OPEN ACCESS Freely available online

PLOS ONE

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

Simon Olsson[1]*, Jes Frellsen[1], Wouter Boomsma[2], Kanti V. Mardia[3], Thomas Hamelryck[1]*

1 Bioinformatics Centre, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark, **2** Structural Biology and NMR Laboratory, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark, **3** Department 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 $\tau(\mathbf{x})$ 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.

Citation: Olsson S, Frellsen J, Boomsma W, Mardia KV, Hamelryck T (2013) Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data. PLoS ONE 8(11): e79439. doi:10.1371/journal.pone.0079439

Editor: Narcis Fernandez-Fuentes, Aberystwyth University, United Kingdom

Received March 27, 2013; **Accepted** September 24, 2013; **Published** November 7, 2013

Funding: SO and JF acknowledge funding from the Danish Council for Independent Research. http://fivu.dk/en/research-and-innovation/councils-and-commissions/the-danish-council-for-independent-research/the-council-1/the-danish-council-for-independent-research-technology-and-production-sciences?set_language=en&cl=en (Technology and Production Sciences, FTP, 274-09-0184). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: solsson@binf.ku.dk (SO); thamelry@binf.ku.dk (TH)

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_{exp} that brings in the experimental data is combined with an approximative physical forcefield E_{phys}. The term E_{exp} is typically a straight-forward 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 posterior 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_{exp} and E_{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 one *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 weighed 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 [8,21–23] 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 heterogenous 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 $p(\mathbf{x}|\mathbf{d})$, where $\mathbf{x}$ denotes a protein's conformation and $\mathbf{d}$ denotes spatially and temporally averaged, experimental data. The variable $\mathbf{x}$ represents a positional microstate in atomic detail. Through a forward model $f(\mathbf{x})$ we can calculate a coarse-grained representation, $\mathbf{f}$, of a protein conformation $\mathbf{x}$. That is, our forward model is a mapping, $f:\mathbb{R}^{3N}\rightarrow\mathbb{R}^{M}$, of the N atoms of $\mathbf{x}$ to an $M<3N$-dimensional coarse grained representation, $\mathbf{f}$. Conceptually, $\mathbf{f}$ may be interpreted as the instantaneous 'experimental data' back-calculated from a positional micro-state, $\mathbf{x}$. However, as $\mathbf{d}$ represents an averaged quantity we need to introduce a variable, $\mathbf{e}$, to represent an *ensemble average* of the simulated experimental data $\mathbf{f}$. Consequently, our full posterior distribution becomes $p(\mathbf{x},\mathbf{f},\mathbf{e}|\mathbf{d})$.

We clarify the relation between $\mathbf{f}$, $\mathbf{e}$ and $\mathbf{d}$ 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 $\mathbf{f}$ is a vector related to pairwise distances between atoms in a protein conformation $\mathbf{x}$. In one case, this is simply a vector of these distances. The variable $\mathbf{e}$ is an average of $\mathbf{f}$ vectors from an ensemble of protein conformations. The experimental NOE data $\mathbf{d}$ can be interpreted as a noisy observation of the vector of averages, $\mathbf{e}$. In general, there is no simple relationship between the vector $\mathbf{f}$ and the averaged vector $\mathbf{e}$, 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, $\mathbf{f}$, $\mathbf{e}$ and $\mathbf{d}$, without considering the fine-grained representation, $\mathbf{x}$. Following the Bayesian probability calculus, we formulate a posterior distribution,

$$\begin{aligned} p(\mathbf{f},\mathbf{e}|\mathbf{d}) &\propto p(\mathbf{d}|\mathbf{f},\mathbf{e})\pi(\mathbf{f},\mathbf{e}) \\ &= p(\mathbf{d}|\mathbf{e})\pi(\mathbf{f},\mathbf{e}), \end{aligned} \tag{1}$$

where the first term is the likelihood and the second term is the prior distribution. Note that the prior of $\mathbf{d}$ is in variant during inference and left out, hence the proportionality. The equality is due to the redundancy of $\mathbf{f}$ in the evaluation of the likelihood function – $\mathbf{d}$ is a noisy observation of $\mathbf{e}$, which does not involve $\mathbf{f}$. 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

$$p(\mathbf{f},\mathbf{e}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{e})\pi_f(\mathbf{f}|\mathbf{e})\pi_e(\mathbf{e}). \tag{2}$$

$\pi_f(\mathbf{f}|\mathbf{e})$ is the prior distribution of the simulated data $\mathbf{f}$ given their averaged value $\mathbf{e}$, and $\pi_e(\mathbf{e})$ is the prior distribution over the simulated ensemble averaged data $\mathbf{e}$.

