On the significance of the reference ratio method in inferential structure determination of biomolecules.

Simon Olsson and Thomas Hamelryck

Bioinformatics Centre, University of Copenhagen

Abstract

The inference of biomolecular structure from biophysical data is an important task in molecular biology. Rigorous Bayesian inference requires the formulation of a joint posterior distribution regarding structure and nuisance parameters given the observed data. The relationship between biomolecule, typically a vector of atomic coordinates, and data is typically many-to-one. Consequently, *forward models* are needed to relate a given biomolecular structure to its associated data. Thus, the calculation of the likelihood involves projecting a high dimensional manifold to a lower dimensional one, respectively concerning structure and data. This projection is a *reduced* or *coarse grained representation* of the structure.

Given the nature of the data obtained from the biophysical experiments, the use of prior distributions concerning biomolecular structure is indispensable. A prior on biomolecular structure necessarily also induces a prior on its reduced or coarse-grained representation. We call this induced prior the *reference distribution*. The reference distribution induced by a fine-grained prior is typically assumed to be suitable for the coarse-grained variable. Often, this assumption is invalid. Here, we quantify the impact of the induced reference distribution on the posterior distribution and discuss its possible implications.

1 Background

Bio-molecular function is closely connected to structure. Consequently, the inference of structure from biophysical experiments is an important problem in molecular biology. However, the procedure is complicated by the nature of the experimental data which, are incomplete, averaged and subject to experimental noise. This fact makes it difficult, if not impossible, to use experimental observations to determine structures without the use of strong prior information.

The determination of bio-molecular structure usually concerns an atomic-level, or *fine-grained* representation, $\mathbf{x} \in \mathcal{N}$. The experimental observations, $\mathbf{d}$, provide information of some projection $\mathbf{f}$ of $\mathbf{x}$. The relationship between $\mathbf{f}$ and $\mathbf{x}$ is given by a *forward model* as $\mathbf{f} = \mathcal{F}(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{M}$ where $\mathcal{F} : \mathcal{N}, \mathcal{O} \rightarrow \mathcal{M}, \dim(\mathcal{N}) >> \dim(\mathcal{M})$, with nuisance parameters $\boldsymbol{\theta} \in \mathcal{O}$. That is, there is a deterministic relationship between the fine-grained $\mathbf{x}$ and coarse-grained representation $\mathbf{f}$, through $\mathcal{F}$.

In a practical structure inference setting, prior information on the fine-grained space $\mathcal{N}$ is typically introduced to compensate for the noisy and incomplete nature of experimental data. This inadvertently results in a prior distribution on the coarse-grained space $\mathcal{M}$, as a consequence of the variables' deterministic relationship. While the prior information about $\mathcal{N}$ may seem entirely appropriate, this is not guaranteed for the resulting induced prior distribution on $\mathcal{M}$. Moreover, if the induced prior is inappropriate, other free parameters, such as $\boldsymbol{\theta} \in \mathcal{O}$, may compensate for this. This will in turn introduce a bias which may hamper the direct interpretation of $\boldsymbol{\theta}$. Here, the aim is to demonstrate that the prior information used needs to be appropriate with respect to both $\mathcal{N}$ and $\mathcal{M}$.

2 Theory

The isolated spin-pair approximation (ISPA)

In nuclear magnetic resonance spectroscopy a key observable is the nuclear Overhauser enhancement (NOE), I. This parameter reflects the transfer of magnetization in between atomic nuclei, which is related to the inter-atomic distance, r as,

$$I = \gamma[r^{-6}] \tag{1}$$

if we assume the pair of nuclei is unaffected by other nuclei. The parameter γ is known as the equilibration parameter, and it relates the arbitrary scale of the nuclear Overhauser enhancement to the scale of the interatomic distance. Equation (1) is known as the isolated spin-pair approximation and constitutes one of the most common forward models in biomolecular structure determination from nuclear magnetic resonance data. For a set of n observed NOEs I_{obs} we can define a coarse-graining function as $\mathcal{F}(\mathbf{x}, \gamma) = \gamma\{r_1^{-6}, r_2^{-6}, \ldots, r_n^{-6}\} \in \mathbb{R}^{+n}$, where $\mathbf{x}$ defines an all-atom representation of a biological macromolecule, with inter-atomic distances r_i corresponding to the observed values.

Bayesian structure determination

Nilges and co-workers introduced a probabilistic, Bayesian treatment of macromolecular structure determination using NMR data, called inferential structure determination (ISD). Here, a posterior probability distribution of the unknown protein structure ($\mathbf{x}$) and nuisance parameters (γ, σ) is formulated given experimental data ($\mathbf{d}$),

$$p(\mathbf{x}, \gamma, \sigma \mid \mathbf{d}) \quad \propto \quad p(\mathbf{d} \mid \mathcal{F}(\mathbf{x}, \gamma), \sigma)\pi(\sigma, \gamma, \mathbf{x}) \tag{2}$$

where σ is the standard deviation of the likelihood. In the original formulation, the variables are conveniently assumed to be independent and in the case of σ and γ improper priors were used, resulting in $\pi(\sigma, \gamma, \mathbf{x}) = \tau_{\mathbf{x}}(\mathbf{x})\sigma^{-1}\gamma^{-1}$ (Rieping *et al.*, 2005a,b). Analogous assumptions were made for other experimental data sources, i.e. for different $\mathcal{F}$.

