. The posteriors obtained here essentially "flatten out" the possible side-chain conformations for a backbone, and the distribution that we compute should therefore be interpreted as the posterior over backbones rather over complete homo-oligomeric structures including side-chains.

In future work, we would like to better account for biophysical plausibility by incorpo-

rating a Boltzmann prior representing molecular modeling energies. The key challenge is to efficiently and tightly bound such a prior over an SCS cell. This is analogous to the move from energy minimization after pruning rotamers with Dead-End Elimination (DEE) [13], which loses the global minimum energy guarantee of DEE, to minimized-DEE [17], which accounts for possible energy minimization when considering pruning and thus regains the provable guarantee.

Acknowledgement

This work was supported by the following grant from the National Institutes of Health: R01 GM-65982 to B.R.D. We would like to thank Tim Tregubov for significant help with the compute cluster in general and with Hadoop in particular.

4. SIMULTANEOUS DETERMINATION OF SUBUNIT AND COMPLEX STRUCTURES OF SYMMETRIC HOMO-OLIGOMERS FROM AMBIGUOUS NMR DATA

Abstract

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

demonstrate that it can handle the difficult configuration space search problem and substantial ambiguity, effectively pruning the configuration spaces and characterizing the actual diversity of structures supported by the data.

4.1 Introduction

Symmetric homo-oligomers are comprised of subunits that are identical in sequence and highly similar in structure and are arranged symmetrically; we study here cyclic symmetry, in which the subunits are placed like spokes on a wheel (Fig. 4.1). Symmetric homo-oligomers are thought to make up a majority of proteins; they play important roles in biological processes that include ion transport and regulation, signal transduction, and transcriptional regulation [18]. They are therefore a valuable target for structural studies, and nuclear magnetic resonance spectroscopy (NMR) provides the ability to analyze their structures and dynamics in solution.

Nuclear Overhauser Enhancement Spectroscopy (NOESY), which measures through-space interactions, provides the main source of structural information in standard NMR protocols. Nuclear Overhauser Effect (NOE) intensities are converted into distance restraints and assigned to pairs of protons, giving upper bounds on their interatomic distances. NOE distance restraints are typically used to frame NMR structure determination as an optimization problem combining biophysical modeling terms with pseudo-energy encodings of the restraints [6, 20, 21]. Statistical inference techniques have also been developed to combine modeling and experimental terms and characterize the resulting posterior distributions of structures [8, 48].

In a symmetric homo-oligomer, the high structural similarity of the subunits yields highly similar chemical environments for their atoms, rendering an atom in one subunit indistinguishable from the corresponding atom in another subunit under standard NMR experiments. Consequently, an NOE may involve two atoms in the same subunit or in two

Fig. 4.1: Symmetric homo-oligomer (*B. subtilis* anti-TRAP trimer, pdb id 2ko8) with ambiguous interpretation of a distance restraint: intra-subunit (red), "clockwise" (yellow), and "counter-clockwise" (black).

different subunits (with multiple pairs of subunits possible for symmetry higher than two) or both, with no inherent means to resolve this ambiguity (Fig. 4.1). Some experimental strategies have been devised to separate intra- vs. inter-subunit restraints [24, 32, 63, 71], but they are difficult and have met with limited success.

The problem of determining the structure of a symmetric homo-oligomer from ambiguous NOEs was formulated by Nilges and coworkers following the standard NMR approach of optimizing a pseudo-energy function combining biophysical modeling terms and symmetry-enforcing restraints [42]. A simulated annealing protocol is employed to optimize the pseudo-energy, initially considering both intra- and inter-subunit interpretations for the restraints. After a round of optimization, the intra- vs. inter-subunit interpretations are reassessed; those that are inconsistent with the identified structures are eliminated. Another round of optimization is then initialized from the structures and reduced set of restraint disambiguations. The process is repeated in order to identify mutually consistent structures and disambiguations. This approach, called ARIA, has been used to determine a number of symmetric homo-oligomeric structures from ambiguous NOE data (see `aria.pasteur.fr/aria-links/pdb-structures-calculated-using-aria`), and has continued to expand in functionality, e.g., incorporating additional types of restraints such as RDC. The method, however, fails to give any assurances on the correctness

of the obtained structures or to account for the possibility of missing structures. There are no guarantees that the heuristics to escape local minima will work, or that the greedy selections of disambiguations will lead to the native structures.

An alternative approach is based on Rosetta [54]. It first computes monomer structures that satisfy chemical shifts and backbone NOEs; critically, it requires these NOEs already to have been assigned to specific atoms, with intra- vs. inter-subunit ambiguity resolved. It then constructs a complex structure by docking the monomers subject to a symmetric constraint, while also incorporating RDC-derived orientational constraints. Energy refinement is subsequently performed to generate final structures. This method is clearly susceptible to being trapped in local minima by the two-stage process, as well as by the use of stochastic search methods in each stage.

In order to correctly and completely characterize the diversity of structures consistent with the data, we present here a method that employs a branch-and-bound search over a configuration space representation of the subunit and complex structures. The approach builds on our earlier work in structure determination of symmetric homo-oligomers [8, 14, 44], which handled special cases where special experimental techniques successfully resolved the intra- vs. inter-subunit ambiguity. The method presented here goes significantly further, simultaneously determining the subunit structure and the arrangement of the subunits into a complex structure, while partially disambiguating the intra- vs. inter-subunit interpretations of the input NOE distance restraints. Since multiple structures may be consistent with the data, we provide a guarantee of completeness to within a user-specified resolution: our method generates a set of representative structures such that any other structure that is sufficiently consistent with the data has a sufficiently similar representative. As we discuss further below, we determine just the backbone structure within secondary structure elements (SSEs), though the resulting SSE-based structures could readily be used as inputs for loop closure algorithms using additional experimental data [60], with side-chain packing and energy minimization algorithms then employed to obtain the best full confor-

mations.

4.2 Methods

Our overall goal is to take as input a protein sequence and experimental NMR data, and produce as output the complex structure. We leverage earlier work on determination of SSE backbone structures from residual dipolar coupling (RDC) data: RDC-Panda [70] employs a tree-based search algorithm to find sequences of backbone torsion angles in the SSEs that best explain the experimental RDC data. The RDC data also allows determination of the orientation (but not position) of the symmetry axis (uniquely for 3-fold and higher, one of the three eigenvectors of the Saupe Matrix for 2-fold) [2].

Thus we focus here on the problem in which the input includes the SSE backbone structures, the symmetry axis orientation, and a set of ambiguous NOE distance restraints, and the output is a structure placing the SSE backbone structures relative to each other and to the symmetry axis (Fig. 4.2), thereby generating the subunit and complex structure. As discussed in the introduction, there is not likely to be a unique such structure best satisfying the restraints, so the output is actually a set of representatives, such that any other structure satisfying a sufficient number of restraints is sufficiently similar. We define "sufficiently similar" in terms of root mean squared distance (RMSD), and a "sufficient number of restraints" relative to the best solution. Thus our method can be seen as complete to within a user-specified resolution, and does not suffer from problems of myopia, local minima, and so forth.

