Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data

Simon Olsson¹*, Jes Frøllsen¹, Wouter Boomsma², Kanti V. Mardia³, Thomas Hamelryck¹*

¹Bioinformatics Centre, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark, ²Structural Biology and NMR Laboratory, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark, ³Department of Statistics, School of Mathematics, University of Leeds, Leeds, United Kingdom

Abstract

We present the theoretical foundations of a general principle to infer structure ensembles of flexible biomolecules from spatially and temporally averaged data obtained in biophysical experiments. The central idea is to compute the Kullback-Leibler optimal modification of a given prior distribution $\pi(x)$ with respect to the experimental data and its uncertainty. This principle generalizes the successful inferential structure determination method and recently proposed maximum entropy methods. Tractability of the protocol is demonstrated through the analysis of simulated nuclear magnetic resonance spectroscopy data of a small peptide.

Introduction

The rigorous analysis of experimental data probing the structure of biological macromolecules forms the foundation of many biophysical studies [1]. The sources of experimental data include nuclear magnetic resonance spectroscopy spectroscopy (NMR) and small-angle X-ray- and neutron scattering. This article addresses several issues which often make inference of biomolecular structure from such data particularly challenging. First, in these experiments, the time-scale of acquisition typically exceeds that of molecular fluctuations. Second, the samples studied are often near molar concentrations. Third, data is frequently incomplete, or even sparse, and subject to experimental noise. Consequently, data obtained from such techniques yield incomplete, noisy, spatially and temporally averaged information on the Boltzmann ensemble of the observed system. Thus, such data are ideally analyzed through models that take these properties into account. While this fact has long been recognized, the analysis of these types of data has revolved predominantly around structure determination — that is, fitting a single conformation to fulfill all derived geometrical restraints [2]. Such structure determination methods do not adequately handle sparse, noisy and averaged data. Here, we propose an alternative method which addresses these shortcomings.

Typically, structure determination from experimental data proceeds through hybrid energy minimization [3]. In this method, an energy function $E_{\text{exp}}$ that brings in the experimental data is combined with an approximate physical forcefield $E_{\text{phys}}$. The term $E_{\text{exp}}$ is typically a straightforward combination of a forward- and an error-model. A forward-model relates a protein conformation to experimental data, whereas an error-model concerns experimental errors. Alternatively, a Bayesian formulation known as inferential structure determination (ISD) has been proposed, formulating structure determination in a rigorous probabilistic framework [4]. In ISD, a prior distribution is constructed by combining a data likelihood with prior distributions on conformational and nuisance parameters. The likelihood and the prior concerning biomolecular structure correspond to $E_{\text{exp}}$ and $E_{\text{phys}}$, respectively. This Bayesian approach extends the common hybrid energy minimization by solving the important problems of choosing appropriate error-models, treating model-parameters coherently and performing inference through posterior sampling rather than minimization. However, by construction, these approaches assume that conformational variability can be represented through uncorrelated, homoscedastic fluctuations around an average structural representation. Consequently, the conformational heterogeneity present in the posterior distribution reflects the quality and completeness of the experimental data and the prior distributions, but not necessarily any physical fluctuations [5]. Despite this well-known limitation, the approximation tends to yield good results for well-folded proteins when conformational fluctuations are modest.

Early attempts to model ensemble NMR data involved averaging along molecular dynamics trajectories [6,7]. In these protocols, a memory function specifies an averaging time-span which is used to obtain a time-averaged representation of the experimental data. While this approach displayed initial promise, the short timescales accessible through routine molecular dynamics limit its use [8]. An alternative approach, which involves explaining the data using an average of several conformations,
emerged around the same time [9]. This approach has since shown to be more viable.

