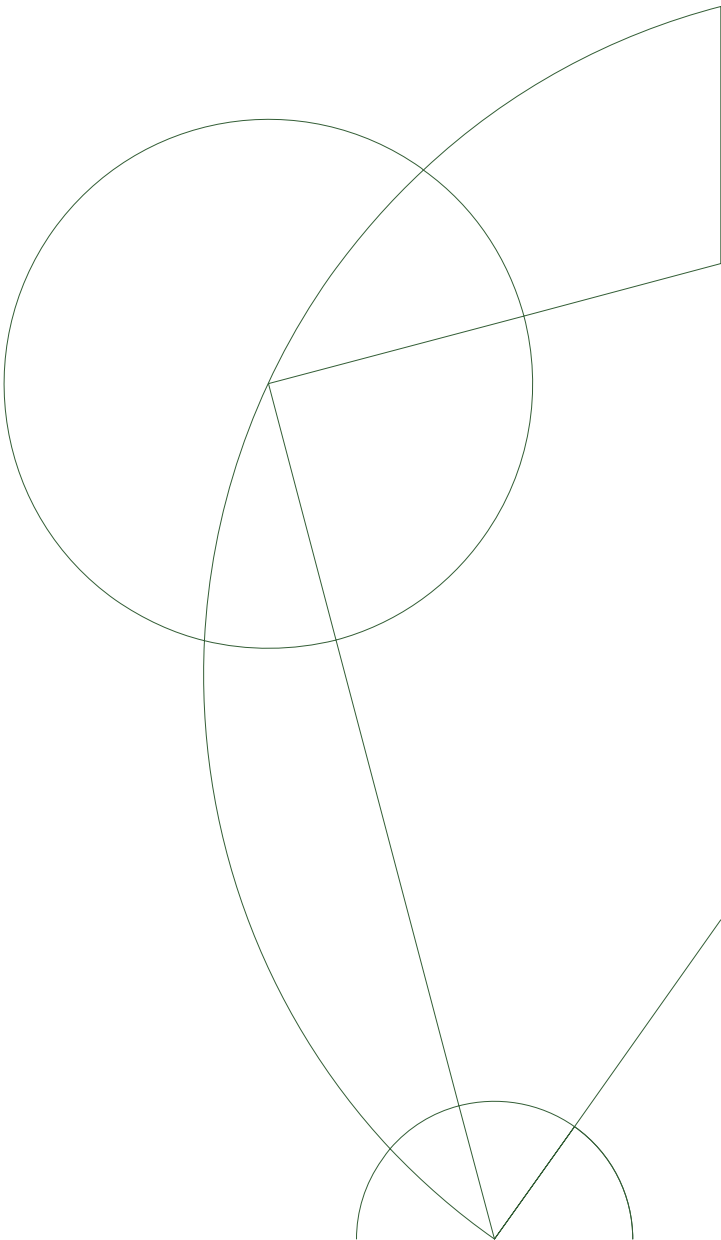
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PhD Thesis

Anders S. Christensen

Inferential Protein Structure Determination Using Chemical Shifts



Academic supervisor: Jan H. Jensen

January 29, 2014

Acknowledgements

I would like to thank the following people

- Jan Jensen
- Casper Steinmann
- More people

Publication list

List of publications:

1. Anders S. Christensen, Stephan P. A. Sauer, Jan H. Jensen (2011) Definitive benchmark study of ring current effects on amide proton chemical shifts. *Journal of Chemical Theory and Computation*, 7:2078-2084.

2. Wouter Boomsma, Jes Frellsen, Tim Harder, Sandro Bottaro, Kristoffer E. Johansson, Pengfei Tian, Kasper Stovgaard, Christian Andreetta, Simon Olsson, Jan B. Valentin, Lubomir D. Antonov, Anders S. Christensen, Mikael Borg, Jan H. Jensen, Kresten Lindorff-Larsen, Jesper Ferkinghoff-Borg, Thomas Hamelryck (2013) PHAISTOS: A framework for Markov chain Monte Carlo simulation and inference of protein structure. *Journal of Computational Chemistry*, 34:1697-1705.