Here, we point out that the application of the forward model $\mathcal{F}$ to the fine-grained model $\mathbf{x}$ inherently introduces a new, coarse-grained, parameter $\mathbf{f}$. We may therefore recast the equation 2 into,

$$\begin{aligned} p(\mathbf{x}, \mathbf{f}, \gamma, \sigma \mid \mathbf{d}) \quad &\propto \quad p(\mathbf{d} \mid \mathbf{f}, \mathbf{x}, \sigma, \gamma)\pi(\sigma, \gamma, \mathbf{x}, \mathbf{f}) && (3) \\ &= \quad p(\mathbf{d} \mid \mathbf{f}, \sigma, \gamma)\pi(\sigma, \gamma, \mathbf{x}, \mathbf{f}), && (4) \end{aligned}$$

where we have used the conditional independence relationship, $\mathbf{d} \perp \mathbf{x} \mid \mathbf{f}$, in step two. If we again assume our prior knowledge of σ and γ is mutually independent and furthermore independent of $\mathbf{f}$ and $\mathbf{x}$ we get,

$$p(\mathbf{x}, \mathbf{f}, \gamma, \sigma \mid \mathbf{d}) \quad \propto \quad p(\mathbf{d} \mid \mathbf{f}, \sigma, \gamma)\pi(\mathbf{x}, \mathbf{f})\sigma^{-1}\gamma^{-1}. \tag{5}$$

It has been shown that the joint prior $\pi(\mathbf{f}, \mathbf{x})$ may be written as $\frac{\pi_{\mathbf{f}}(\mathbf{f})}{\tau_{\mathbf{f}}(\mathbf{f})}\tau_{\mathbf{x}}(\mathbf{x})$ (Hamelryck *et al.*, 2010). Here, $\tau_{\mathbf{f}}(\mathbf{f})$ is the distribution on $\mathcal{M}$ *induced* by the prior $\tau_{\mathbf{x}}(\mathbf{x})$ on $\mathcal{N}$, $\frac{\tau_{\mathbf{x}}(\mathbf{x})}{\tau_{\mathbf{f}}(\mathbf{f})}$ corresponds to the conditional prior $\pi_{\mathbf{x}|\mathbf{f}}(\mathbf{x} \mid \mathbf{f})$ on $\mathcal{N}$ and $\pi_{\mathbf{f}}(\mathbf{f})$ represents the prior distribution on $\mathcal{M}$. It now becomes evident that there is an additional assumption made implicitly in ISD: the desired distribution, $\pi_{\mathbf{f}}(\mathbf{f})$, on $\mathcal{M}$ is equal to the induced distribution $\tau_{\mathbf{f}}(\mathbf{f})$.