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

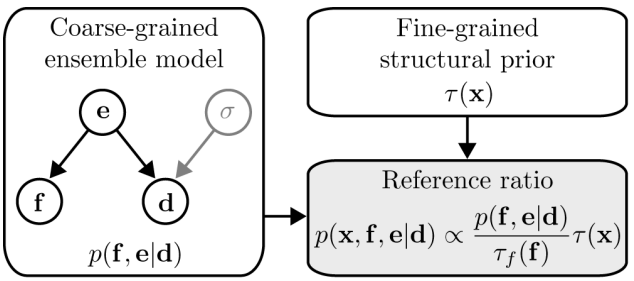

Figure 1. A directed graphical model of the ensemble model (on the left) and its interplay with a fine-grained conformational prior distribution (top right) through the reference ratio method, (bottom right). In the graphical model, the black circles are random variables, and the arrows determine their conditional independencies. The parameter σ, marked in grey on the left, is fixed and given, and denotes the experimental error in this particular example. $\tau_f(\mathbf{f})$ denotes the reference distribution.
doi:10.1371/journal.pone.0079439.g001

tion $p(\mathbf{x},\mathbf{f},\mathbf{e}|\mathbf{d})$ which features atomic detail, we need to combine (2) with a fine-grained physical forcefield or a probability distribution, $\tau(\mathbf{x})$. This can be done by using the *reference ratio method* [32,33],

$$p(\mathbf{x},\mathbf{f},\mathbf{e}|\mathbf{d}) \propto \frac{p(\mathbf{d}|\mathbf{e})\pi_f(\mathbf{f}|\mathbf{e})\pi_e(\mathbf{e})}{\tau_f(\mathbf{f})}\tau(\mathbf{x}). \quad (3)$$

$\tau_f(\mathbf{f})$ is called the reference distribution, and is the distribution *induced* in the coarse-grained space by the fine-grained prior, $\tau(\mathbf{x})$. That is, the prior distribution of $\mathbf{x}$, directly implies a prior distribution on $\mathbf{f}$, due to the parameters deterministic relationship through the forward model, $f(\cdot)$. This induced prior is called the reference distribution.

The reference ratio method yields the Kullback-Leibler optimal modification of the fine-grained model $\tau(\mathbf{x})$ with respect to the coarse-grained information (for proof, see chapter 4 in [34]). Kullback-Leibler optimality is closely linked to the maximum entropy principle of Jaynes [35]. In essence, our approach can be seen as a maximum entropy solution given the noisy observation of an ensemble average. It should be noted that even if the distribution given by Equation (2) is unimodal, the posterior given by Equation (3) can still be multimodal due to the nature of the conformational prior, $\tau(\mathbf{x})$.

The relationship to other methods

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

$$p(\mathbf{x},\mathbf{f}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{f})\tau(\mathbf{x}) \quad (4)$$

if we choose the Dirac delta function $\delta(\mathbf{f}-\mathbf{e})$ for $\pi_f(\mathbf{f}|\mathbf{e})\pi_e(\mathbf{e})$ and assume that $\tau_f(\mathbf{f})$ is uniform. Choosing the Dirac delta function corresponds to assuming the Boltzmann distribution is infinitely narrow. Hence, our model can be seen as a generalization of ISD. The choice of the uniform distribution for $\tau_f(\mathbf{f})$ corresponds to assuming that $\tau(\mathbf{x})$ implies a suitable prior for $\mathbf{f}$ as well. This may be inappropriate in some cases (see below).