3. Anders S. Christensen, Troels E. Linnet, Mikael Borg, Wouter Boomsma, Kresten Lindorff-Larsen, Thomas Hamelryck, Jan H. Jensen (2013) Protein Structure Validation and Refinement Using Amide Proton Chemical Shifts Derived from Quantum Mechanics. *PLoS ONE* 8:e84123.

4. Anders S. Christensen, Thomas Hamelryck, Jan H. Jensen (2014) FragBuilder: An efficient Python library to setup quantum chemistry calculations on peptides models. *PeerJ* (accepted).

5. Anders S. Christensen, Lars Bratholm, Simon Olsson, Thomas Hamelryck, Jan H. Jensen (2014) Weighting of chemical shift evidence in Monte Carlo simulation of proteins. *ShareLatex* (unpublished).

6. Torus-DBN-CS-LARS

7. J-coupling Casper/Kongen

List of public code:

1. FragBuilder (BSD license) `https://github.com/jensengroup/fragbuilder`

2. CamShift module (BSD license) `https://github.com/jensengroup/camshift-phaistos`

3. ProCS module (BSD license) `https://github.com/jensengroup/procs-phaistos`

4. Phaistos (GPL license) `https://svn.code.sf.net/p/phaistos/code/trunk`

5. GAMESS patch FMO-RHF:MP2 (GAMESS license/free) `https://github.com/andersx/fmo-rhf-mp2`

List of other publications:

1. Casper Steinmann, Kristoffer L. Blædel, Anders S. Christensen, Jan H. Jensen (2013) Interface of the polarizable continuum model of solvation with semi-empirical methods in the GAMESS program. *PLoS ONE* 8:e67725.

2. Anders S. Christensen, Casper Steinmann, Dmitri G. Fedorov, Jan H. Jensen (2013) Hybrid RHF/MP2 geometry optimizations with the Effective Fragment Molecular Orbital Method. *PLoS ONE* (accepted).

3. HF-3c Jimmy

4. PM6 Jimmy

5. h-bond Jimmy

Preface

This is the introduction

Contents

Chapter 1

Chemical shifts in a probabilistic framework

This section introduces the formalism for Monte Carlo simulations which includes both physical energy terms as well as a probabilistic energy terms based on experimentally observed chemical shifts. These equations presented are not new, but have not been published in the form in which they are presented here. The intention is to present the equations in the form in which they are implemented in PHAISTOS, so that they can easily be re-implemented in other programs by others.

1.1 Defining an energy function from Bayes' theorem

A simplistic approach to this problem is to is to define a hybrid energy by defining a penalty function that describes the agreement between experimental data and data calculated from a proposed protein structure with a physical energy (such as from a molecular mechanics force field). A structure is then determined by minimizing

$$E_{\text{hybrid}} = w_{\text{data}}\ E_{\text{data}} + E_{\text{physical}}. \tag{1.1}$$

This approach, however, does not uniquely define neither shape nor weight of E_{data}. Chemical shifts have been combined with physical energies in a multitude of ways, e.g., weighted RMSD values or harmonic constraints. Vendruscolo and co-workers implemented a "square-well soft harmonic potential", and corresponding molecular gradients and were able to run a chemical shift-biased MD simulation. In all cases the parameters and weights of E_{data} had to be carefully tweaked by hand, and it is not clear how to choose optimal parameters.

The inferential structure determination (ISD) principles introduced by Rie-ping, Habeck and Nigles [Rieping et al., 2005]defines a Bayesian formulation of Eq XX. In the following section the equations of an ISD approach are derived for combining the knowledge of experimental chemical shifts with a physical energy. First remember Bayes' theorem which relates a conditional probability (A given B) with its inverse:

$$p\left(A|B\right) = \frac{p\left(B|A\right)p\left(A\right)}{p\left(B\right)} \tag{1.2}$$

Now consider a set of chemical shifts $\{\delta_i\}$, and the uncertainty to which these can be predicted $\{\sigma_i\}$ from a structure, $\boldsymbol{X}$ (the experimental uncertainty is negligibly small compared to this). We have to make the basic assumption, that the error, given as $\Delta\delta_i = \left|\delta_i^{\text{predicted}} - \delta_i^{\text{experimental}}\right|$, approximately follows a Gaussian distribution with some standard deviation, but we need not

hand-pick and assign any numeric value to the standard deviation.Furthermore, the Gaussian distribution is the least biasing distribution according to the principle of maximum entropy.