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

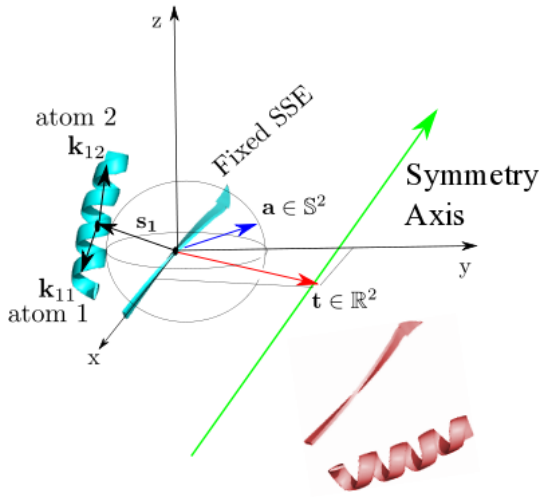


Fig. 4.2: Configuration space representation for simultaneous determination of subunit and complex structure of symmetric homo-oligomers. The backbone SSE structures and symmetry axis orientation are determined from RDC data, so our problem is to place the SSEs relative to each other and to the symmetry axis based on ambiguous NOE distance restraints. The configuration space is parameterized by placing the center of one "fixed" SSE at the origin, and specifying the 3D translations $\mathbf{s}_i$ of the centers of the other SSEs, as well as the position $\mathbf{t}$ of the intersection of the symmetry axis with the x-y plane (its orientation $\mathbf{a}$ is precomputed). Specifying the $\mathbf{s}_i$ then generates a subunit structure (here cyan), while rotating it around the axis $(\mathbf{a}, \mathbf{t})$ generates a complex structure (here a dimer, with the second subunit in red). In assessing restraints, we also use the distances $\mathbf{k}_{ij}$ of atom j in SSE i to the center of the SSE. For side-chain atoms, this requires determination of the side-chain conformation, which we choose from a set of rotamers.

4.2.1 Representation

As summarized in Fig. 4.2, we place the global origin at the center of mass of the backbone atoms of one of the SSEs (the "fixed" SSE, #0) in one of the subunits (the "fixed" subunit, #0). These choices are arbitrary in terms of experimental information. The center of mass of the ith SSE is then expressed as a 3D vector $\mathbf{s}_i$, with $\mathbf{s}_0 = \mathbf{0}$. The position of the symmetry axis is specified by the translation $\mathbf{t}$ of its intersection with the x-y plane; its orientation $\mathbf{a}$ is known.

The coordinates of the atom j within SSE i are specified by translation $\mathbf{k}_{ij}$ from the SSE center. For backbone atoms, the translation is fixed, but for side-chain atoms, they depend on the side-chain conformation, which in turn is restrained by the data and guided

to avoid steric clash. We adopt a rotamer-based representation of side-chain conformations, allowing them to vary over discrete sets of low-energy representations mined from experimental structures. Our results are based on the "penultimate rotamer library" [34], but the method can use any such rotamer library. The side-chain translation vector for atom j in SSE i is thus a member of a set K_{ij} precomputed from the rotamer library. For simplicity of notation, we also use such a set, containing a single member, for backbone atoms.

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

$$Q_{0ij} = \{\mathbf{s}_i + \mathbf{k}_{ij} \mid \mathbf{k}_{ij} \in K_{ij}\} \tag{4.1}$$

$$Q_{rij} = \{R_{\mathbf{a}}(\alpha)(\mathbf{q}_{0ij} - \mathbf{t}) + \mathbf{t} \mid \mathbf{q}_{0ij} \in Q_{0ij}\} \qquad r > 0 \tag{4.2}$$

where $R_{\mathbf{a}}(\alpha)$ is a three dimensional rotation by angle α around axis $\mathbf{a}$, where $\alpha = r2\pi/c$ for c subunits.

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

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

$$\exists\ \mathbf{p} \in P,\ \mathbf{q} \in Q\ s.t.\ \|\mathbf{p} - \mathbf{q}\| \leq \delta \tag{4.3}$$

where P and Q are sets of atom positions for atoms p and q respectively and $\|\cdot\|$ is the Euclidean distance. With respect to our representation, we know the atom and SSE indices of p and q; if we also knew the subunit indices, then P and Q would be determined as one of

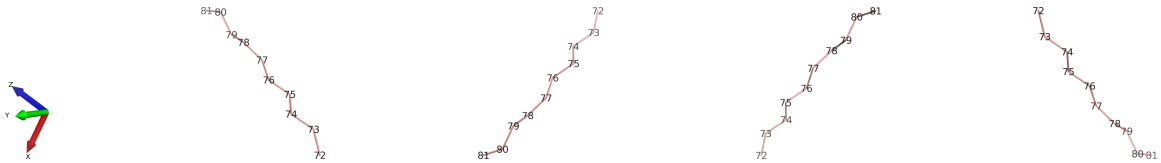


Fig. 4.3: The four orientations possible for an example SSE (residues 72–81) in MinE. The orientations are, in order: original, rotation by 180 degrees around x, around y, and around z.

the Q_{rij}. The fact that we do not know which subunits are involved in each restraint is the key confounding factor in simultaneous determination of subunit and complex structure.

Let us consider how to represent this ambiguity regarding the interpretation of a restraint: is it within a single subunit ("intra") or between two subunits ("inter"), and if between two, which two. We note that experimentally the restraints are "mirrored"—an intra restraint is satisfied in all subunits, and an inter restraint is satisfied in all pairs of subunits the same spacing apart around the cycle. (This is why the choice of fixed subunit doesn't matter.) In our approach, we consider all such interpretations. Only if the interatomic distance is too large in all interpretations (Eq. 4.3) do we consider the restraint to be violated. Thus we essentially express ambiguity as a logical OR over the interpretations.

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

One final source of ambiguity in our representation actually comes with the input SSE backbone structures, each of which is subject to a 180 degree rotation around each of the axes in a manner that yields 4 images of the structure, called "orientations" [70] (Fig. 4.3). We do not directly represent this ambiguity, but instead simply solve for each combination independently.

4.2.2 Cell-based restraint analysis

The configuration space gives a compact representation for all possible structures; our goal is to find within it representatives for all structures that are sufficiently consistent with the data. To consider the feasible region in SCS, let us for a moment ignore the various sources of ambiguity. The SCS-to-Euclidean conversion (Eq. 4.1 and 4.2) is a linear transformation, and each distance restraint (Eq. 4.3) specifies a ball in Euclidean space. The pseudoinverse of the transformation (which is indeed of rank 3) transforms the feasible ball to an ellipsoid in SCS. A feasible configuration is the ellipsoid crossed with the the null space, which makes it an infinite-length cylinder with an ellipsoidal cross-section. Thus we need to compute the intersections of these cylinders. Once we incorporate the ambiguity, however, we would have to compute an exponential number of such intersections.