We also observe that Equation (2) may be reduced to the previously proposed maximum entropy restraining methods [27–29]. This is evident if we consider the case where $p(\mathbf{d}|\mathbf{e})$ is the normal distribution and $\pi_f(\mathbf{f}|\mathbf{e})\pi_\mathbf{e}(\mathbf{e})$ is a log-linear model $\mathcal{G}(\cdot)$ with a linear *link-function*, $\ell(\mathbf{A},\mathbf{b})=\mathbf{Ab}$. The link function allows us to include the Lagrange multipliers used to relate the coarse-grained variable $\mathbf{f}$ to the mean value $\bar{\mathbf{f}}$ [36]. Thus, $\mathcal{G}(\mathbf{f}|\Lambda,\mathbf{e}) = \exp[c+\mathbf{f}^{\mathrm{T}}\ell(\Lambda,\mathbf{e})] \propto \exp(\mathbf{f}^{\mathrm{T}}\Lambda\mathbf{e})$. We have

$$p(\mathbf{f},\mathbf{e}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{e})\pi_f(\mathbf{f}|\mathbf{e})\pi_e(\mathbf{e}) = \mathcal{N}(\mathbf{d}|\mathbf{e},\sigma)\mathcal{G}(\mathbf{f}|\Lambda,\mathbf{e}) \quad (5)$$

where Λ is a diagonal matrix of Lagrange multipliers. If we now consider the limit where the experimental noise vanishes we obtain,

$$\lim_{\sigma\to 0}\mathcal{N}(\mathbf{d}|\mathbf{e},\sigma)\mathcal{G}(\mathbf{f}|\Lambda,\mathbf{e}) = \mathcal{G}(\mathbf{f}|\Lambda,\mathbf{d}). \quad (6)$$

In minus log-space Equation 6 is proportional to minus $\mathbf{f}^{\mathrm{T}}\Lambda\mathbf{d}$. 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 $\tau_f(\mathbf{f})$ when combining the empirical term in the coarse-grained space with a fine-grained prior distribution, $\tau(\mathbf{x})$. However, if the prior $\tau_f(\mathbf{f})$ is appropriate, then the Lagrange multipliers Λ may provide the necessary means for minor adjustments.

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 hair-pin 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 $\mathrm{NOE}\propto\langle r^{-6}\rangle$ [37]. 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 [38]. The choice for the prior $\pi_f(\mathbf{f}|\mathbf{e})$ is less obvious and depends on the type of experimental data. Here, we use the exponential distribution with mean β, $\mathcal{E}(x|\beta)=e^{-\frac{x}{\beta}}\beta^{-1}$, since it constitutes the least biasing continuous distribution on the positive real axis, when no higher order moments are observed

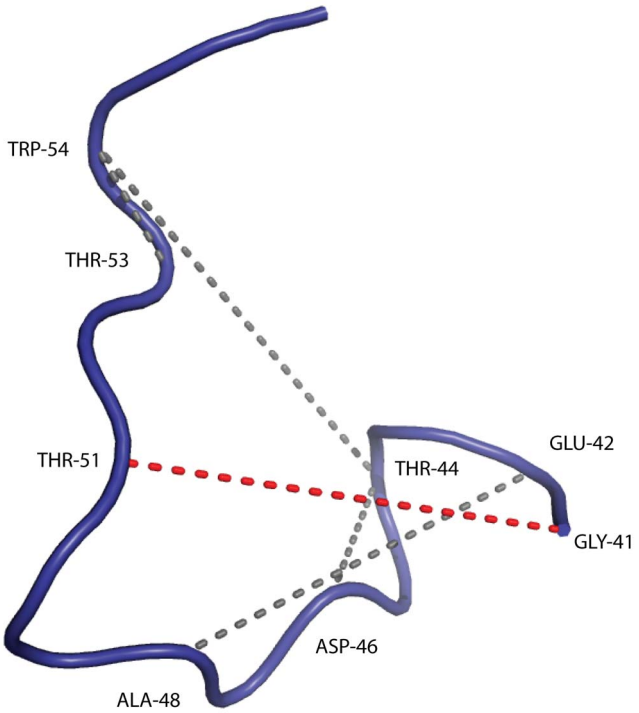

Figure 2. A random backbone conformation of the GB1 hairpin. The restraints listed in Table 1 are shown as dashed lines. The distance shown in red is used as the reaction coordinate $\mathbf{f}_0$ used in Figures 3 and 4. This figure was created using PyMOL (DeLano Scientific LCC).
doi:10.1371/journal.pone.0079439.g002

Table 1. Synthetic datasets used in this study.