During the past two decades there has been an increasing interest in biomolecules that undergo significant conformational fluctuations, such as natively unfolded and partially unfolded proteins [10]. Consequently, there have been many efforts to overcome the limitations of structure determination procedures with respect to the flexibility of these molecular systems. Prevalently, conformational fluctuations are represented by finite *ensembles*: the data is explained by a weighted average of $N > 1$ conformations, introduced above. In effect, this corresponds to discretizing the Boltzmann ensemble. Such discrete ensembles may be constructed in a multitude of ways, including database-derived explicit ensembles [11,12], data-optimized explicit ensembles [13–17], fragment based ensemble construction [18–20] and multi-conformer refinement, molecular dynamics and Monte Carlo methods [21–23] and maximum entropy methods [24]. Another important approach uses multiple replicas in the calculation of the hybrid energy used in restrained molecular simulations [8,25,26]. However, the discretization of the conformational ensemble is inherently problematic because determining the optimal ensemble size $N$, and its associated uncertainty, is difficult.

Restraining simulations using an average of multiple replicas is a sensible solution, as it was recently shown that multiple replica restrained simulations constitute the least biased method when the number of replicas goes to infinity in the absence of experimental noise [27–29]. However, a measurable bias is introduced when the number of replicas used is too small [28]. Since the use of large numbers of replicas may prove to be computationally intractable or impossible, the development of approaches which are independent of this discretization is highly desirable.

In this work, we approach the problem of modeling sparse, spatially and temporally averaged data through the principles of Bayesian statistics and information theory. Unlike the previous Bayesian efforts [4,16], we explicitly take into account the experimental data as noisy, average quantities of an underlying heterogeneous ensemble in continuous space. We derive a general posterior distribution from first principles which imposes the least necessary bias on our prior knowledge to fulfill the experimental data.

We outline a number of general, theoretical advances concerning biomolecular structure determination and restrained molecular simulations. To ensure a focused and concise presentation, we limited the number of practical examples. However, one example given uses synthetic data of a small idealized peptide GB1 generated using the PROFASI forcefield at high temperature [30]. This choice allows us to carefully evaluate the theory presented by avoiding confounding variables. Finally, our findings are compared to existing methodology and is shown to generalize these.

**Results and Discussion**

A hierarchical model of spatially and temporally averaged restraints

Ultimately, our aim is to sample from the conditional probability distribution $p(x|d)$, where $x$ denotes a protein's conformation and $d$ denotes spatially and temporally averaged, experimental data. The variable $x$ represents a positional microstate in atomic detail. Through a forward model $f(x)$ we can calculate a coarse-grained representation, $f$, of a protein conformation $x$. That is, our forward model is a mapping, $f : \mathbb{R}^{3N} \rightarrow \mathbb{R}^M$, of the $N$ atoms of $x$ to an $M < 3N$-dimensional coarse grained representation, $f$. Conceptually, $f$ may be interpreted as the instantaneous 'experimental data' back-calculated from a positional micro-state, $x$. However, as $d$ represents an averaged quantity we need to introduce a variable, $e$, to represent an *ensemble average* of the simulated experimental data $f$. Consequently, our full posterior distribution becomes $p(x, f, e|d)$.

We clarify the relation between $f, e$ and $d$ using the example we will present later on. In the case of nuclear Overhauser enhancement (NOE) data obtained from an NMR experiment [31], the coarse-grained variable $f$ is a vector related to pairwise distances between atoms in a protein conformation $x$. In one case, this is simply a vector of these distances. The variable $e$ is an average of $f$ vectors from an ensemble of protein conformations. The experimental NOE data $d$ can be interpreted as a noisy observation of the vector of averages, $e$. In general, there is no simple relationship between the vector $f$ and the averaged vector $e$, but a simple probabilistic model that relates them can be developed, as we discuss next.

We start by considering the coarse-grained representations of the distribution, $f$, $e$ and $d$, without considering the fine-grained representation, $x$. Following the Bayesian probability calculus, we formulate a posterior distribution,

$$p(f,e|d) \propto p(d|f,e)p(f,e) = p(d|e)p(f|e),$$

with the first term the likelihood and the second term is the prior distribution. Note that the prior of $d$ in the evaluation of the likelihood function – $d$ is a noisy observation of $e$, which does not involve $f$. The independence assumptions of the model are shown in the corresponding graphical model in Figure 1.

Applying the product rule of probability theory to equation (1), we obtain

$$p(f,e|d) \propto p(d|e)p(f|e)p(e),$$

where $p(f|e)$ is the prior distribution of the simulated data $f$ given their averaged value $e$, and $p(e)$ is the prior distribution over the simulated ensemble averaged data $e$.

