JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 12, 2011
© Mary Ann Liebert, Inc.
Pp. 1757–1775
DOI: 10.1089/cmb.2010.0327

NMR Structural Inference of Symmetric Homo-Oligomers

HIMANSHU CHANDOLA,[1] ANTHONY K. YAN,[2] SHOBHA POTLURI,[1,3]
BRUCE R. DONALD,[2,4] and CHRIS BAILEY-KELLOGG[1]

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 Nuclear Magnetic Resonance (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 under-sample 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.

Key words: Bayesian inference, distance restraint, posterior distribution characterization, protein complex structure determination.

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 (Güntert et al., 1991, Brünger,

[1] Department of Computer Science, Dartmouth College, Hanover, New Hampshire.
[2] Department of Computer Science, Duke University, Durham, North Carolina.
[3] Pfizer Global Research.
[4] Department of Biochemistry, Duke University Medical Center, Durham, North Carolina.

1993, Güntert et al., 1997), 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 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 (Rieping et al., 2005b). 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 Nuclear Overhauser Effect (NOE) restraints (Rieping et al., 2005b).

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* (Goodsell and Olson, 2000); 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 (Ikura and Bax, 1992, Lee et al., 1994, Zwahlen et al., 1997, Walters et al., 1997). 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 (Oxenoid and Chou, 2005, Schnell and Chou, 2008, Wang et al., 2009). Our problem (Fig. 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-

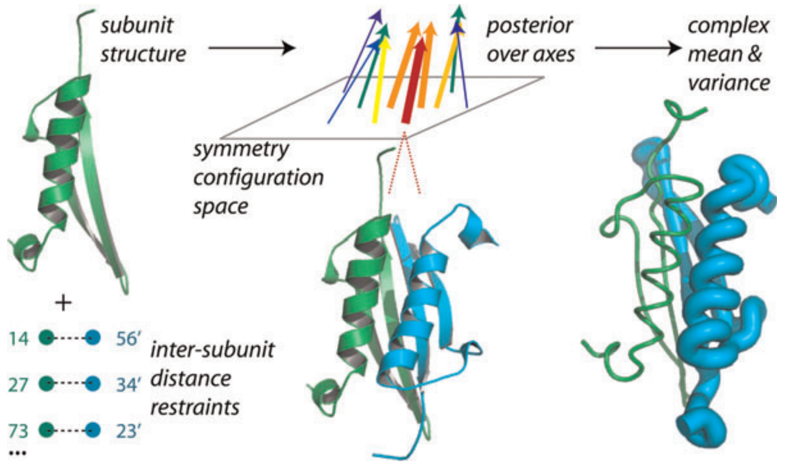


FIG. 1. Structural inference of symmetric homo-oligomers. Given a subunit structure and set of inter-subunit distance restraints, we compute the posterior distribution over all possible complex structures, represented in terms of a configuration space of symmetry axes. The posterior distribution evaluates 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.

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 (Potluri et al., 2006, 2007), but this article 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 (Rieping et al., 2005b, Habeck et al., 2005), which may under-sample the high-dimensional and very rugged posterior distribution of a 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.

2. METHODS

As mentioned, the subunit structure as it exists in complex can be obtained prior to complex structure determination (Oxenoid and Chou, 2005, Schnell and Chou, 2008, Wang et al., 2009). Thus, we are given the subunit structure (Euclidean coordinates $\mathbf{p}_i$ for each atom i) and a set of n distance restraints $R = \{r_1, \ldots, r_n\}$ (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 our previous work (Potluri et al., 2006) 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. 2a). 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 $E(\mathbf{q}_{ij} \mid R)$ and variance $var(\mathbf{q}_{ij} \mid R)$ of the atomic coordinates q_{ij} for each atom i in each rotated subunit j (Fig. 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) to find the high-quality cells and provide guarantees

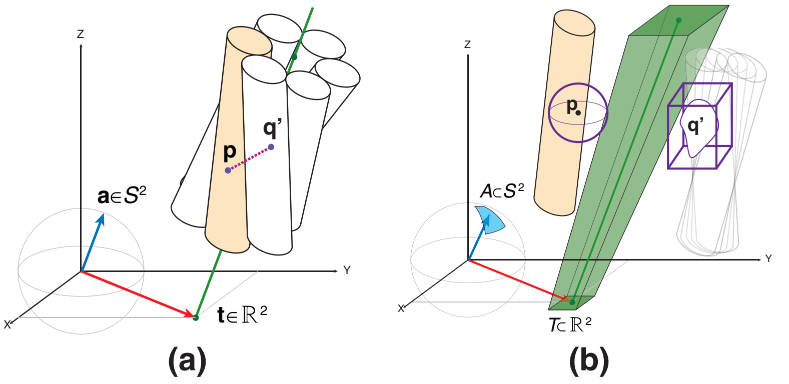


FIG. 2. Symmetry configuration space (SCS). **(a)** Each structure is defined by a point $\mathbb{S}^2 \times \mathbb{R}^2$in the configuration space of symmetry axes. Each subunit is depicted by a cylinder; the structure is obtained by rotating the fixed subunit (shaded cylinder) by the angle of symmetry around the symmetry axis (line). A distance restraint is shown between an atom at position p on the fixed subunit and one at position 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 q' over these structures.