Thus instead of trying to exactly compute the feasible portion of the configuration space, we subdivide the space into regions containing relatively similar structures and evaluate the discretized regions. In particular, we use a cell-based representation of configuration space regions, where a "cell" is an axis-aligned box $T \times S_1 \times \ldots S_{m-1}$ with T an axis-aligned rectangle in $\mathbb{R}^2$ containing the symmetry axis translations and S_i an axis-aligned cuboid in $\mathbb{R}^3$ containing the translations of SSE i. In evaluating a cell, we want to know how many restraints its various structures satisfy.

First let us bound the positions of the atoms. Recall that Eq. 4.1 and 4.2 define the possible positions based on specific choices for $\mathbf{t}$ and $\mathbf{s}_i$; $\mathbf{k}_{ij}$ is constant when assessing a particular rotamer. Our cell representation allows $\mathbf{t}$ and $\mathbf{s}_i$ to range across axis-aligned boxes T and S_i. Thus for an atom in the fixed subunit, extending Eq. 4.1 over $\mathbf{s}_i \in S_i$ simply displaces the box S_i by the vector $\mathbf{k}_{ij}$. For an atom in another subunit, Eq. 4.2,

when expanded out, becomes:

$$\begin{aligned} Q_{rij} &= \{R_{\mathbf{a}}(\alpha)(\mathbf{s}_i + \mathbf{k}_{ij} - \mathbf{t}) + \mathbf{t} \,|\, \mathbf{t} \in T, \mathbf{s}_i \in S_i\} \\ &= \{R_{\mathbf{a}}(\alpha)(\mathbf{s}_i - \mathbf{t}) + \mathbf{t} + R_{\mathbf{a}}\mathbf{k}_{ij} \,|\, \mathbf{t} \in T, \mathbf{s}_i \in S_i\} \end{aligned} \tag{4.4}$$

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

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

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

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

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

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

While these tests allow us to evaluate the two extreme cases for each restraint, the

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

The algorithm is described formally in Alg. 2. We form position-specific sets such that R_i contains all restraints in which some atom from residue i participates. (Since each restraint has two atoms, it can appear in two such sets.) We work from N terminus to C terminus. When considering residue i, we examine each of its possible rotamers, identifying the one (a) that supports the satisfaction of the most restraints. (We don't allow a restraint to be considered satisfied for both its different residues.) To do so, for each remaining restraint involving an atom from residue i, we consider all possible rotamers b for the other atom, and evaluate the resulting interatomic distance over the given cell. Here function d computes Euclidean distance after applying the configuration space transformation for parameters $\mathbf{c}$ and using the intra-SSE distance vectors $\mathbf{k}$ for rotamers a and b, as in Eq. 4.1 and 4.2. After identifying the best such rotamer, we add to our list all the restraints it satisfies, and continue.

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

Theorem 4.2.1. *The size of set S at iteration i of Alg. 2 is an overestimate of the size of the optimal set of restraints from $R_1 \cup \ldots \cup R_i$ that can be satisfied when choosing for each position 1 through i a unique rotamer.*

Algorithm 2 Simultaneous restraint satisfaction bound

Input: Sets R_i ($1 \leq i \leq n$) of restraints with at least one atom from residue i
Input: Sets A_i ($1 \leq i \leq n$) of rotamers for residue i
Input: Cell C.
Output: $|S|$: bound on number of satisfied restraints

$S \leftarrow \emptyset$
for $i = 1 \rightarrow n$ **do**
  $U \leftarrow R_i \backslash S$
  $a \leftarrow \arg\max_{a \in A_i} |\{(p, q, \delta) \in U \mid \exists\, b \in A_{\text{resi}(q)},\ \mathbf{c} \in C\ s.t.\ d(p, q;\, a, b, \mathbf{c}) \leq \delta\}|$
  $S \leftarrow S \cup \{(p, q, \delta) \in U \mid \exists\, b \in A_{\text{resi}(q)},\ \mathbf{c} \in C\ s.t.\ d(p, q;\, a, b, \mathbf{c}) \leq \delta\}$
end for

Proof. Let the set computed by our algorithm be denoted by X_i and the optimal set (with a unique rotamer per position) by O_i. The proof is by induction. For the base case $i = 1$, our algorithm finds the maximum number of satisfied restraints using any rotamer at position 1, an overestimate of the actual number which would restrict the other side of the restraint; i.e., $|X_i| \geq |O_i|$.

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

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

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

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

As a corollary, when it terminates at $i = n$, Alg. 2 produces an overestimate of the size of the optimal set of consistently satisfied restraints.

4.2.3 Cell structural uniformity assessment

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

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

$$
\begin{aligned}
SD &= \|(R_{\mathbf{a}}(\alpha)(\mathbf{s}_i + \mathbf{k}_{ij} - \mathbf{t}) + \mathbf{t}) - \\
&\qquad (R_{\mathbf{a}}(\alpha)((\mathbf{s}_i + \Delta\mathbf{s}_i) + \mathbf{k}_{ij} - (\mathbf{t} + \Delta\mathbf{t})) + (\mathbf{t} + \Delta\mathbf{t}))\|^2 \\
&= \|R_{\mathbf{a}}(\alpha)(\Delta\mathbf{s}_i - \Delta\mathbf{t}) + \Delta\mathbf{t}\|^2 \\
&= \left\| \begin{pmatrix} R_{\mathbf{a}}(\alpha) & I - R_{\mathbf{a}}(\alpha) \end{pmatrix} \begin{pmatrix} \Delta\mathbf{s}_i \\ \Delta\mathbf{t} \end{pmatrix} \right\|^2 \\
&\leq \left\| \begin{pmatrix} R_{\mathbf{a}}(\alpha) & I - R_{\mathbf{a}}(\alpha) \end{pmatrix} \right\|^2 \left\| \begin{pmatrix} \Delta\mathbf{s}_i \\ \Delta\mathbf{t} \end{pmatrix} \right\|_2^2
\end{aligned}
\tag{4.5}
$$

Note that the quantities are independent of individual atom coordinates, so the total contribution is proportional to the number of atoms. Let n_i be the number of atoms in SSE

i (from 1 to $m-1$, omitting fixed SSE number 0). Let z_k be the matrix norm in Eq. 4.5 for subunit k (from 1 to $c-1$, omitting fixed subunit 0); it is given by the highest singular value in the matrix. Now we can compute an upper bound on $g_x(y)$ as

$$g_{\mathbf{x}}(\mathbf{y}) \leq \frac{1}{c\sum_{i=1}^{m-1} n_i} \sum_{i=1}^{m-1} n_i \left(\|\Delta \mathbf{s}_i\|^2 + \sum_{k=1}^{c-1} z_k^2 \left\| \begin{pmatrix} \Delta \mathbf{s}_i \\ \Delta t \end{pmatrix} \right\|^2 \right) \tag{4.6}$$

Let $u_{\mathbf{x}}(\mathbf{y})$ denote this upper bound. We now show that we can bound $g_{\mathbf{x}}(\mathbf{y})$ for any two points in a cell by $u_{\mathbf{a}}(\mathbf{b})$ for corner points $\mathbf{a}$ and $\mathbf{b}$.