In this case, the most likely structure, $\boldsymbol{X}$, and optimal choice of $\{\sigma_i\}$ is found by maximizing (via Bayes' theorem)

$$p\Big(\boldsymbol{X}, \{\sigma_i\}\Big|\{\delta_i\}\Big) = \frac{p\Big(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\Big)p\Big(\boldsymbol{X}, \{\sigma_i\}\Big)}{p\Big(\{\delta_i\}\Big)}. \tag{1.3}$$

Here, *marginal distribution* of $p\left(\{\delta_i\}\right)$ merely serves as a normalizing factor and the *likelihood* of $p\left(\{\delta_i\}|\boldsymbol{X}, \{\sigma_i\}\right)$, is obtained as the product of the individual, Gaussian probabilities over all n single chemical shift measurements. Nuclei of the same atom-type, here denoted by index j, (e.g. C^{α}, H^{α}, etc.) are assumed to carry the same uncertainly denoted by σ_j:

$$\begin{aligned}
p\Big(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\Big) &\simeq \prod_{i=0}^{n} p\left(\Delta\delta_i|\boldsymbol{X}, \sigma_i\right) && (1.4)\\
&= \prod_{j=0}^{m}\prod_{i_j=0}^{n_j} p\left(\Delta\delta_{i_j}|\boldsymbol{X}, \sigma_j\right) && (1.5)\\
&= \prod_{j=0}^{m}\prod_{i_j=0}^{n_j} \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left(-\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) && (1.6)\\
&= \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(\sum_{i_j=0}^{n_j}-\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) && (1.7)\\
&= \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(\frac{-\chi_j^2(\boldsymbol{X})}{2\sigma_j^2}\right) && (1.8)
\end{aligned}$$

Furthermore, $p\left(\boldsymbol{X}, \{\sigma_j\}\right)$ can be simplified as

$$\begin{aligned}
p\Big(\boldsymbol{X}, \{\sigma_j\}\Big) &\propto p\Big(\{\sigma_j\}\Big|\mathbf{X}\Big)p\Big(\mathbf{X}\Big) && (1.9)\\
&= p\Big(\{\sigma_j\}\Big)p\Big(\boldsymbol{X}\Big), && (1.10)
\end{aligned}$$

where it is assumed that the errors in the chemical shift prediction model are independent of the particular protein structure and *vice versa*. The *prior* distribution of $p\left(\{\sigma_j\}\right)$ is accounted for by proposing updates from a log-normal distribution (see next subsection). $p(\boldsymbol{X})$ of the molecular protein structure is here simply the Boltzmann distribution, i.e.

$$p(\boldsymbol{X}) = \frac{1}{Z}\exp\left(-\frac{E(\boldsymbol{X})}{k_{\mathrm{B}}T}\right) \tag{1.11}$$

where $E(\boldsymbol{X})$ is the (physical) potential energy of the protein structure, most often described by a molecular mechanics force field. k_{B} is the Boltzmann constant and T is the temperature of interest. Luckily we need not calculate the partition function, Z, because the relative energy landscape is invariant under choice of normalization constant. Note that $p(\boldsymbol{X})$ also can be introduced via conformational sampling from a biased distribution, such as for example TorusDBN or BASILISK (mimicking the Ramachandran plot and side chain rotamer distributions, respectively).

Neglecting normalization constants, the total probability to be maximized is thus proportional to:

$$p\Big(\boldsymbol{X},\{\sigma_i\}\Big|\{\delta_i\}\Big) \propto p\Big(\{\delta_i\}\Big|\boldsymbol{X},\{\sigma_i\}\Big)p\Big(\boldsymbol{X}\Big)p\Big(\{\sigma_i\}\Big) \tag{1.12}$$

$$\propto \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(-\frac{1}{2\sigma_j^2}\chi_j^2\right)\exp\left(-\frac{E(\boldsymbol{X})}{k_{\mathrm{B}}T}\right)p\Big(\{\sigma_j\}\Big) \tag{1.13}$$