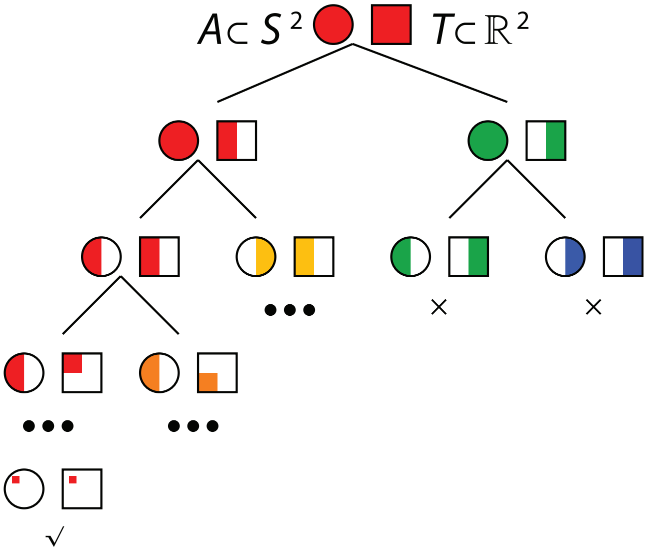


FIG. 3. Hierarchical subdivision of SCS. The 4-dimensional SCS is depicted as two two-dimensional regions: a sphere representing the orientation space $\mathbb{S}^2$ and a rectangle representing the translation space $\mathbb{R}^2$. We compute a bound for the best posterior of a configuration in the 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.

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 (Section 2.1). We then present our inferential framework based on this representation see (Section 2.2), our error bounds (Section 2.3), and our hierarchical subdivision algorithm for computing the posterior and performing the inference (Section 2.4).

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 SCS, $\mathbb{S}^2 \times \mathbb{R}^2$ (Potluri et al., 2006), with orientations from the two-sphere $\mathbb{S}^2$ and translations from the *xy* plane $\mathbb{R}^2$ (Fig. 2a).

Given a symmetry axis $c = (\mathrm{a}, \mathrm{t}) \in \mathbb{S}^2 \times \mathbb{R}^2$ and an angle of rotation $\alpha = 2\pi j/m$ for subunit $j \in \{1, \ldots, (m-1)\}$ (treating the 0th subunit as the fixed one), we compute the coordinates $\mathbf{q}'$ for an atom in a rotated subunit from corresponding coordinates $\mathbf{q}$ for the same atom in the fixed subunit:

$$\mathbf{q}' = \mathcal{T}(c, \mathbf{q}, \alpha) = \mathcal{R}_{\mathbf{a}}(\alpha)(\mathbf{q} - \mathbf{t}) + \mathbf{t} \tag{1}$$

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

In our algorithm for computing the posterior, we consider simultaneously a set of axes in a *cell* of the SCS (Fig. 2b). 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 (Potluri et al., 2006).

$$\mathbf{q}' \in \mathcal{B}(C, \mathbf{q}, \alpha) \tag{2}$$

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 $c \in C$:

$$l_i(C) \le \min_{c \in C} \|\mathbf{p}_i - \mathcal{T}(c, \mathbf{q}_i, \alpha)\| \ge u_i(C) \tag{3}$$

For the geometric computations giving these bounds, we refer the reader to Potluri et al. (2006). In Section 2.3, we apply these bounds to derive upper and lower bounds on the posterior $p(c \mid R)$.

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 (Rieping et al., 2005b). 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 (Lipari and Szabo, 1982) and spin diffusion (Macura and Ernst, 1980).

$$p(c \mid R) = \int p(c, \sigma \mid R)\, \mathrm{d}\sigma \tag{4}$$

$$\propto \int p(R \mid c, \sigma) p(c) p(\sigma)\, \mathrm{d}\sigma \tag{5}$$

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.

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

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

$$p(R \mid c, \sigma) = \prod_{i=1}^{n} p(r_i | c, \sigma) \tag{6}$$

To evaluate a single restraint r_i, we adopt the log-normal distribution advocated by Nilges and co-workers (Rieping et al., 2005a, Habeck et al., 2006) 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 (Nilges et al., 2006); 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:

$$p(r_i \mid c, \sigma) = \frac{1}{\sqrt{2\pi}\sigma d_i} \exp\left(-\frac{1}{2\sigma^2} \log^2 g_i(c)\right), \tag{7}$$

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

$$g_i(c) = \frac{d_i}{\|\mathbf{p}_i - \mathcal{T}(c, \mathbf{q}_i, 2\pi/m)\|}. \tag{8}$$

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 $2\pi/m$).