Proposition 4.2.2. *For a cell C, $\max_{\mathbf{x}\in C} \max_{\mathbf{y}\in C} u_{\mathbf{x}}(\mathbf{y}) = u_{\mathbf{a}}(\mathbf{b})$ for a pair $\mathbf{a}, \mathbf{b}$ of corner points of C.*

Proof. Assume for contradiction that the maximum is at some $\mathbf{d}$ and $\mathbf{e}$ in C, one or both of which is not a corner, and that this value is strictly greater than the values between all pairs of corners. First let us consider $u_{\mathbf{d}}(\mathbf{x})$, for $\mathbf{x} \in C$. This is a convex function, since it is a sum of convex functions. Therefore, its maximum is attained at one of the corners of C, say $\mathbf{a}$; thus $u_{\mathbf{d}}(\mathbf{a}) \geq u_{\mathbf{d}}(\mathbf{e})$. Now let us consider $u_{\mathbf{a}}(\mathbf{x})$, also a convex function, with maximum at some corner $\mathbf{b}$ of C. Thus $u_{\mathbf{a}}(\mathbf{b}) \geq u_{\mathbf{a}}(\mathbf{d})$. Since $u_{\mathbf{a}}(\mathbf{d}) = u_{\mathbf{d}}(\mathbf{a})$ and we showed that $u_{\mathbf{d}}(\mathbf{a}) \geq u_{\mathbf{d}}(\mathbf{e})$, it follows that $u_{\mathbf{a}}(\mathbf{b}) \geq u_{\mathbf{d}}(\mathbf{e})$, which is a contradiction. □

By inspection we can determine that the lower corner (with smallest value for each element) and upper corner (with largest for each) yield the largest u, as they provide the maximal $\|\Delta \mathbf{t}\|$ and $\|\Delta \mathbf{s}_i\|$. Thus to assess cell uniformity, we evaluate Eq. 4.6 at the lower and upper corners.

4.2.4 Search algorithm

We now develop a branch-and-bound search algorithm that hierarchically subdivides the SCS, using the cell-based evaluations to assess restraint satisfaction within the cells and

to identify terminal cells that need not be further divided. Ultimately the search identifies cells that satisfy the most restraints and that have sufficiently uniform structures. Some of the cells might be similar to each other, so a clustering process yields the final set, from which are generated representatives for all structure sufficiently consistent with the data.

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

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

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

- Cells that are sufficiently uniform (we use a threshold of 1 Å RMSD for our results) are considered terminal and tested by Alg. 2 for restraint satisfaction, updating τ when appropriate.

- Cells that are sufficiently small (we use a threshold of 2 Å in each dimension) are tested for complete satisfaction (the expense of the test is not justified for larger cells). If the number of restraints that aren't completely satisfied is at most τ, then the cell appears good but not sufficiently uniform. Thus we repeatedly subdivide it

until the subcells are sufficiently uniform. We consider them terminal and continue as above.

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

Upon termination, we reevaluate the terminal cells, ordered by the number of violations, against the final τ. Those having fewer violations than the final τ are tested for steric clash within the structure represented by the cell center. Those that pass are considered *accepted cells*. The accepted cells are clustered to reduce redundancy while still ensuring that all satisfying structures are represented by a sufficiently close solution. The clustering is performed using the Euclidean distance metric on a KD-Tree constructed from the cell centers. The RMSD upper bound used to assess structural uniformity in the cell (Sec. 4.2.3) also enables us to use Euclidean distances on SCS points after scaling the coordinates appropriately.

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

4.3 Results

We applied our approach to three test cases with experimental NOE restraints: (1) MinE [29] (PDB id 1EV0), a dimer with one α-helix and two β-strands restrained by 1109 NOEs (926 intra + 183 inter); (2) *B. subtilis* Anti-TRAP (PDB id 2KO8), a trimer with one α-helix and two β-strands (along with a third that is unrestrained and thus not considered here) restrained by 863 NOEs (378 intra + 485 inter); (3) the cytoplasmic domain structure of BM2 proton channel from influenza B virus [66] (PDB id 2KJ1), a tetramer with two α-helices restrained by 400 NOEs (340 intra + 60 inter). NOEs were obtained from the BioMagResBank (BMRB) [61] and the intra vs. inter resolution was ignored. We used

the deposited SSE backbone structures and axis orientation, since these proteins lacked the RDC data necessary to determine them by RDC-Panda. The RMSD cut-off for cell uniformity was set to 1 Å and the initial maximum NOE restraint violation τ was set to ten percent of the total number of restraints.

We first determined which restraints supported only an inter interpretation, as the involved atoms were in the same SSE but no rotamer choice could place them close enough. For MinE, 104 restraints were classified as inter, all consistent with the deposited interpretation. For Anti-TRAP, 173 restraints were classified as inter, but 60 of them were actually intra according to the deposited interpretation. For BM2, 8 restraints were classified as inter, 4 of which were actually intra. The misclassifications were due to the use of discrete rotamers, which did not come sufficiently close; possible fixes include relaxing the distance threshold or using rotamer "voxels" [16]. Note that while the preprocessing forced some incorrect interpretations, they are due to the geometric model and the same interpretations would ultimately have resulted from the search algorithm; the preprocessing is simply a time-saving measure.