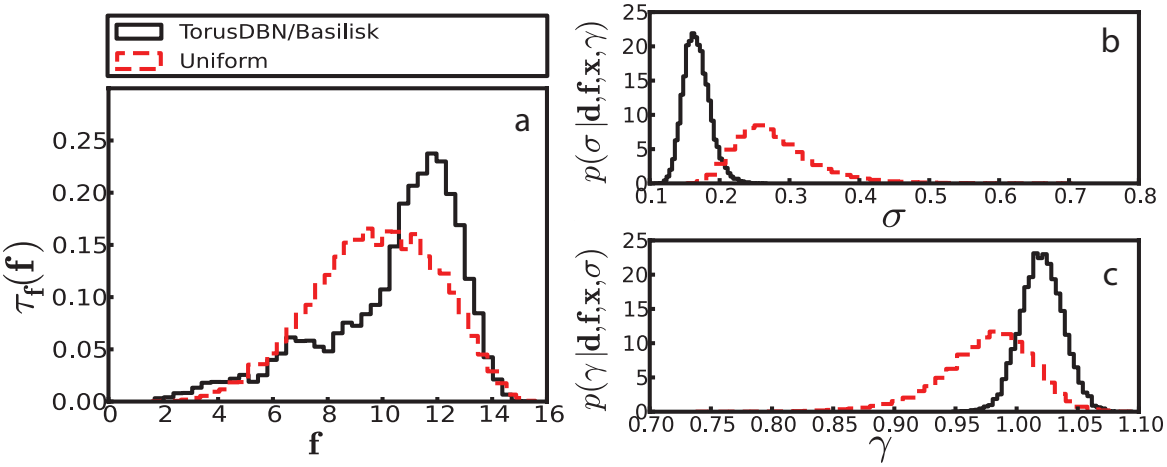

Figure 1: Illustration of how the marginal posterior distributions of model parameters σ and γ are affected by two different fine-grained priors, respectively involving TorusDBN/Basilisk (black) and the uniform distribution (red/dashed). Sub-figure a) shows a plot of $\tau_{\mathbf{f}}(\mathbf{f})$, where $\mathbf{f}$ corresponds to the distance between the atom pair (3-H$^{\alpha}$, 19-H$^{\beta 2}$). Sub-figures b) and c) show plots of $p(\sigma \mid \mathbf{d}, \mathbf{f}, \mathbf{x}, \gamma)$ and $p(\gamma \mid \mathbf{d}, \mathbf{f}, \mathbf{x}, \sigma)$, respectively.

In a previous study (Olsson *et al.*, 2011), we found that the marginal posterior distributions of the nuisance parameters, γ and σ, are affected by the choice of $\tau_{\mathbf{x}}(\mathbf{x})$. This suggests that invalidity of the assumption $\pi_{\mathbf{f}}(\mathbf{f}) = \tau_{\mathbf{f}}(\mathbf{f})$ may be of practical significance.

We illustrate these ideas using an example in the next section.

3 Example

We considered the small protein TRP-cage, using all unambiguous NOE data previously reported (Neidigh *et al.*, 2002). The posterior density of equation (2) was simulated as previously described (Olsson *et al.*, 2011) using two different fine-grained prior distributions, $\tau_{\mathbf{x}}(\mathbf{x})$. In the first case, the probabilistic models of protein dihedral angles TorusDBN (Boomsma *et al.*, 2008) and Basilisk (Harder *et al.*, 2010) were used. In the second case, uniform priors on these same dihedral angles were employed. In addition, for both priors, atomic positions were not allowed to overlap, due to *Van der Waals repulsion*. This was enforced using a simple binary condition. We also performed a set of 'reference' simulations were the fine-grained priors were simulated in absence of the likelihood, $p(\mathbf{d} \mid \mathbf{f}, \sigma, \gamma)$.

The two fine-grained priors yield realistic structure on a local length scale. The uniform prior on the dihedral angles effectively becomes a realistic model due to its combination with the binary exclusion condition (Ramachandran *et al.*, 1963). TorusDBN and Basilisk are probabilistic models, based on Bayesian networks, which were trained from experimental data. Although both models represent suitable priors for the local structure of proteins, they induce significantly different distributions on $\mathcal{M}$. An example of this is shown in Figure 1a.

4 Discussion

The aim here is to investigate how the marginal posterior distributions of model parameters such as σ and γ are affected by using similar, yet different fine-grained prior information. We approached this problem by simulating two joint posterior distributions using two different

fine-grained prior distributions, namely TorusDBN/Basilisk and a uniform prior (see details above). We found a significant difference in the induced prior distributions $\tau_{\mathbf{f}}(\mathbf{f})$ using the two different priors, see Figure 1a. These differences manifested themselves in the marginal posterior distribution of σ and γ, Figure 1b,c. This result is in agreement our previous report, where a discrepancy in the marginal posterior values of nuisance parameters was observed, compared to a previous study (Olsson *et al.*, 2010). This suggests that the interpretation of σ as experimental uncertainty and, to a lesser extent, γ as the equilibration parameter, is not straightforward. Rather, it seems σ is related to empirical weighing factors used in *hybrid energy minimization* heuristics (Habeck *et al.*, 2006). While these parameters are intrinsically nuisance parameters, that is, parameters of little practical interest, it is important to note, that this is not always the case. Consequently, caution is required when there is direct interest in these, or related, parameters.

Acknowledgements

SO acknowledges funding from the Danish Council for Independent Research (FTP09-066546).

References

Boomsma, W., Mardia, K.V., Taylor, C.C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008) A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. Sci. USA*, **105**, 8932–8937.

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

Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W., Bottaro, S. and Ferkinghoff-Borg, J. (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized *PLoS One*, **5**, e13714

Harder, T., Boomsma, W., Paluszewski, M., Frellsen, J., Johansson. K.E. and Hamelryck, T. (2010) Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics*, **11**, 306.

Neidigh, J.W., Fesinmeyer, R.M. and Andersen, N.H. (2002) Designing a 20-residue protein. *Nat.Struct.Biol.* **9**, 425-430.

Olsson, S., Boomsma, W., Frellsen, J., Bottaro, S., Harder, T., Ferkinghoff-Borg, J. and Hamelryck, T. (2011). Generative probabilistic models extend the scope of inferential structure determination *J. Magn. Reson.*, **213**, 182–186.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations *J. Mol. Biol.* **7**, 959.

Rieping, W., Habeck, M., and Nilges, M. (2005a). Inferential structure determination *Science*, **309**, 303–306.

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