2.2.2. 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 (Jeffreys, 1946):

$$p(\sigma) = 1/\sigma. \tag{9}$$

2.2.3. Prior $p(c)$.

Laplace postulated that in the absence of sufficient reason, each point in the parameter space should be assigned a uniform prior (Kass and Wasserman, 1996). 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.

$$p(c) = \begin{cases} 0, & \text{if complex structure has steric clash} \\ 1, & \text{otherwise.} \end{cases} \tag{10}$$

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

Section 2.2.5 below details how to appropriately define (and integrate) such a probability distribution over the SCS parameterization.

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

$$\begin{aligned} p(c \mid R) &= \int_0^\infty p(c, \sigma \mid R) \mathrm{d}\sigma \\ &\propto p(c) \int_0^\infty \sigma^{-(n+1)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \log^2 g_i(c) \right) \mathrm{d}\sigma \\ &\propto p(c) \left(\sum_{i=1}^n \log^2 g_i(c) \right)^{-n/2}. \end{aligned} \tag{11}$$

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. 11.

2.2.5. 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 Section 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 over-represents 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 (Moran and Kendall, 1963; Santaló, 2002). Applying that approach with SCS parameters gives an infinitesimal invariant volume $\mathrm{d}\mu$:

$$\mathrm{d}\mu = |\cos\theta| \sin\theta \; \mathrm{d}\theta \; \mathrm{d}\phi \; \mathrm{d}x \; \mathrm{d}y \tag{12}$$

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.

2.2.6. Posterior p(c | R). Finally, to define the posterior probability, we divide Eq. 11 by normalization factor Z:

$$p(c \mid R) = \frac{1}{Z} p(c) \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \tag{13}$$

where

$$Z = \int_{\Omega} p(c) \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \mathrm{d}\mu \tag{14}$$

Ω denotes the SCS and, as discussed above, the integration is with respect to the invariant volume $\mathrm{d}\mu$ (Eq. 12).

2.2.7. Inference using posterior. Given the posterior (Eq. 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. 1):

$$\mathrm{E}(\mathbf{q}' \mid R) = \int_{\Omega} \mathbf{q}' p(c \mid R)\ \mathrm{d}\mu \tag{15}$$

$$\mathrm{var}(\mathbf{q}' \mid R) = \int_{\Omega} (\mathbf{q}' - \mathrm{E}(\mathbf{q}'))^T (\mathbf{q}' - \mathrm{E}(\mathbf{q}'))\ p(c \mid R)\ \mathrm{d}\mu, \tag{16}$$

where $\mathbf{q}' = \mathcal{T}(c, q, \alpha)$, Ω is again the SCS and $\mathrm{d}\mu$ the invariant volume. var is the covariance matrix since $\mathbf{q}'$ 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 Section 3 for the sum of the diagonal elements in this matrix.

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 (Potluri et al., 2006) to bound the individual factors of the posterior distribution in Eq. 13. This lets us compute upper and lower bounds to the unnormalized probability density inside a cell. Since the normalization factor Z in Eq. 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.

2.3.1. SCS cell volume. The invariant volume for a cell is given by the integral of the infinitesimal invariant volume (Eq. 12). If $\cos\theta$ is positive in the range $[\theta_l, \theta_h]$, then we have:

$$\begin{aligned}\int_C \mathrm{d}\mu &= \int_C \mid \cos\theta \mid \sin\theta\ \mathrm{d}\theta\ \mathrm{d}\phi\ \mathrm{d}x\ \mathrm{d}y \\ &= [-\frac{1}{2}\cos^2\theta]_{\theta_l,\phi_l,x_l,y_l}^{\theta_h,\phi_h,x_h,y_h} \\ &= \frac{1}{2}(\cos^2\theta_l - \cos^2\theta_h)(\phi_h - \phi_l)(x_h - x_l)(y_h - y_l)\end{aligned} \tag{17}$$

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. 17. If $\cos\theta$ changes signs in this range, we split the integral accordingly and evaluate each part.

2.3.2. Upper bound on the posterior within a cell. Let us first compute an upper bound on the value of $p(c \mid R)$ (Eq. 13) for an axis c in an SCS cell C (a contiguous set of axes; see again Fig. 2b). 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. 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.

$$\forall c \in C: \; p(c \mid R) \leq \frac{1}{Z_l} \max_{c \in C} p(c) \max_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \tag{18}$$

$$Z \geq Z_l = \sum_{C \in \Omega} \min_{c \in C} p(c) \min_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \int_C \mathrm{d}\mu \tag{19}$$

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

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. 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 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.

$$\max_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \leq \left(\sum_{i=1}^{n} \min_{c \in C} \log^2 g_i(c) \right)^{-n/2} \tag{20}$$

$$\min_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \geq \left(\sum_{i=1}^{n} \max_{c \in C} \log^2 g_i(c) \right)^{-n/2} \tag{21}$$