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

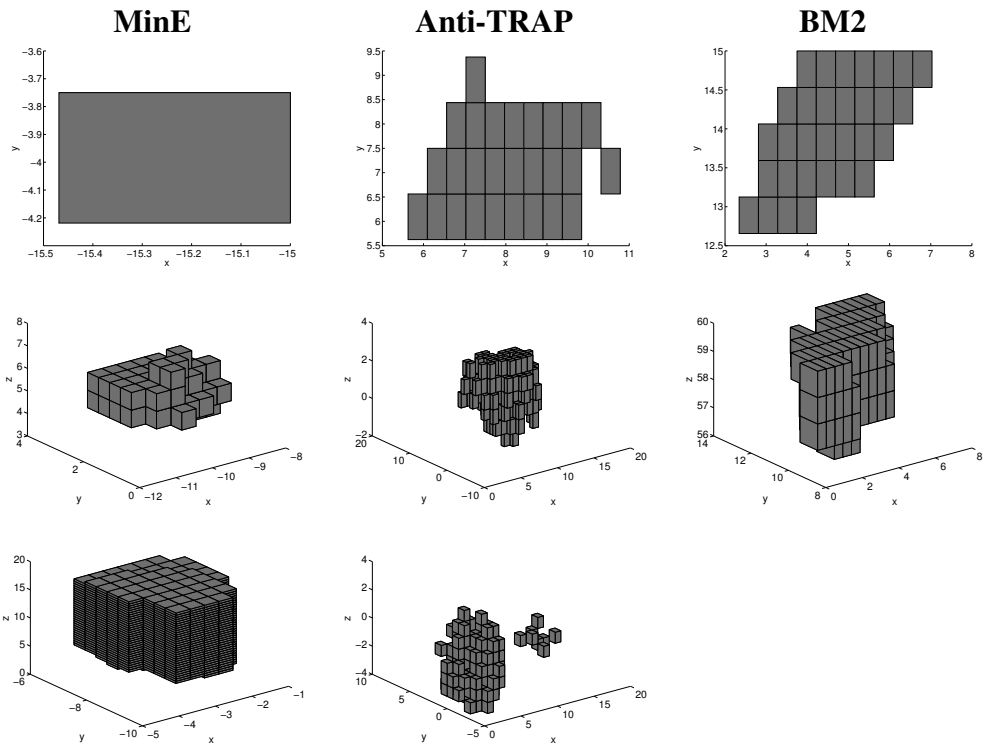


Fig. 4.4: Accepted SCS cells for the test cases. (top) The translation of the x–y intersection of the symmetry axis. (middle, bottom) The translation of the non-fixed SSEs. For MinE, one β-strand is fixed while the other β-strand (middle) and the α-helix (bottom) are translated. For Anti-TRAP, the α-helix is fixed and the β-strands (middle and bottom) translated. For BM2, one α-helix is fixed and the other (middle) translated.

4.3.1 Configuration space search

The configuration space search yields a set of accepted cells representing the feasible regions; each cell specifies the 2D translation of the symmetry axis relative to the fixed subunit (T) and the 3D translation of each SSE relative to the fixed SSE (S_i). Recall that we do the search independently for each set of SSE orientations Fig. 4.3. Fig. 4.4 illustrates the accepted cells for the most populated orientation set (i.e., the one with the largest volume) for each of our three test cases. Note that while the different components of the cells are displayed separately, not all combinations of these components are accepted.

For MinE, eight of the sixteen SSE orientation combinations led to accepted cells. The two most populated combinations were nearly equal in number and different from the deposited structure. The first combination (34% of the remaining volume) had the sheet

rotated around z axis and the second (33% of the remaining volume) had both beta strand and alpha helix rotated around the z axis. The combination of the deposited structure contained 13% of the remaining volume, as did another combination with the alpha helix rotated around the z axis. We later discuss the resulting conformations, but wanted to point out here that the differences in orientations in the configuration space indeed lead to differences in conformations; e.g., the average RMSD for samples in the two most populated orientation sets was 4.6 Å with a maximum of 12.8 Å. The remaining volume of the translational component of the symmetry axis was 2×10^{-5} that of the initial volume. The symmetry axis was highly constrained by inter-subunit restraints. For the SSEs, the relative volumes of the translational components were 1×10^{-5} for SSE 1 and 2×10^{-4} for SSE 2. The β-strand was restrained to the fixed SSE by 237 NOEs, yielding a relatively restricted remaining translational component (10 Å^3). The α-helix, in contrast, was relatively unrestrained, with only 26 NOEs to the fixed SSE, resulting in much more translational uncertainty (32 Å^3). There are no intra restraints between the two non-fixed SSEs and the inter restraints are therefore valuable in pinning down the structure.

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

groups, with the second SSE much further away from the fixed one in one than in the other.

For BM2 only one SSE orientation combination, that in the deposited structure, produced accepted cells. 7×10^{-4} of the symmetry axis translation volume was accepted, while 1×10^{-5} of the SSE translation volume remained. Though there are only four inter-subunit restraints between the fixed SSE in each subunit, the axis translation was tightly characterized (7 Å^2), and likewise the two restraints from SSE 1 to the fixed SSE sufficed to reduce its translational uncertainty to 47 Å^3. The NOEs acted in concert with backbone steric clash to drastically prune the configuration space.

4.3.2 Example structures

To illustrate the diversity of structures represented by the final accepted SCS cells, we performed agglomerative clustering on all of the cells (from all orientation combinations) and selected an example from each of the most distinct groups. See Fig. 4.5.

MinE. The top level of the dendrogram represents an RMSD of 10 Å, indicating substantial diversity in the structures. However, chopping the tree into 8 clusters yields compact groups, each with no more than 1.5 Å RMSD among its members. Fig. 4.5(top) illustrates one sample from each cluster. Between these samples, the non-fixed alpha helix for the first subunit had an RMSD up to 24 Å, while the beta sheet had an RMSD up to 16 Å.

Anti-TRAP. The accepted cells yielded much more similar structures, with a maximum RMSD at the top of the dendrogram of 4 Å. There are only three clusters that have a maximum RMSD of 3.5 Å in structures within them. The example structures in Fig. 4.5(middle) illustrate this relative uniformity of identified representatives. The SSE 2a had the most variance between these samples, as much as 23 Å.

BM2. This structure was the best determined, with an RMSD of only 0.7 Å at the top of the dendrogram, resulting from the relatively compact set of accepted cells. The six