Cα-pair	$\langle r_i^{-6} \rangle$	$\langle r_i \rangle$
41–51	$4.799 \cdot 10^{-7}$	19.40
42–48	$1.32 \cdot 10^{-6}$	14.36
44–46	$2.07 \cdot 10^{-5}$	6.06
44–54	$2.17 \cdot 10^{-7}$	19.39
53–54	$13.29 \cdot 10^{-4}$	3.51

First colum: Cα atoms involved in the pairwise distance. Second and last columns: averaged and power-averaged pairwise distances, respectively.
doi:10.1371/journal.pone.0079439.t001

[39]. The prior on $\mathbf{f}$ thus becomes: $\pi_f(\mathbf{f}|\mathbf{e}) = \pi_f(\mathbf{f}|\mathbf{e},\mathbf{w}) = \mathbf{\Pi}_i \mathcal{E}(\mathbf{f}_i|\mathbf{e}_i\mathbf{w}_i)$, where the product runs over all data-points and the scale vector $\mathbf{w}$ is a free parameter (discussed below). It follows that

$$p(\mathbf{x},\mathbf{f},\mathbf{e}|\mathbf{d}) \propto \frac{\mathcal{N}(\ln \mathbf{d}|\ln \mathbf{e},\sigma)\pi_f(\mathbf{f}|\mathbf{e},\mathbf{w})\pi_e(\mathbf{e})}{\tau_f(\mathbf{f})}\tau(\mathbf{x}), \qquad (7)$$

where $\boldsymbol{\sigma}$ is the experimental error, which is fixed and given, and $\mathcal{N}(\cdot)$ is the normal distribution. As prior on the ensemble average $\mathbf{e}$ we choose $\pi_e(\mathbf{e}) \propto \mathbf{e}^{-1}$. 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 $\mathbf{w}$, which relates $\mathbf{f}$ to $\mathbf{e}$ in $\pi_f(\mathbf{f}|\mathbf{e},\mathbf{w})$.

In the ideal case, an optimal choice for $\mathbf{w}$ results in the desired distribution for $\mathbf{f}$ as calculated from the structures $\mathbf{x}$. More precisely, it results in a marginal posterior distribution of Equation (7) for $\mathbf{x}$, such that, when $\mathbf{e}$ is fixed, the expectation of $\mathbf{f}$ is equal to $\mathbf{e}$. In practice, a satisfactory point estimate for $\mathbf{w}$ can be obtained in an iterative manner, using an empirical Bayes approach (see Materials and Methods). The parameter $\mathbf{w}$ compensates for the approximate nature the reference distribution $\tau_f(\mathbf{f})$, which is difficult to estimate accurately [33]. The introduction of $\mathbf{w}$ 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 $\tau(\mathbf{x})$ 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 $\mathbf{f_0}$ between atoms $C\alpha^{41}$ and $C\alpha^{51}$ – 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 (Figure3). 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 ± 1.5Å. The ensembles obtained with our method from the power averaged and linearly averaged data resulted in a slightly higher average (10.30 ± 1.8Å and 10.19 ± 1.6Å, 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 ± 0.6Å). Finally, sampling from the prior distribution alone results in an average radius of gyration of 11.34 ± 1.8Å, 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 chosing the reference distribution $\tau_f(\mathbf{f})$ and the scale vector, $\mathbf{w}$, 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 $\tau_f(\mathbf{f})$ 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 $\tau_f(\mathbf{f})$ uniform and $\mathbf{w}$ equal to the 1-vector, severely skews the distribution of the distances (green line in figure 4). When the scale vector $\mathbf{w}$ is estimated, while still assuming $\tau_f(\mathbf{f})$ 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 $\mathbf{w}$ provides a good fit (red line).

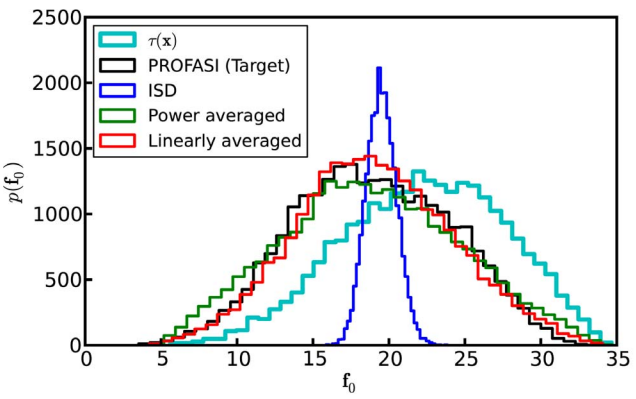

Figure 3. Histograms, $p(\mathbf{f}_0)$, of a representative pairwise distance $\mathbf{f}_0$ (between $C\alpha^{41}$-$C\alpha^{51}$, in A) in the ensembles. The black and blue lines are obtained from the PROFASI and ISD ensembles respectively, while the cyan line represent the prior $\tau(\mathbf{x})$. 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