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. 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, we can write the bounds for the log terms as:

$$\min_{c \in C} \log^2 g_i(c) = \begin{cases} 0, & \text{if } l_i(C) \leq \|\mathbf{p}_i - \mathcal{T}(c, \mathbf{q}_i, \alpha)\| \leq u_i(C) \\ \min\left(\log^2 \frac{d_i}{l_i(C)}, \log^2 \frac{d_i}{u_i(C)}\right), & \text{otherwise} \end{cases} \tag{22}$$

$$\max_{c \in C} \log^2 g_i(c) = \max_{c \in C} \left(\log^2 \frac{d_i}{l_i(C)}, \log^2 \frac{d_i}{u_i(C)} \right) \tag{23}$$

Note that we have computed lower bounds on the individual probability terms in Eq. 18. This enables us to define the lower bound on the unnormalized probability density ρ that we use in Section 2.4. The lower bound can be written as:

$$\forall c \in C: \rho(c \mid R) \geq \min_{c \in C} p(c) \min_{c \in C} \left(\sum_{i=1}^{n} \log^2 g_i(c) \right)^{-n/2} \tag{24}$$

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

2.3.3. 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. 18. The upper bound on the posterior can be written as:

$$\begin{aligned} P(C \mid R) &= \int_C p(c \mid R) \mathrm{d}\mu \\ &\leq \left(\max_{c \in C} p(c \mid R) \right) \int_C \mathrm{d}\mu \end{aligned} \tag{25}$$

2.3.4. 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. 15).

In the derivations that follow, we represent the unnormalized conditional probability density by ρ that we introduced in Section 2.3.2. Thus:

$$\rho(c \mid R) = p(c)\left(\sum_{i=1}^{n} \log^2 g(c, \mathbf{q}_i, \alpha)\right)^{-n/2} \tag{26}$$

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:

$$\delta(C, \mathbf{q}) = \left\| \frac{\int_\Omega \mathcal{T}(c, \mathbf{q}, \alpha)\rho(c \mid R)\mathrm{d}\mu}{\int_\Omega \rho(c \mid R)\mathrm{d}\mu} - \frac{\int_{\Omega\setminus C} \mathcal{T}(c, \mathbf{q}, \alpha)\rho(c \mid R)\mathrm{d}\mu}{\int_{\Omega\setminus C} \rho(c \mid R)\mathrm{d}\mu} \right\| \tag{27}$$

We can write the integrals in the first term of Eq. 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:

$$\delta(C, \mathbf{q}) \le \frac{\|\mathrm{E}(\mathcal{T}(c, \mathbf{q}, \alpha) \mid R)\| \max_{c\in C} \rho(c \mid R) \int_C \mathrm{d}\mu + \max_{c\in C} \|\mathcal{T}(c, \mathbf{q}, \alpha)\| \max_{c\in C} \rho(c \mid R) \int_C \mathrm{d}\mu}{\int_{\Omega\setminus C} \rho(c \mid R)\mathrm{d}\mu} \tag{28}$$

The algorithm we present in the next section will compute the denominator. The geometric bounds (Eq. 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:

$$\|\mathrm{E}(\mathcal{T}(c, \mathbf{q}, \alpha) \mid R)\| = \frac{\| \int_\Omega \mathcal{T}(c, \mathbf{q}, \alpha)\rho(c \mid R)\mathrm{d}\mu \|}{Z} \tag{29}$$

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 $\Omega\backslash C$. 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:

$$\left\| \int_\Omega \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\| \le \max_{c\in C}\ \|\mathcal{T}(c, \mathbf{q}, \alpha)\| \max_{c\in C} \rho(c \mid R) \int_C \mathrm{d}\mu + \left\| \int_{\Omega\setminus C} \mathcal{T}(c, \mathbf{q}, \alpha)\ \rho(c \mid R)\ \mathrm{d}\mu \right\| \tag{30}$$

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. 28 gives us the final inequality for the error in expectation:

$$\begin{aligned}\delta(C, \mathbf{q}) \le \frac{1}{\int_{\Omega\setminus C} \rho(c \mid R)\mathrm{d}\mu} \Bigg(&\max_{c\in C}\ \|\mathcal{T}(c, \mathbf{q}, \alpha)\| \max_{c\in C} \rho(c \mid R) \int_C \mathrm{d}\mu + \frac{\max_{c\in C} \rho(c \mid R) \int_C \mathrm{d}\mu}{Z_l} \\ &\cdot \left(\max_{c\in C}\ \|\mathcal{T}(c, \mathbf{q}, \alpha)\| \max_{c\in C} \rho(c \mid R) \int_C \mathrm{d}\mu + \left\| \int_{\Omega\setminus C} \mathcal{T}(c, \mathbf{q}, \alpha)\rho(c \mid R)\mathrm{d}\mu \right\| \right) \Bigg)\end{aligned} \tag{31}$$