Equation (2) is a probabilistic model of the relationship between noisy, ensemble averaged data, and conformational micro-states in a coarse-grained space. However, to obtain a probability distribu-

**Figure 1. A directed graphical model of the ensemble model (on the left) and its interplay with a fine-grained conformational prior distribution (top right) through the reference ratio method, (bottom right).** In the graphical model, the black circles are random variables, and the arrows determine their conditional independencies. The parameter $\sigma$, marked in grey on the left, is fixed and given, and denotes the experimental error in this particular example. $\tau(f)$ denotes the reference distribution.

doi:10.1371/journal.pone.0079439.g001
We use the logarithmic distribution as an appropriate error model for pairwise distances derived from NOEs, which is the approach adopted by ISD [39]. The choice for the prior \( \pi_f(f) \) is less obvious and depends on the type of experimental data. Here, we use the exponential distribution with mean \( \beta \), \( \mathcal{E}(\beta) = e^{-\beta} \), since it constitutes the least biasing continuous distribution on the positive real axis, when no higher order moments are observed.

Reconstructing a high temperature ensemble from sparse data

To test the presented theory, we use synthetic NOE data, obtained from an ensemble of the GB1 hairpin simulated at 400K in the Profasi forcefield [30]. This simple, idealized system was chosen to minimize the chances of undersampling, as well as to avoid confounding associated with experimental data.

The restraints used here are visualized on a random conformation of the GB1 hairpin in Figure 2. Historically, NOEs constitute one of the most important sources of semi-quantitative information in NMR structure determination. Under the isolated spin-pair approximation for rigid molecules, NOEs are related to an interatomic distance \( r \) as \( \text{NOE} \propto \langle r^{-6} \rangle \). As an example, we will apply Equation (2) to two cases of averaged pairwise distance data – these two cases involve the arithmetic mean \( \langle r \rangle \), and the power-averaged mean \( \langle r^{-6} \rangle \), respectively. They represent two different averaging processes that are common in biophysical data.

We use the log-normal distribution as an appropriate error model for pairwise distances derived from NOEs, which is the approach adopted by ISD [39]. The choice for the prior \( \pi_f(f) \) is less obvious and depends on the type of experimental data. Here, we use the exponential distribution with mean \( \beta \), \( \mathcal{E}(\beta) = e^{-\beta} \), since it constitutes the least biasing continuous distribution on the positive real axis, when no higher order moments are observed.

In minus log space Equation 6 is proportional to minus \( f^T \Lambda d \). This corresponds to the empirical term of the previously reported maximum entropy method in absence of experimental uncertainty [27]. We note that this method does not explicitly account for the reference distribution \( \pi_f(f) \) when combining the empirical term in the coarse-grained space with a fine-grained prior distribution, \( \tau(x) \). However, if the prior \( \pi_f(f) \) is appropriate, then the Lagrange multipliers \( \Lambda \) may provide the necessary means for minor adjustments.

Figure 2. A random backbone conformation of the GB1 hairpin. The restraints listed in Table 1 are shown as dashed lines. The distance shown in red is used as the reaction coordinate \( f_0 \) used in Figures 3 and 4. This figure was created using PyMOL (DeLano Scientific LLC). doi:10.1371/journal.pone.0079439.g002

The relationship to other methods

The model given by Equation (3) may be reduced to the ISD framework [4]

\[
p(x, f, d) \propto p(d|e)p_f(f)\pi_e(e) \pi_s(s) \pi_d(d) \tau(x).
\]
Inference of Biomolecular Structure Ensembles

<table>
<thead>
<tr>
<th>Table 1. Synthetic datasets used in this study.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cα-pair</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>41–51</td>
</tr>
<tr>
<td>42–48</td>
</tr>
<tr>
<td>44–46</td>
</tr>
<tr>
<td>44–54</td>
</tr>
<tr>
<td>53–54</td>
</tr>
</tbody>
</table>

First column: Cα atoms involved in the pairwise distance. Second and last columns: averaged power-averaged pairwise distances, respectively. doi:10.1371/journal.pone.0079439.g001