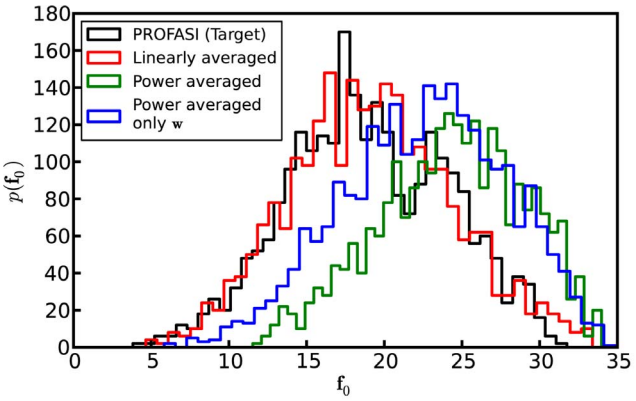

Figure 4. The influence of $\tau_f(\mathbf{f})$ and $\mathbf{w}$ on the ensembles. The figure shows histograms, $p(\mathbf{f}_0)$, of a representative pairwise distance $\mathbf{f}_0$ (between $C\alpha^{41}$-$C\alpha^{51}$, in Å) in the ensembles obtained without the reference distribution $\tau_f(\mathbf{f})$ or the scale vector $\mathbf{w}$. The black line denotes the PROFASI target ensemble; the red and green lines denote the ensembles obtained using the linearly and the power averaged data, respectively. The blue line denotes the case of the power averaged data without $\tau_f(\mathbf{f})$, but with $\mathbf{w}$.
doi:10.1371/journal.pone.0079439.g004

If we again consider the gyration radii as providing complementary views of the ensembles, we find that the power-averaged ensemble with uniform $\tau_f(\mathbf{f})$ 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 $\tau_f(\mathbf{f})$ and $\mathbf{w}$ are required for a satisfactory distribution. In the case of the linearly averaged data, our results suggest that $\mathbf{w}$ and $\tau_f(\mathbf{f})$ may be approximated by the 1-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 $\tau_f(\mathbf{f})$ is considerable. When we assumed $\tau_f(\mathbf{f})$ to be uniform, we were unable to construct sufficiently accurate distribution of pair-wise 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 $3.5{\cdot}10^8$ 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 Tabel 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 $I_i = \langle r_i^{-6} \rangle$, where r_i is an inter-atomic distance and the angular-brackets denote ensemble averaging [46]. We used an experimental uncertainty for the power averaged dataset σ_I of the same relative amplitude as for the average restraint set σ_d, by enforcing the signal-to-noise ratio to be constant. Hence, $\frac{\overline{d}}{\sigma_d} = \frac{\overline{I}}{\sigma_I} \Rightarrow \sigma_I = \frac{\overline{I}}{\overline{d}}$, as $\sigma_d = 1$, where $\overline{I}$ and $\overline{d}$ denote the datapoints corresponding to the largest average distance in the power and linearly averaged datasets, respectively. Noise with standard-deviation σ_I was added to the power-averaged data.

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

This section describes the estimation of the reference distribution $\tau_f(\mathbf{f})$ and the vector $\mathbf{w}$ needed for the posterior distribution:

$$p(\mathbf{x},\mathbf{f},\mathbf{e}|\mathbf{d}) \propto \frac{\mathcal{N}(\ln \mathbf{d}|\ln \mathbf{e},\sigma)\pi_{\mathbf{f}}(\mathbf{f}|\mathbf{e},\mathbf{w})\pi_{\mathbf{e}}(\mathbf{e})}{\tau_f(\hat{\mathbf{f}})}\tau(\mathbf{x}). \quad (8)$$

In the case of the power-averaged data, the reference distribution $\tau_f(\hat{\mathbf{f}})$ was approximated by a product of exponential distributions:

$$\tau_f(\hat{\mathbf{f}}) \propto \prod_i \mathcal{E}(-\hat{\mathbf{f}}_i/\mathbf{b}_i). \quad (9)$$