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 $C \in \mathcal{C}$:

$$\begin{aligned}\left\| \int_{\mathcal{C}} \mathcal{T}(c, \mathbf{q}, \alpha)\rho(c \mid R)\mathrm{d}\mu \right\| &= \left\| \sum_{C\in\mathcal{C}} \int_C \mathcal{T}(c, \mathbf{q}, \alpha)\rho(c \mid R)\mathrm{d}\mu \right\| \\ &\le \sum_{C\in\mathcal{C}} \left\| \int_C \mathcal{T}(c, \mathbf{q}, \alpha)\rho(c \mid R)\mathrm{d}\mu \right\|\end{aligned} \tag{32}$$

The upper bound in Eq. 28 can be rewritten in terms of the individual cells by using Eq. 32.

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 Figure 3, and pseudocode is provided in Algorithm 1. While we also used hierarchical subdivision in our earlier approach (Potluri et al., 2006), 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 (Potluri et al., 2006). 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.

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: $\zeta_0$: maximum pruned probability mass
Output: $P$: posterior distribution, a set of (cell, posterior) pairs
  $P \leftarrow \emptyset$
  $\mathcal{C} \leftarrow \mathcal{C}_0$  // cells for the next level
  $\zeta \leftarrow \zeta_0$  // remaining allowed error
  $Z_l \leftarrow$ lower bound on $Z$ for $\mathcal{C}_0$  // Eq. 19
  while $\mathcal{C}$ is not empty do  // expand the next level
    $V \leftarrow \sum_{C \in \mathcal{C}} \int_C \mathrm{d}\mu$  // invariant volume, Eq. 17
    $\mathcal{C}' \leftarrow \emptyset$  // cells for the next level
    for $C \in \mathcal{C}$ do
      $u \leftarrow$ upper bound on $P(C \mid R)$ using current $Z_l$  // Eq. 25, Eq. 18
      if $u < (\zeta/V) \int_C \mathbf{d}\mu$ then  // prune cell
        $\zeta \leftarrow \zeta - u$
      else if $C$ is small enough then  // accept cell
        $p \leftarrow \rho(c \mid R) \int_C \mathrm{d}\mu$ for the centroid $c$ of $C$  // unnormalized Eq. 13
        add to $P$ the pair $(C, p)$
      else
        subdivide $C$ into $C_1$ and $C_2$
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \{C_1, C_2\}$
        // Update $Z_l$ for subdivision, using $\rho_{\min}$ from Eq. 24
        $Z_l \leftarrow Z_l - \rho_{\min}(C \mid R) \int_C \mathrm{d}\mu + \rho_{\min}(C_1 \mid R) \int_{C_1} \mathrm{d}\mu + \rho_{\min}(C_2 \mid R) \int_{C_2} \mathrm{d}\mu$
      end if
    end for
end while
```

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. 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. 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. 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 τ_0 (e.g., 1 Å) (Potluri et al., 2006).

To subdivide a cell, we split one of the dimensions in half, employing the heuristic from our earlier method (Potluri et al., 2006). 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 (Section 2.3.4). 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 (Dean and Ghemawat, 2004), which provides a framework for parallelizing the code, taking care of machine failure, scheduling jobs, and partitioning the data.