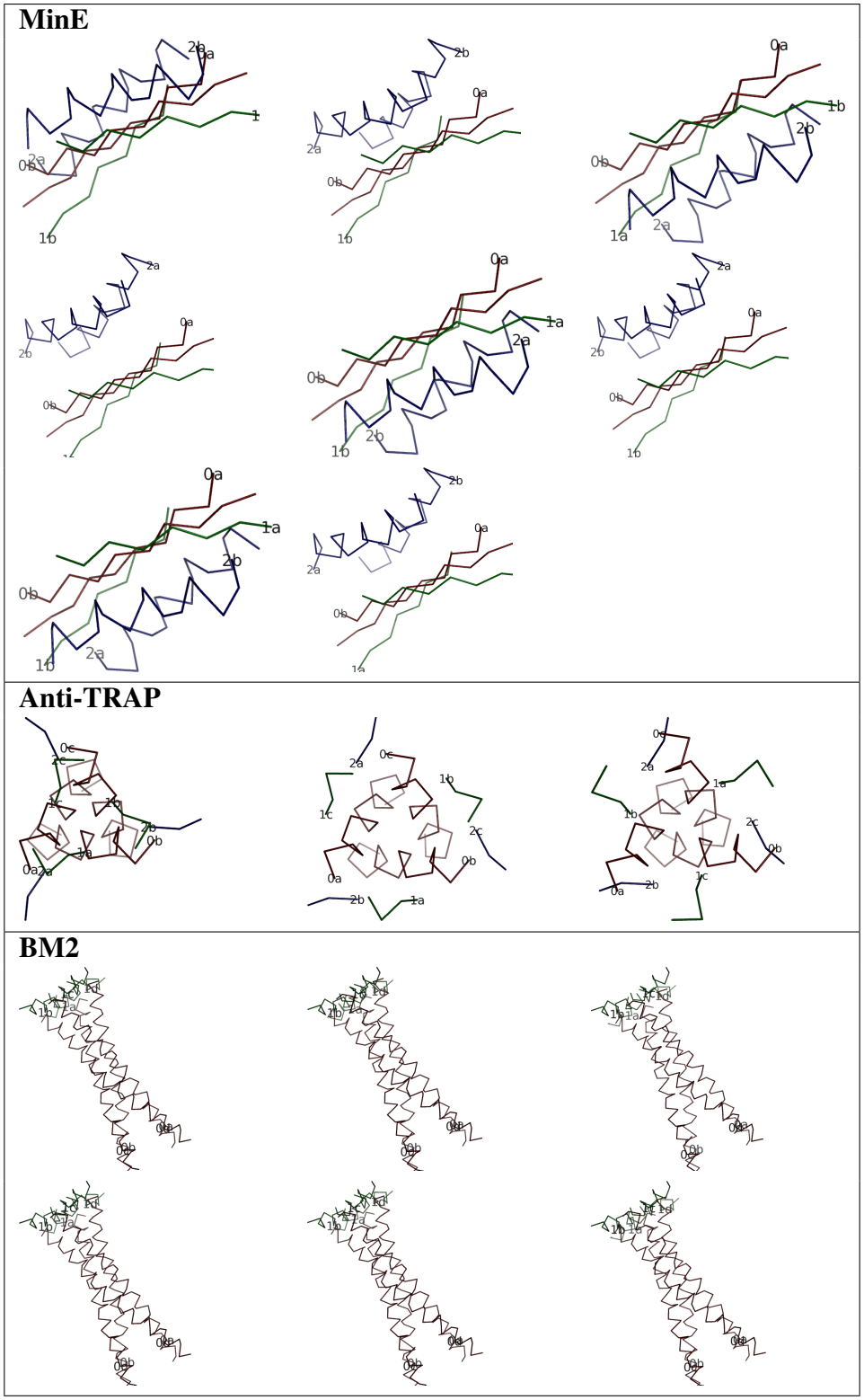


Fig. 4.5: Diverse example structures from satisfying SCS cells. The SSEs are labeled 0a, 1b etc., where the number indexes the SSE and the character the subunit (e.g., 0a is the fixed SSE in the first subunit).

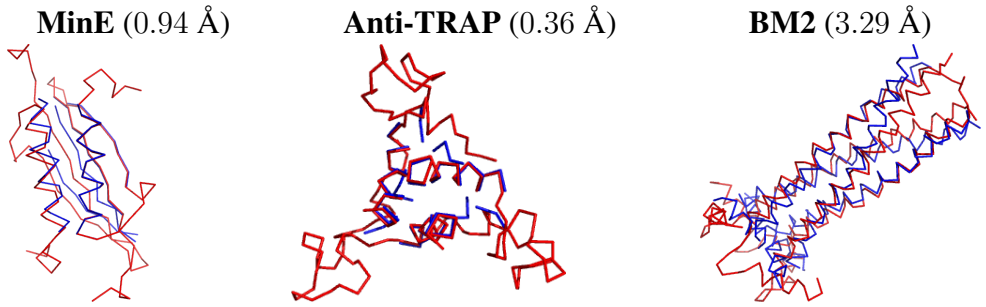


Fig. 4.6: Superpositions of lowest-energy deposited structure (red) and closest representative from our search (blue). Parenthesized numbers are RMSDs.

example structures from the top-most clusters (Fig. 4.5, bottom) emphasize this point.

4.3.3 Comparison to structures from previous methods

The structures identified by our method represent the deposited structures well. Fig. 4.6 shows that for MinE, the minimum-energy deposited structure is 0.94 Å away from the closest member in our ensemble. Similarly, the minimum-energy deposited structure for Anti-TRAP is represented by one representative with only 0.36 Å RMSD. On the other hand, BM2's representative is 3.29 Å RMSD, larger than we would have expected. However, the deposited structure violates ten restraints, whereas our ensemble was comprised of structures violating only one restraint. Relaxing the allowed number of violations would have enabled us to find the deposited structure (along with many others violating more restraints than those we found with the tight restriction).

Going beyond the lowest energy structures in the deposited samples, we can see that each deposited structure is represented by one of ours about as well as the lowest-energy one is (Fig. 4.7, top). However, our structures capture much more diversity (Fig. 4.7, bottom), as some of them are quite different from their most similar counterpart among the deposited structures. Of course, the feasible region we obtain is an overestimate of the true feasible region, since we have relaxed the evaluation of consistency of constraints individually (via bounds) and simultaneously, are using a cell-based discretization of the space,

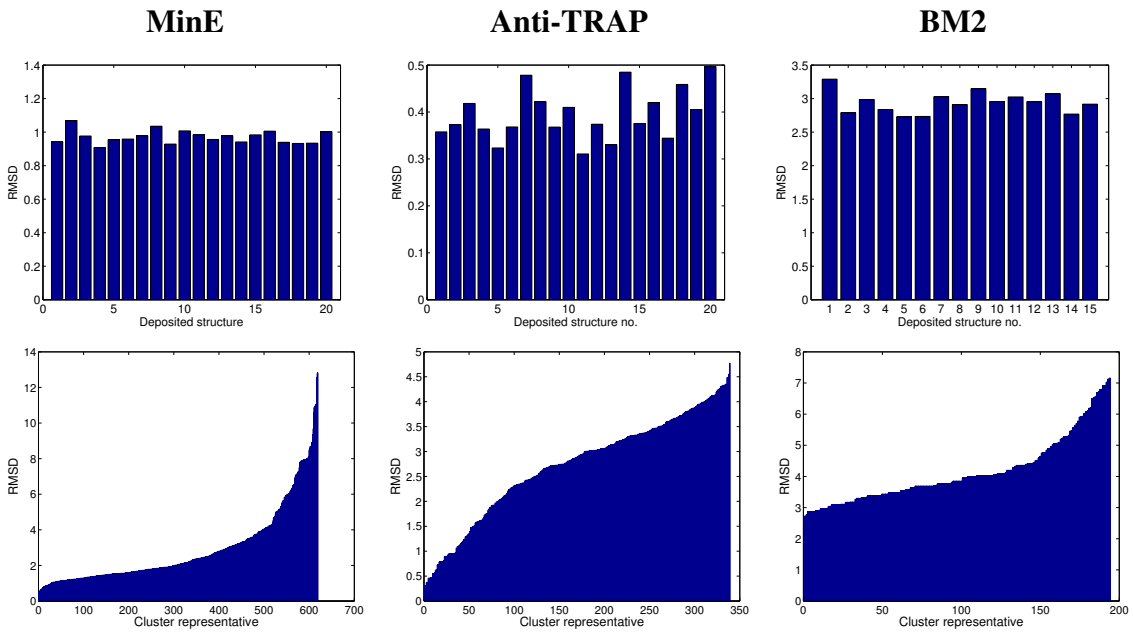


Fig. 4.7: RMSDs between deposited structures (in order within the pdb file) and those identified by our search. (top) The closest of ours to the minimum energy structure in the deposited ensemble. (bottom) The closest deposited structure to each of our models (in order of RMSD).

and are focused just on the SSE structures (which would be further filtered by restraints to and packing with loops). A complete protocol for structure determination could proceed from this point by incorporating loops and refining in conformation space.