[39]. The prior on $f$ thus becomes: $p\gamma(f|e) = p\gamma(f|w) = \Pi_i \mathcal{N}(f_i|e_i, w_i)$, where the product runs over all data-points and the scale vector $w$ is a free parameter (discussed below). It follows that

$$p(x,f,e|d) \propto \mathcal{N}(\ln d|\ln e, \sigma) p\gamma(f|e) p\gamma(e) \tau_f(x),$$

(7)

where $\sigma$ is the experimental error, which is fixed and given, and $\mathcal{N}(\cdot)$ is the normal distribution. As prior on the ensemble average $e$ we choose $p\gamma(e) \propto e^{-1}$. This prior has previously been shown to provide good results for variables confined to the positive real axis [40].

Equation (7) provides a general solution to the problem of modeling averaged NOE data subject to experimental uncertainty. The only parameter to be estimated is the scale vector $w$, which relates $f$ to $e$ in $p\gamma(f|e, w)$.

In the ideal case, an optimal choice for $w$ results in the desired distribution for $f$ as calculated from the structures $x$. More precisely, it results in a marginal posterior distribution of Equation (7) for $x$, such that, when $e$ is fixed, the expectation of $f$ is equal to $e$. In practice, a satisfactory point estimate for $w$ can be obtained in an iterative manner, using an empirical Bayes approach (see Materials and Methods). The parameter $w$ compensates for the approximate nature the reference distribution $\tau_f(\cdot)$, which is difficult to estimate accurately [33]. The introduction of $w$ provides a simple, yet effective measure to compensate for this.

We use equation (7) to model synthetic pairwise distance restraints in the GB1 hairpin. For $\tau(x)$ we use probabilistic models of the conformational space of the main chain [41] and the side chains [42], as these models recently yielded excellent results when combined with the ISD method [43]. As the prior distribution and the likelihood concern local and nonlocal features of protein structure, respectively, their information content shows little overlap. More informative priors, for example based on physical energy functions, can be envisaged, but this is beyond the scope of this article.

Evaluation of inferred ensembles

The prior distribution used in this study concerns protein structure on a local length scale, and thus does not model long range distances accurately. Consequently, as a reaction coordinate, we chose a representative distance $f_0$ between atoms C$\alpha$-$C_{\gamma1}$ and C$\alpha$-$C_{\gamma1}$ which are separated farthest in sequence – to illustrate the long-range properties of the eight different ensembles considered here. Histograms of this pair-wise distance in the different ensembles are shown in Figures 3 and 4. This pair-wise distance is highlighted with yellow color in Figure 2.

The conformational prior and PROFASI, which was used to generate the averaged data, result in different distance distributions (Figure 3). However, if we modify the prior using the reference ratio method as described above, we obtain good fits with the PROFASI distribution for both linearly and power-averaged data. The ISD ensemble, which does not take the ensemble nature of the data into account, is overly tightly peaked around the (correct) mean.

A similar pattern is observed for the distribution of the gyration radii $R_g$. The gyration radii are not used in the estimation of the probability distributions, and can thus be used for cross-validation. The average and standard deviation of the gyration radii of the ensemble used to generate the data is 9.71 ± 1.5 Å. The ensembles obtained with our method from the power averaged and linearly averaged data resulted in a slightly higher average (10.30 ± 1.8 Å and 10.19 ± 1.6 Å, respectively), but essentially the correct standard deviation. This is an excellent result, as a perfect fit is not expected due to the sparse and noisy nature of the data. Again, the ISD ensemble provides an overly narrow distribution (9.88 ± 0.6 Å). Finally, sampling from the prior distribution alone results in an average radius of gyration of 11.34 ± 2.8 Å, which is considerably too high.

In some cases, compensating for the bias introduced by the reference distribution is not critical to obtain good results. As it constitutes an additional obstacle in terms of estimation and simulation time, we evaluate its significance on the obtained results. In the power averaged case we achieved this by choosing the reference distribution $\tau_f(\cdot)$ and the scale vector, $w$, in equation (7) to be the uniform distribution and the unit vector, respectively. In the linearly averaged case, the scale vector was kept fixed equal to the 1-vector, as $\tau_f(\cdot)$ was assumed to be uniform in the results presented above. The results are shown in Figure 4. In the case of the power averaged data, with $\tau_f(\cdot)$ uniform and $w$ equal to the 1-vector, severely skews the distribution of the distances (green line in figure 4). When the scale vector $w$ is estimated, while still assuming $\tau_f(\cdot)$ uniform, the fit improves (blue line), but without resulting in a satisfactory distribution. In the linearly averaged case, we find that a 1-vector for $w$ provides a good fit (red line).