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 *Escherichia coli* MinE (King et al., 2000) has 50 residues per subunit with 183 inter-subunit NOE restraints, the homo-trimeric coiled-coil domain of chicken cartilage matrix protein (CCMP) (Wiltscheck et al., 1997) has 47 residues per subunit and 49 inter-subunit NOE restraints and a transmembrane peptide of Glycophorin A (GpA transmembrane peptide) (MacKenzie et al., 1996) 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) (Berman et al., 2000)—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) (Seavey et al., 1991). The restraints are fairly well-dispersed in the structures (Fig. 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.

3.1. 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 dμ, and thus independent of the coordinate frame. The top row of Figure 5 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

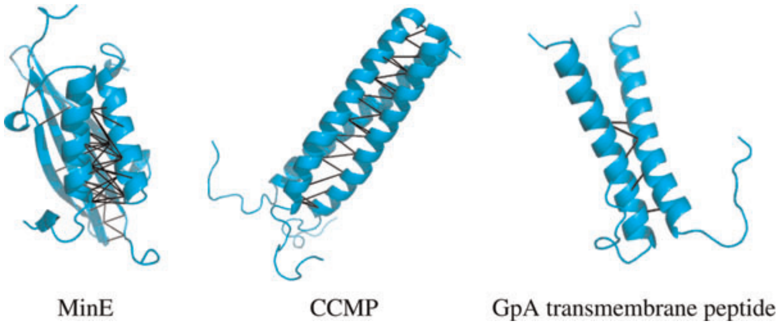


FIG. 4. Reference structures (cyan) and inter-subunit distance restraints (black) for MinE (183 restraints), CCMP (49), and GpA transmembrane peptide (6). For CCMP, restraints are only shown between chains A and B, to avoid clutter.

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 root-mean-square deviation (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. 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 Figure 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 generally acceptable variation in conformations (as represented by axes).

We also compared our results against those obtained from our earlier ‘‘binary’’ approach (Potluri et al., 2006), 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. Figure 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. 6, middle). The maximum backbone RMSD is between the backbone C^{α}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

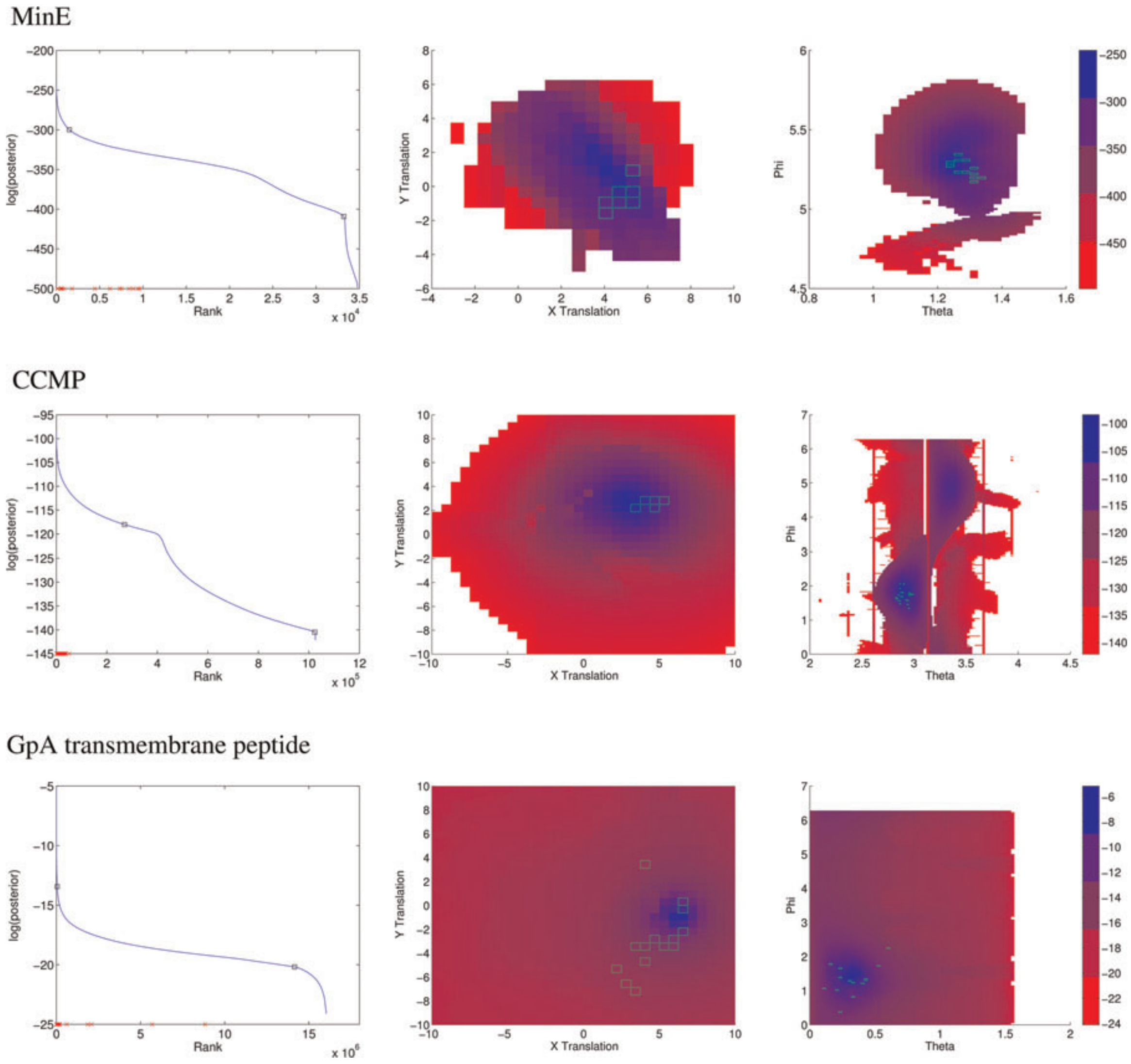


FIG. 5. Posterior distributions. **(Left)** Unnormalized log posterior for accepted cells. Red points on the x-axis indicate posteriors computed for members of the reference ensembles. **(Right)** Projections of SCS onto translation and orientation components, colored by posterior (different scales for 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).

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. Figure 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

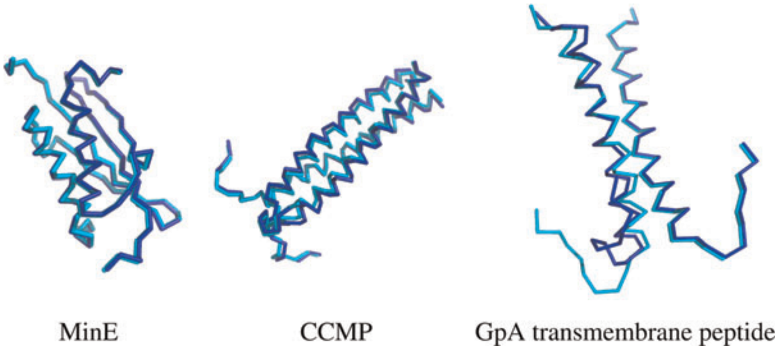


FIG. 6. MAP structures (cyan) superimposed with closest member of reference ensembles (blue).

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 Figure 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.

3.2. *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.

Figure 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. 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 there. Finally, Figure 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. 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 Figure 4 (right). Therefore, this part of the helix in the second subunit shows the least variance.

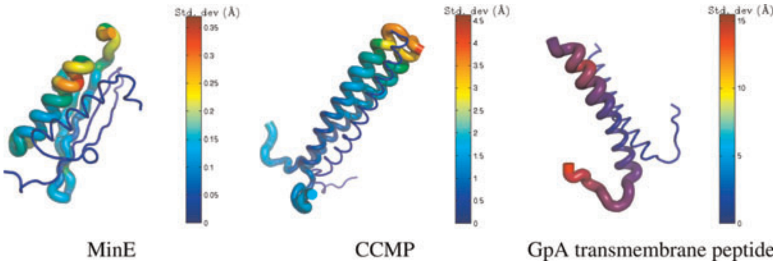


FIG. 7. Inferred means and standard deviations in atomic coordinates, represented as sausage plots. The fixed subunit is shown in blue with a zero standard deviation. The color and thickness of the adjacent subunit represent the standard deviation in the positions of the backbone atoms. Note that the standard deviations are on different scales for different proteins.