4.4 Conclusion

We have developed an approach to fully account for intra vs. intersubunit ambiguity in NOE data for symmetric homo-oligomers, simultaneously determining the subunit and complex structures most consistent with the data. In contrast to search approaches that are heuristic in nature and can get trapped in local minima, our approach partitions a configuration space that represents all possible structures, using a set of restraint satisfaction tests to identify the regions that best satisfy the data. This search procedure enables us to provide guarantees on the results, namely that any structure sufficiently consistent with the data is sufficiently close to one of the identified representatives. We demonstrated with three test cases that

the approach effectively prunes the search space and identifies diverse structures consistent with the data. In future work, we can take into account additional ambiguities, including NOE assignment (similar to [45]), as well as multiple possible symmetry axes [37].

Acknowledgement

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

5. SUMMARY AND FUTURE WORK

This thesis has developed methods that are able to determine the association model from input solution scattering data and symmetric homo-oligomer structure from NOE restraints. The method for determining the association model from SAS data was (1) robust to noise and (2) robust to the presence of contaminants. The structure determination methods were (1) complete in identifying all conformations (within a similarity threshold) that were consistent with NOE restraints (2) data driven in that we kept solutions that satisfied the data without having steric clashes and did not model the biophysical energy of the structures.

5.1 Future work

Association model In our work with contaminants in the association model, we incorporate constraints for non-negativity and best fit to the data while optimizing for the smoothness of the curve. Such formulation, can at times lead to non physical scattering curves getting obtained and we may end up with infeasible optimization problems. Future work should incorporate algebraic approximations of a physical scattering curve that can retain the convex nature of the optimization problem.

Another possibility in improving the method is to use heuristics like simulated annealing when doing searches in three dimensions. This can lead to faster runtimes of the method because grid searches in dimensions greater than two are slow.

Homo-oligomeric structure from NMR In the structural inference work on symmetric homo-oligomers, we chose a weak prior that was uniform for all structures exhibiting no steric clash. We would like to, however, better account for biophysical plausibility by incorporating a Boltzmann prior that represents molecular modeling energies. Bounding such a prior and then using it in the framework that we created, however, remains a significant challenge. One can pick energy functions that are more coarse grained but can be bounded and fit in the inferential structure determination framework that we developed.

In the work with structure determination from ambiguous NOE restraints, the thesis computed the symmetry axis and used it in a branch and bound algorithm. In order to make the method more robust to experimental noise, orientations can be sampled from a distribution centered at the computed axis instead of picking just one axis. This would make it less sensitive to noisy RDC data. During branching, instead of splitting a cell into two equal halves, better heuristics can be used to split it into two unequal halves. One with higher probability of satisfaction and the other with lesser.

Bibliography

[1] T. Akutsu. NP-hardness results for protein side-chain packing. *Genome Inform.*, 1997.

[2] H. M. Al-Hashimi, P. J. Bolon, and J. H. Prestegard. Molecular symmetry as an aid to geometry determination in ligand protein complexes. *Journal of Magnetic Resonance*, 142(1):153–158, 2000.

[3] A. K. Attri and A. P. Minton. Composition gradient static light scattering: a new technique for rapid detection and quantitative characterization of reversible macromolecular hetero-associations in solution. *Anal. Biochem.*, 346:132–138, 2005.

[4] 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 Res.*, 28:235–242, 2000.

[5] P. Bernadó, Y. Pérez, J. Blobel, J. Fernández-Recio, D. I. Svergun, and M. Pons. Structural characterization of unphosphorylated STAT5a oligomerization equilibrium in solution by small-angle X-ray scattering. *Protein Science*, 18(4), 2009.

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

[7] P. Chacón, F. Morán, J. F. Diaz, E. Pantos, and J. M. Andreu. Low-resolution structures of proteins in solution retrieved from x-ray scattering with a genetic algorithm. *Biophys. J.*, 74(6):2760–2775, 1998.

[8] H. Chandola, A. K. Yan, S. Potluri, B. R. Donald, and C. Bailey-Kellogg. NMR structural inference of symmetric homo-oligomers. *J. Comp. Biol.*, 12:1757–1775, 2011.

[9] L. Chen, K. O. Hodgson, and S. Doniach. A lysozyme folding intermediate revealed by solution X-ray scattering. *J. Mol. Biol.*, 261:658–671, 1996.

[10] S. G. Codreanu, L. C. Thompson, D. L. Hachey, H. W. Dirr, and R. N. Armstrong. Influence of the dimer interface on glutathione transferase structure and dynamics revealed by amide H/D exchange mass spectrometry. *Biochemistry*, 44:10605–10612, 2005.

[11] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, 2004.

[12] D. G. Dervichian, G. Fournet, and A. Guinier. X-ray scattering study of the modifications which certain proteins undergo. *Biochim. Biophys. Acta.*, 8:145–149, 1952.

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

[14] B. R. Donald and J. Martin. Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 55(2):101–127, 2009.

[15] L. A. Feigin and D. I. Svergun. *Structure analysis by small-angle x-ray and neutron scattering*. Plenum Press, New York, 1987.

[16] I. Georgiev, R. Lilien, and B. R. Donald. Improved pruning algorithms and divide-and-conquer strategies for dead-end elimination, with application to protein design. *Bioinformatics*, 22:174–183, 2006.

[17] I. Georgiev, R. H. Lilien, and B. R. Donald. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *J. Comput. Chem.*, 29:1527–1542, 2008.

[18] D. S. Goodsell and A. J. Olson. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.*, 29:105–153, 2000.

[19] A. Guinier and G. Fournet. *Small-Angle Scattering of X-rays*. Wiley, New York, 1955.

[20] 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.

[21] 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.

[22] M. Habeck, M. Nilges, and W. Rieping. Replica-exchange Monte Carlo scheme for Bayesian data analysis. *Phys. Rev.*, 94:01805, 2005.

[23] M. Habeck, W. Rieping, and M. Nilges. Weighting of experimental evidence in macromolecular structure determination. *PNAS*, 103:1756–1761, 2006.

[24] M. Ikura and A. Bax. Isotope-filtered 2D NMR of a protein peptide complex-study of a skeletal-muscle myosin light chain kinase fragment bound to calmodulin. *J. Am. Chem. Soc.*, 114:2433–2440, 1992.

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

[26] K. Kameyama and A. P. Minton. Rapid quantitative characterization of protein interactions by composition gradient static light scattering. *Biophys. J.*, 90:2164–2169, 2006.

[27] R. E. Kass and L. A. Wasserman. The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.*, 91:1343–1370, 1996.

[28] 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.

[29] 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.