The mean β was estimated using a Monte Carlo scheme similar to that used to form the synthetic datasets, but only using the prior $\tau(\mathbf{x})$, 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, $\hat{\mathbf{f}}$, was the inverse pairwise distance between the Cα atoms listed in Table 1. For the linearly-averaged data, $\tau_f(\hat{\mathbf{f}})$ 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 (8), 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:

$$\mathbf{w_{n+1,i}} = \mathbf{w_{n,i}} \frac{\overline{\mathbf{f}}_\mathbf{i}}{\overline{\mathbf{e}}_\mathbf{i}}.$$

$\overline{\mathbf{f}}$ and $\overline{\mathbf{e}}$ 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 $\overline{\mathbf{f}}$ were within the experimental uncertainty. Each step in the algorithm runs for $2.5 \cdot 10^6$ MCMC steps, and a final production ensemble is produced using $25 \cdot 10^6$ MCMC steps.

Sampling of $\mathbf{e}$

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

$$\alpha(\mathbf{e} \rightarrow \mathbf{e}\Delta) = \min\left(1, \frac{p(\mathbf{e}\Delta)}{p(\mathbf{e})}\right),$$

$$\text{where } p(\cdot) = \frac{p(\mathbf{e}|\mathbf{f},\mathbf{x},\mathbf{d})}{\pi_e(\mathbf{e})}.$$