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 (≈50%, 25%, and 20%), randomly choosing the restraints from the entire set. We generated five such datasets for each number of restraints. Similarly, for CCMP we generated random subsets of sizes 27 and 16 from the original 49 (≈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.

Table 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. Figure 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 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.

Figure 9 plots the mean atomic variances for the C$^\alpha$ 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 Figure 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. 7, middle).

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

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

Table 1. Effects of Missing Restraints on Inference

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

Pruned%, percentage of SCS volume pruned; RMSD$_1$, reduced-restraint MAP versus full-restaint MAP; RMSD$_2$, reduced-restraint MAP versus reference ensemble; RMSD$_3$, reduced-restraint mean versus full-restraint mean; RMSD$_4$, reduced-restraint mean versus reference ensemble. All RMSDs are computed with backbone atoms. The RMSD to closest structure in reference ensemble is shown.

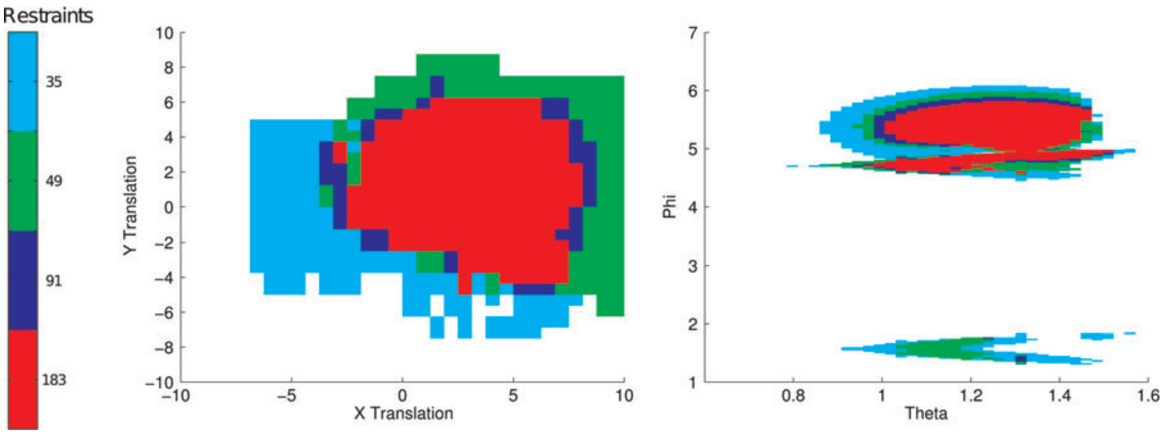


FIG. 8. Translation and orientation parameters of accepted MinE cells for different subsets of restraints (one dataset for each number of restraints). Colored cells are those eliminated with more restraints but not with fewer restraints.

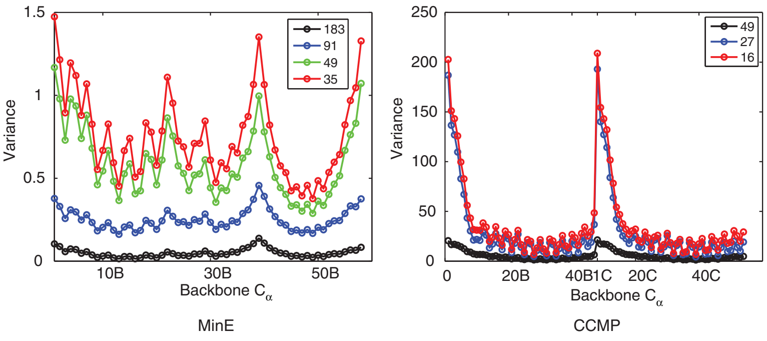
uncertainty in distance estimation, algorithms must also be able to handle violations. We simulated noise in a manner that reflects realistic systematic structural variation and uncertainty, instead of simply adding random noise. In addition to the representative structure whose first chain we used as 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,

FIG. 9. Mean variance (y-axis, Å^2) in C^α 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.



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 out-number them. Hence, the noisy restraints do not impact the eventual posterior distribution in CCMP to the extent observed for MinE.

Our original binary algorithm (Potluri et al., 2006) 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) (Potluri et al., 2007). 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.

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 conformations. 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 (Potluri et al., 2006), 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 incorporating 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) (Desmet et al., 1992), which loses the global minimum energy guarantee of DEE, to minimized-DEE (Georgiev et al., 2008), which accounts for possible energy minimization when considering pruning and thus regains the provable guarantee.

ACKNOWLEDGMENTS

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

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Berman, H.M., Westbrook, J., Feng, Z., et al. 2000. The Protein Data Bank. *Nuclic Acids Res.* 28, 235–242.

Brünger, A.T. 1993. *XPLOR: A System for X-Ray Crystallography and NMR*. Yale University Press, New Havern, CT.

Dean, J., and Ghemawat, S. 2004. MapReduce: simplified data processing on large clusters. Proc. *OSDI'04: Sixth Symp. Operating System Design Implement*.