Figure 3. Histograms, $p(f_0)$, of a representative pairwise distance $f_0$ (between C$\alpha$-$C_{\gamma1}$ in Å) in the ensembles. The black and blue lines are obtained from the PROFASI and ISD ensembles respectively, while the cyan line represent the prior $\tau(x)$. Finally, the green and red lines respectively represent ensembles obtained from the power-averaged and linearly averaged data. doi:10.1371/journal.pone.0079439.g003
Inference of Biomolecular Structure Ensembles

If we again consider the gyration radii as providing complementary views of the ensembles, we find that the power-averaged ensemble with uniform \( \tau(f) \) and unit scale vector yields an overly extended ensemble, 11.94 ± 1.63Å. The results in the linearly averaged case compare to those with an estimated scale-vector, 10.34 ± 1.69Å, presented above.

To summarize, in the power averaged case, both \( \tau(f) \) and \( w \) are required for a satisfactory distribution. In the case of the linearly averaged data, our results suggest that \( w \) and \( \tau(f) \) may be approximated by the l-vector and uniform distribution, respectively. This is particularly interesting as it may be a general feature of applying other kinds of linearly averaged data. This may make the use of these types of data for restraining easier.

Conclusions

In conclusion, we present the theoretical foundations of a Bayesian principle to infer ensembles of protein structures from noisy experimental data subject to ensemble and time averaging. We demonstrate the principle constitutes a generalization of ISD and previously proposed maximum entropy restraining approaches. Finally, the principle is successfully evaluated using synthetic experimental data of a small idealized system.

Our approach combines a coarse-grained Bayesian model of the data with a fine-grained model of protein conformational space. The combination is accomplished using the reference ratio method [33], which corresponds to a maximum entropy solution in the presence of experimental noise. The role of the reference distribution \( \tau(f) \) is considerable. When we assumed \( \tau(f) \) to be uniform, we were unable to construct sufficiently accurate distribution of pairwise distance geometry, in the case of power-averaged data.

The Bayesian model may in principle be applied to denser and/or ambiguous [44] datasets and to data from other sources such as small angle X-ray- or neutron scattering, or other NMR experiments. Also, low-resolution data may be combined with more sophisticated physical prior distributions such as those embodied in force fields. The presented method is thus a general method to obtain physically sound ensemble models of solution and endogenous states of biomolecules, given appropriate experimental data. Practical implementation of protocols for other data sources and larger systems clearly is necessary. Possible issues arising with this methodology include insufficient sampling of the conformational space and difficult estimation procedures for reference distributions and scale parameters alike. However, the work presented herein provides the guiding principles for these future developments.

Materials and Methods

Synthetic datasets

A synthetic dataset was created for the GB1 hairpin [Protein data bank identifier: 1LE3; sequence variant [Y45W, F52W, V54W]]. The data were generated by simulating the protein at 400K with the PROFASI forcefield [30], using Engh-Huber parameters for bond-angles and bond-lengths [45]. The high temperature was used to emulate the effect of a denatured, disordered state. A total of 3.5×10⁶ steps were performed using the Metropolis-Hastings algorithm in the PHAISTOS Markov chain Monte Carlo framework (http://www.phaistos.org). We used a Monte Carlo move set previously described [43]. Samples were saved in intervals of 5000 steps. These samples were used to form five non-redundant, averaged \( C\alpha – C\alpha \) distance restraints (see Table 1).