Acknowledgments

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

Author Contributions

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

References

1. Brünger AT, Nilges M (1993) Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. Q Rev Biophys 26: 49–125.
2. Jardetzky O (1980) On the nature of molecular conformations inferred from highresolution NMR. Biochim Biophys Acta 621: 227 - 232.
3. Jack A, Levitt M (1978) Refinement of large structures by simultaneous minimization of energy and *R* factor. Acta Crystallogr A 34: 931–935.
4. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 309: 303–306.
5. Habeck M (2011) Statistical mechanics analysis of sparse data. J Struct Biol 173: 541– 548.
6. Torda AE, Scheek RM, van Gunsteren WF (1989) Time-dependent distance restraints in molecular dynamics simulations. Chem Phys Lett 157: 289 - 294.
7. Torda AE, Scheek RM, van Gunsteren WF (1990) Time-averaged nuclear overhauser effect distance restraints applied to tendamistat. J Mol Biol 214: 223 - 235.
8. Lindorff-Larsen K, Kristjansdottir S, Teilum K, Fieber W, Dobson CM, et al. (2004) Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. J Am Chem Soc 126: 3291-3299.
9. Kim Y, Prestegard JH (1989) A dynamic model for the structure of acyl carrier protein in solution. Biochemistry 28: 8792-7.
10. Teilum K, Olsen JG, Kragelund BB (2011) Protein stability, flexibility and function. Biochim Biophys Acta 1814: 969-976.
11. Jha AK, Colubri A, Freed KF, Sosnick TR (2005) Statistical coil model of the unfolded state: Resolving the reconciliation problem. Proc Natl Acad Sci USA 102: 13099-13104.
12. Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RWH, et al. (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proc Natl Acad Sci USA 102: 17002-17007.
13. Chen Y, Campbell SL, Dokholyan NV (2007) Deciphering protein dynamics from NMR data using explicit structure sampling and selection. Biophys J 93: 2300 - 2306.
14. Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, et al. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J Am Chem Soc 131: 17908-17918.
15. Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, et al. (2010) Conformational space of flexible biological macromolecules from average data. J Am Chem Soc 132: 13553-13558.
16. Fisher CK, Huang A, Stultz CM (2010) Modeling intrinsically disordered proteins with bayesian statistics. J Am Chem Soc 132: 14919-14927.
17. Guerry P, Salmon L, Mollica L, Ortega Roldan JL, Markwick P, et al. (2013) Mapping the population of protein conformational energy sub-states from NMR dipolar couplings. Angew Chem Int Ed Engl 52: 3181-3185.
18. Wang L, Donald BR (2006) A data-driven, systematic search algorithm for structure determination of denatured or disordered proteins. Comput Syst Bioinformatics Conf : 67-78.
19. Shehu A, Clementi C, Kavraki LE (2006) Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. Proteins 65: 164-79.
20. Donald BR (2011) Algorithms in Structural Molecular Biology. The MIT Press.
21. Kemmink J, van Mierlo CP, Scheek RM, Creighton TE (1993) Local structure due to an aromatic-amide interaction observed by 1H-nuclear magnetic resonance spectroscopy in peptides related to the N terminus of bovine pancreatic trypsin inhibitor. J Mol Biol 230: 312 - 322.
22. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433: 128-132.
23. Lange OF, Lakomek NA, Farès C, Schröder GF,Walter KFA, et al. (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320: 1471-5.
24. Groth M, Malicka J, Czaplewski C, O ldziej S, Lankiewicz L, et al. (1999) Maximum entropy approach to the determination of solution conformation of flexible polypeptides by global conformational analysis and NMR spectroscopy–application to DNS1-c-[DA2, bu2,Trp4,Leu5]enkephalin and DNS1-c-[D-A2bu2,Trp4,D-Leu5]enkephalin. J Biomol NMR 15: 315-30.
25. Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM (2005) Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. J Am Chem Soc 127: 476-477.
26. Bonvin AMJJ, Boelens R, Kaptein R (1994) Time- and ensemble-averaged direct NOE restraints. J Biomol NMR 4: 143-149.
27. Pitera JW, Chodera JD (2012) On the use of experimental observations to bias simulated ensembles. J Chem Theory Comput 8: 3445.
28. Cavalli A, Camilloni C, Vendruscolo M (2013) Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. J Chem Phys 138: 094112.
29. Roux B, Weare J (2013) On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. J Chem Phys 138: 084107.
30. Irbäck A, Mitternacht S, Mohanty S (2009) An effective all-atom potential for proteins. PMC Biophys 2: 2.
31. Cavanagh J, Fairbrother WJ, III AGP, Skelton NJ, Rance M (2006) Protein NMR Spectroscopy: Principles And Practice. Academic Press, 2 edition.
32. Diaconis P, Zabell SL (1982) Updating subjective probability. J Am Statist Assoc 77: 822-830.
33. Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, et al. (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLoS One 5: e13714.
34. Hamelryck T, Mardia K, Ferkinghoff-Borg J, editors (2012) Bayesian Methods in Structural Bioinformatics. Springer.
35. Kullback S (1968) Information Theory and Statistics. Dover.
36. McCullagh P, Nelder JA (1989) Generalized linear models. Chapman & Hall, London, 2 edition.
37. Gronenborn AM, Clore G (1985) Investigation of the solution structures of short nucleic acid fragments by means of nuclear overhauser enhancement measurements. Prog Nucl Magn Reson Spectrosc 17: 1 - 32.
38. Rieping W, Habeck M, Nilges M (2005) Modeling errors in NOE data with a log-normal distribution improves the quality of NMR structures. J Am Chem Soc 127: 16026–16027.
39. Jaynes E (2003) Probability theory: The logic of science. Cambridge.

40. Habeck M, Rieping W, Nilges M (2006) Weighting of experimental evidence in macromolecular structure determination. Proc Natl Acad Sci USA 103: 1756–1761.
41. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, et al. (2008) A generative, probabilistic model of local protein structure. Proc Natl Acad Sci USA 105: 8932–8937.
42. Harder T, Boomsma W, Paluszewski M, Frellsen J, Johansson KE, et al. (2010) Beyond rotamers: a generative, probabilistic model of side chains in proteins. BMC Bioinformatics 11: 306.
43. Olsson S, Boomsma W, Frellsen J, Bottaro S, Harder T, et al. (2011) Generative probabilistic models extend the scope of inferential structure determination. J Magn Reson 213: 182-186.
44. Nilges M, O'Donoghue SI (1998) Ambiguous NOEs and automated NOE assignment. Prog Nucl Magn Reson Spectrosc 32: 107–139.
45. Engh RA, Huber R (1991) Accurate bond and angle parameters for x-ray protein structure refinement. Acta Cryst A47: 392-400.
46. Tropp J (1980) Dipolar relaxation and nuclear Overhauser effects in nonrigid molecules: The effect of fluctuating internuclear distances. J Chem Phys 72: 6035–6043.
47. Boomsma W, Frellsen J, Harder T, Bottaro S, Johansson K, et al. (2013) PHAISTOS: A framework for Markov Chain Monte Carlo simulation and inference of protein structure. J Comp Chem 34: 1697–1705.