When $p(\{\sigma_j\})$ is introduced via biased sampling, the associated hybrid-energy to be evaluated is (again neglecting constant terms)

$$E_{\mathrm{hybrid}}\left(\boldsymbol{X}\right) = -k_{\mathrm{B}}T\ln\left(p\Big(\boldsymbol{X},\{\sigma_i\}\Big|\{\delta_i\}\Big)\right) \tag{1.14}$$

$$= E(\boldsymbol{X}) - k_{\mathrm{B}}T\sum_{j=0}^{n_j} n_j \ln\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right) + \frac{\chi_j^2}{2\sigma_j^2} \tag{1.15}$$

1.1.1 Jeffreys' prior (general, one-parameter case)

The prior distribution of the nuisance parameter is inherently unknown. In such cases, it is necessary to use a prior distribution that will have only very little influence on the sampled value. One such *uninformative prior* could for instance be a flat distribution over the positive real line. The concept of Jeffreys' priors are a generalization of flat priors. In the one parameter case the Jeffrey's prior is given as

$$p(\theta) \propto \sqrt{\mathbf{I}(\theta)}, \tag{1.16}$$

where $\mathbf{I}(\theta)$ is the *Fisher information* defined (in the one parameter case) as

$$\mathbf{I}(\theta) = \left\langle\left(\frac{\partial}{\partial\theta}\ln p(x|\theta)\right)^2\right\rangle. \tag{1.17}$$

1.1.2 Jeffreys' prior (Gaussian and Cauchy distributions)

Here we derive Jefferys' prior for the uncertainty of a Gaussian distribution, i.e. a distribution on the form

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)}{2\sigma^2}\right). \tag{1.18}$$

This immediately gives us the Jeffreys' prior:

$$\begin{aligned} p(\sigma) &\propto \sqrt{\left\langle\left(\frac{\partial}{\partial\sigma}\ln p(x|\mu,\sigma)\right)^2\right\rangle} \\ &= \sqrt{\left\langle\left(\frac{\partial}{\partial\sigma}\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)}{2\sigma^2}\right)\right]\right)^2\right\rangle} \\ &= \sqrt{\left\langle\left(\frac{(x-\mu)-\sigma^2}{\sigma^3}\right)^2\right\rangle} \\ &= \sqrt{\int_{-\infty}^{\infty} p(x|\mu,\sigma)\left(\frac{(x-\mu)-\sigma^2}{\sigma^3}\right)^2 dx} \\ &= \sqrt{\frac{2}{\sigma^2}} \propto \frac{1}{\sigma} \end{aligned} \tag{1.19}$$

Similarly for the γ parameter of the Cauchy distribution of the form

$$p(x|x_0,\gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}, \tag{1.20}$$

we obtain the following Jeffreys' prior:

$$\begin{aligned} p(\gamma) &\propto \sqrt{\left\langle \left(\frac{\partial}{\partial\gamma}\ln p(x|x_0,\gamma)\right)^2 \right\rangle} \\ &= \sqrt{\left\langle \left(\frac{\partial}{\partial\gamma}\ln\left[\frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}\right]\right)^2 \right\rangle} \\ &= \sqrt{\left\langle \left(-\frac{\gamma^2-(x-x_0)^2}{\gamma^3+\gamma(x-x_0)^2}\right)^2 \right\rangle} \\ &= \sqrt{\int_{-\infty}^{\infty} p(x|x_0,\gamma)\left(-\frac{\gamma^2-(x-x_0)^2}{\gamma^3+\gamma(x-x_0)^2}\right)^2 dx} \\ &= \sqrt{\frac{1}{2\gamma^2}} \propto \frac{1}{\gamma} \end{aligned} \tag{1.21}$$

1.1.3 Sampling of nuisance parameters

Since the nuisance parameters of the energy functions are unknown, they too must be sampled. The move used to update the value of the nuisance parameters must obey detailed balance:

$$p(\theta \to \theta') = p(\theta' \to \theta) \tag{1.22}$$

The simplest Monte Carlo move is simply adding a number from a normal distribution with $\mu = 0$, this clearly obeys detailed balance, since the distribution is symmetric. For the scale parameter, γ and σ, of the Cauchy and Gaussian distributions, respectively, we found a variance of 0.05 in the normal distributed move to converge quickly and stably. Figure 1.1 show a histogram of sampled values of γ and σ for the NMR structure of Protein G (PDB-id: 2OED). 55 C-alpha