To mimic the effect of distance averaging in a dipolar interaction undergoing fast motion compared to the cross-relaxation but slow motion when compared to molecular tumbling, we calculated a power averaged variant of the dataset as \( I = \langle r^{-6} \rangle \), where \( r \) is an inter-atomic distance and the angular-brackets denote ensemble averaging [46]. We used an experimental uncertainty for the power averaged dataset \( \sigma_\ell \) of the same relative amplitude as for the average restraint set \( \sigma_\ell \), by enforcing the signal-to-noise ratio to be constant. Hence, \( \frac{\tilde{\sigma}}{\sigma_\ell} = \frac{7}{\tilde{\sigma}} \rightarrow \sigma_\ell = \frac{7}{\tilde{\sigma}} \) as \( \sigma_\ell = 1 \), where \( \tilde{\sigma} \) and \( \tilde{\sigma} \) denote the datapoints corresponding to the largest average distance in the power and linearly averaged datasets, respectively. Noise with standard-deviation \( \sigma_\ell \) was added to the power-averaged data.

Estimation of \( p(x,e,f|d) \) and scale vector \( w \)

This section describes the estimation of the reference distribution \( \tau(f) \) and the vector \( w \) needed for the posterior distribution:

\[
p(x,e,f|d) \propto \frac{\mathcal{N}(\ln d|\ln e,\sigma)\tau(f|e,w)\pi(e)}{\tau(f)} \tau(x). \tag{8}
\]

In the case of the power-averaged data, the reference distribution \( \tau(f) \) was approximated by a product of exponential distributions:

\[
\tau(f) \propto \prod \mathcal{E}(^{-\tilde{f}_i}/b_i). \tag{9}
\]

The mean \( \beta \) was estimated using a Monte Carlo scheme similar to that used to form the synthetic datasets, but only using the prior \( \tau(X) \), consisting of the probabilistic models TorusDBN [41] and Basilisk [42] along with a simple binary term assuring atoms do not overlap [47]. The coarse graining, \( \tilde{f} \), was the inverse pairwise distance between the \( C\alpha \) atoms listed in Table 1. For the linearly-averaged data, \( \tau(f) \) was approximated by a uniform distribution.
We obtain a point estimate of $\mathbf{w}$ following an empirical Bayes approach. We start by initializing all the elements of $\mathbf{w}$ to unity. Subsequently, we sample an ensemble according to Equation (6), and update $\mathbf{w}$ based on the sampled values of $\mathbf{f}$ and $\mathbf{e}$. To update $\mathbf{w}$ we make use of the moment estimator for the mean of the exponential distribution:

$$w_{n+1 j} = w_{n j} \frac{\bar{f}_j}{\bar{s}_n}$$

$\bar{f}$ and $\bar{e}$ are posterior expectations of the coarse-grained variable and the ensemble averages using scale vector $\mathbf{w}_n$, respectively, and $\mathbf{w}_{n+1}$ is the updated scale vector. This procedure is repeated until convergence. Convergence was assumed when fluctuations in $\bar{f}$ were within the experimental uncertainty. Each step in the algorithm runs for $2.5 \times 10^5$ MCMC steps, and a final production ensemble is produced using $2.5 \times 10^6$ MCMC steps.

**Sampling of $\mathbf{e}$**

To sample $\mathbf{e}$ from the prior $p^{-1}$ we sampled a factor $\mathbf{A}$ from a log-normal distribution $\mathcal{A} \sim \exp([0\mathbf{N}(0,\mathbf{\sigma})]$, where $\mathbf{\sigma}$ has the same order of magnitude as the experimental uncertainty. A change from $\mathbf{e}$ to $\mathbf{e} + \mathbf{A}$ was accepted according to the Metropolis acceptance probability $\pi$:

$$\pi(\Delta \mathbf{e}) = \min\left(1, \frac{p(\mathbf{e} + \mathbf{A})}{p(\mathbf{e})}\right),$$

where $p(\mathbf{e}) = p(\mathbf{f}[\mathbf{x},\mathbf{d}]) \pi_e(\mathbf{e})$.

**Acknowledgments**

We thank Kresten Lindorff-Larsen and Jesper Feringhoff-Borg for valuable comments, discussions and suggestions.

**Author Contributions**

Conceived and designed the experiments: SO TH. Performed the experiments: SO TH. Analyzed the data: SO TH. MF. Contributed reagents/materials/analysis tools: WB KM JF. Wrote the paper: SO TH.

**References**