Desmet, J., De Maeyer, M., Hazes, B., et al. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356, 539–542.

Georgiev, I., Lilien, R.H., and Donald, B.R. 2008. 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.

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

Güntert, P., Braun, W., and Wüthrich, K. 1991. 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.

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

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

Habeck, M., Rieping, W., and Nilges, M. 2006. Weighting of experimental evidence in macromolecular structure determination. *Proc. Natl. Acad. Sci. USA* 103, 1756–1761.

Ikura, M., and Bax, A. 1992. 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.

Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A* 186, 453–461.

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

King, G.F., Shih, Y.L., Maciejewski, M.W., et al. 2000. Structural basis for the topological specificity function of MinE. *Nat. Struct. Biol.* 7, 1013–1017.

Lee, W., Revington, M.J., Arrowsmith, C., et al. 1994. A pulsed-field gradient isotope-filtered 3D C-13 HMQC-NOESY experiment for extracting intermolecular NOE contacts in molecular-complexes. *FEBS Lett.* 350, 87–90.

Lipari, G., and Szabo, A. 1982. 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.

MacKenzie, K.R., Prestegard, J.H., and Engelman, D.M. 1996. 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.

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

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

Nilges, M., Habeck, M., O'Donoghue, S.L., et al. 2006. Error distribution derived NOE distance restraints. *Proteins* 64, 652–664.

Oxenoid, K., and Chou, J.J. 2005. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc. Natl. Acad. Sci. USA* 102, 10870–10875.

Potluri, S., Yan, A.K., Chou, J.J., et al. 2006. 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.

Potluri, S., Yan, A.K., Donald, B.R., et al. 2007. A complete algorithm to resolve ambiguity for inter-subunit NOE assignment in structure determination of symmetric homo-oligomers. *Protein Sci.* 16, 69–81.

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

Rieping, W., Michael, H., and Nilges, M. 2005b. Inferential structure determination. *Science* 309, 303–306.

Santaló, L.A. 2002. *Integral Geometry and Geometric Probability*, 2nd ed. Cambridge University Press, New York.

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

Seavey, B.R., Farr, E.A., Westler, W.M., et al. 1991. A relational database for sequence-specific protein NMR data. *J. Biomol. NMR* 1, 217–236.

Walters, K.J., Matsuo, H., and Wagner, G. 1997. A simple method to distinguish intermonomer nuclear overhauser effects in homodimeric proteins with C2 symmetry. *J. Am. Chem. Soc.* 119, 5958–5959.

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

Wiltscheck, R., Kammerer, R.A., Dames, S.A., et al. 1997. 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.

Zwahlen, C., Legaulte, P., Vincent, S.J.F., et al. 1997. 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.

Address correspondence to:
Dr. Chris Bailey-Kellogg
Department of Computer Science
Dartmouth College
6211 Sudikoff Laboratory
Hanover, NH 03755

E-mail: cbk@cs.dartmouth.edu.