[30] J. Lebowitz, M. S. Lewis, and P. Schuck. Modern analytical ultracentrifugation in protein science: A tutorial review. *Protein Science*, 11:2067–2079, 2002.

[31] J. Lebowitz, M. S. Lewis, and P. Schuck. Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein Sci.*, 11:2067–79, 2002.

[32] W. Lee, M. J. Revington, C. Arrowsmith, and L. E. Kay. A pulsed-field gradient isotope-filtered 3D C-13 HMQC-NOESY experiment for extracting intermolecular NOE contacts in molecular-complexes. *FEBS Letters*, 350:87–90, 1994.

[33] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.*, 104:4546–4559, 1982.

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

[35] 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 carboncarbon couplings and 13C chemical shifts. *J. Biomol. NMR*, 7:256–260, 1996.

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

[37] J. W. Martin, A. K. Yan, C. Bailey-Kellogg, P. Zhou, and B. R. Donald. A graphical method for analyzing distance restraints using residual dipolar couplings for structure determination of symmetric protein homo-oligomers. *Protein Science*, 20:970–985, 2011.

[38] I. L. Mathews, T. J. Kappock, J. Stubbe, and S. E. Ealick. Crystal structure of *Escherichia coli* PurE, an unusual mutase in the purine biosynthetic pathway. *Structure*, 7:1395–1406, 1999.

[39] M. G. Moran and P. A. P. Kendall. *Geometrical Probability*. Charles Griffin & Co. Ltd., 1963.

[40] M. Nilges, A. Bernard, B. Bardiaux, T. Malliavin, M. Habeck, and W. Rieping. Accurate NMR structures through minimization of an extended hybrid energy. *Structure*, 16:1305–1312, 2008.

[41] M. Nilges, M. Habeck, S. L. O'Donoghue, and W. Rieping. Error distribution derived NOE distance restraints. *Proteins*, 64:652–664, 2006.

[42] S.I. O'Donoghue and M. Nilges. Calculation of symmetric oligomer structures from NMR data. In *Biological Magnetic Resonance*, volume 17, pages 131–161. Springer US, 2002.

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

[44] 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*, 65:203–219, 2006.

[45] S. Potluri, A. K. Yan, J. J. Chou, B. R. Donald, and C. Bailey-Kellogg. A complete algorithm to resolve ambiguity for inter-subunit NOE assignment in structure determination of symmetric homo-oligomers. *Protein Science*, 16:69–81, 2007.

[46] S. Potluri, A. K. Yan, B. R. Donald, and C. Bailey-Kellogg. A complete algorithm to resolve ambiguity for inter-subunit NOE assignment in structure determination of symmetric homo-oligomers. *Protein Sci.*, 16:69–81, 2007.

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

[48] W. Rieping, H. Michael, and M. Nilges. Inferential structure determination. *Science*, 309:303–306, 2005.

[49] L. A. Santaló. *Integral Geometry and Geometric Probability*. Cambridge University Press, 2 edition, 2002.

[50] J. R. Schnell and J. J. Chou. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature*, 451:591–595, 2008.

[51] B. R. Seavey, E. A. Farr, W. M. Westler, and J. Markley. A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, 1:217–236, 1991.

[52] D. J. Segel, A. Bachmann, J. Hofrichter, K. O. Hodgson, S. Doniach, and T. Kiefhaber. Characterization of transient intermediates in lysozyme folding with time-resolved small-angle X-ray scattering. *J. Mol. Biol.*, 288:489–499, 1999.

[53] D. J. Segel, A. L. Fink, K. O. Hodgson, and S. Doniach. Protein denaturation: A small-angle X-ray scattering study of the ensemble of unfolded states of cytochrome c. *Biochemistry*, 37:12443–12451, 1998.

[54] N. G. Sgourakis, O. F. Lange, F. DiMaio, I. André, N. C. Fitzkeei, P. Rossi, G. T. Montelione, A. Bax, and D. Baker. Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J. Am. Chem. Soc.*, 133:6288–6298, 2011.

[55] D. I. Svergun. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.*, 76:2879–2886, 1999.

[56] D. I. Svergun, C. Barberato, and M. H. Koch. Crysol: a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.*, 28:768–773, 1995.

[57] D. I. Svergun, M. V. Petoukhov, and M. H. Koch. Determination of domain structure of proteins from x-ray solution scattering. *Biophys. J.*, 80:2946–2953, 2001.

[58] D. I. Svergun and H. B. Stuhrmann. New developments in direct shape determination from small-angle scattering. 1. Theory and model calculations. *Acta Crystallographica Section A*, 47(6):736–744, 1991.

[59] A. Tovchigrechko and Ilya A. Vakser. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Research*, 34:W310–W314, 2006.

[60] C. Tripathy, J. Zeng, P. Zhou, and B. R. Donald. Protein loop closure using orientational restraints from NMR data. *Proteins: Structure, Function, and Bioinformatics*, 80:433–453, 2012.

[61] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36:D402–D408, 2007.

[62] A. Velazquez-Campoy, S. A. Leavitt, and E. Freire. Characterization of protein-protein interactions by isothermal titration calorimetry. *Methods Mol. Biol.*, 261:35–54, 2004.

[63] K. J. Walters, H. Matsuo, and G. Wagner. A simple method to distinguish inter-monomer Nuclear Overhauser Effects in homodimeric proteins with C2 symmetry. *J. Am. Chem. Soc.*, 119:5958–5959, 1997.

[64] Dirk Walther, Fred E. Cohen, and Sebastian Doniach. Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *Journal of Applied Crystallography*, 33(2):350–363, 2000.

[65] J. Wang, R. M. Pielak, M. A. McClintock, and J. J. Chou. Solution structure and functional analysis of the influenza B proton channel. *Nat. Struct. Mol. Biol.*, 16:1267–1271, 2009.

[66] J. Wang, R. M. Pielak, M. A. McClintock, and J. J. Chou. Solution structure and functional analysis of the influenza B proton channel. *Nature Structural & Molecular Biology*, 16:1267–1271, 2009.

[67] Tim E. Williamson, Bruce A. Craig, Elena Kondrashkina, Chris Bailey-Kellogg, and Alan M. Friedman. Analysis of self-associating proteins by singular value decomposition of solution scattering data. *Biophys. J.*, 94:4906–4923, 2008.

[68] 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 Sci.*, 6:1734–1745, 1997.

[69] Y. Xu. Characterization of macromolecular heterogeneity by equilibrium sedimentation techniques. *Biophys. Chem.*, 108:141–163, 2004.

[70] J. Zeng, J. Boyles, C. Tripathy, L. Wang, A. Yan, P. Zhou, and B. R. Donald. High-resolution protein structure determination starting with a global fold calculated from exact solutions to the RDC equations. *Journal of Biomolecular NMR*, 45:265–281, 2009.

[71] C. Zwahlen, P. Legaulte, 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 lambda N-Peptide/boxB RNA complex. *J. Am. Chem. Soc.*, 119:6711–6721, 1997.