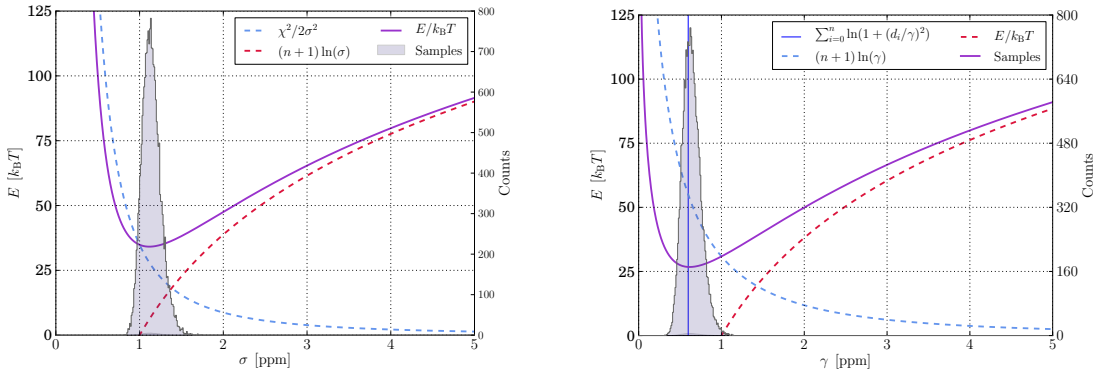
1.2 Prior distributions for protein structure

1.2.1 Molecular mechanics force field

One reasonable prior distribution for protein structure, $p(\mathbf{X})$, is the Boltzmann distributino, e.g.:

$$p(\mathbf{X}) \propto \exp\left(\frac{-E}{k_\mathrm{B}T}\right) \tag{1.23}$$

where E is the energy of the structure, $\mathbf{X}$ and k_B and T are Boltzmann's constant and the temperature, respectively. The energy of the structure is in this context usually approximated by a molecular mechanics force field that is taylor-made for protein simulations. PHAISTOS currently supports two different protein force field: The OPLS-AA/L force field with a GB/SA solvent term, and the PROFASI force field.



(a) Gaussian distribution

(b) Cauchy distribution

Figure 1.1: Sampling of σ and γ for 2OED for Ca-chemical shifts. $n = 55$ and $\chi^2 = 69.7$.

1.2.2 Generative probabilistic models

Another way to introduce the prior distribution for a protein structure is to bias the conformational sampling. Conventional conformational sampling will proposed (ϕ, ψ) backbone angles uniformly (i.e. in the range $[-180°, 180°]$ and let the energy function filter and construct the target distribution (e.g. the canonical ensemble, etc.) via energy evaluation. Since only a fraction of the possible (ϕ, ψ) backbone angles are allowed (i.e. the Ramachandran plot), it is computationally very convenient to only sample from the allowed regions. Using biased sampling, energy evaluation of structures that are obviously in sterically unfavored regions is eliminated with high efficiency.

Taking the biased sampling one step further, it is possible to have the biased sampling via TorusDBN conditioned on a set of chemical shifts. This sampling is carried out via the TorusDBN-CS model by Boomsma *et al.* TorusDBN-CS is trained on all chemical shift data available in the RefDB database, that is 1349 protein structures with their corresponding chemical shifts. This includes both experimental X-ray crystal and NMR structures.

In most cases TorusDBN-CS model is able to restrict the conformational sampling of (ϕ, ψ) angles to not only the Ramachandran plot, but also the correct region (e.g. alpha-helix, beta-sheet), etc. However, since the data set is smaller than that of the TorusDBN (non-CS) model TorusDBN-CS will occasionally be less restrictive than TorusDBN and may sample outside the Ramachandran plot.

TorusDBN

BASILISK

1.2.3 Or both

Bibliography

[Rieping et al., 2005] Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science*, 308:303–306.