- Thank you Jan Jensen.
- Thank you Casper Steinmann

Chapter 1

Introduction

This is the introduction

Chapter 2

Chemical shifts in a probabilistic framework

This section introduces the formalism for Monte Carlo simulations (or optimizations) 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 in, so that they can easily be re-implemented in other programs by others.

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

$$E_{\text{hybrid}} = w_{\text{data}}\ E_{\text{data}} + E_{\text{physical}}. \tag{2.1}$$

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 gradients and were able to run a 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 Rieping, Habeck and Nigles 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:

$$p\left(A|B\right) = \frac{p\left(B|A\right)p\left(A\right)}{p\left(B\right)} \tag{2.2}$$

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)

$$p\Big(\boldsymbol{X}, \{\sigma_i\}\Big|\{\delta_i\}\Big) = \frac{p\Big(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\Big)p\Big(\boldsymbol{X}, \{\sigma_i\}\Big)}{p\Big(\{\delta_i\}\Big)}. \tag{2.3}$$

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:

$$\begin{aligned}
p\Big(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\Big) &\simeq \prod_{i=0}^{n} p\left(\Delta\delta_i|\boldsymbol{X}, \sigma_i\right) && (2.4)\\
&= \prod_{j=0}^{m}\prod_{i_j=0}^{n_j} p\left(\Delta\delta_{i_j}|\boldsymbol{X}, \sigma_j\right) && (2.5)\\
&= \prod_{j=0}^{m}\prod_{i_j=0}^{n_j} \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left(-\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) && (2.6)\\
&= \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(\sum_{i_j=0}^{n_j} -\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) && (2.7)\\
&= \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(\frac{-\chi_j^2(\boldsymbol{X})}{2\sigma_j^2}\right) && (2.8)
\end{aligned}$$

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

$$\begin{aligned}
p\Big(\boldsymbol{X}, \{\sigma_j\}\Big) &\propto p\Big(\{\sigma_j\}\Big|\mathbf{X}\Big)p\Big(\mathbf{X}\Big) && (2.9)\\
&= p\Big(\{\sigma_j\}\Big)p\Big(\boldsymbol{X}\Big), && (2.10)
\end{aligned}$$

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 $p(\{\sigma_j\})$ 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.

$$p(\boldsymbol{X}) = \frac{1}{Z} \exp\left(-\frac{E(\boldsymbol{X})}{k_{\mathrm{B}}T}\right) \tag{2.11}$$

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:

$$p\Big(\boldsymbol{X}, \{\sigma_i\}\Big|\{\delta_i\}\Big) \propto p\Big(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\Big)p\Big(\boldsymbol{X}\Big)p\Big(\{\sigma_i\}\Big) \tag{2.12}$$

$$\propto \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j} \exp\left(-\frac{1}{2\sigma_j^2}\chi_j^2\right) \exp\left(-\frac{E(\boldsymbol{X})}{k_{\mathrm{B}}T}\right)p\Big(\{\sigma_j\}\Big) \tag{2.13}$$

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

$$E_{\mathrm{hybrid}}(\boldsymbol{X}) = -k_{\mathrm{B}}T \ln\left(p\Big(\boldsymbol{X}, \{\sigma_i\}\Big|\{\delta_i\}\Big)\right) \tag{2.14}$$

$$= E(\boldsymbol{X}) - k_{\mathrm{B}}T\sum_{j=0}^{n_j} n_j \ln\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right) + \frac{\chi_j^2}{2\sigma_j^2} \tag{2.15}$$

2.0.1 Sampling of $\{\sigma_j\}$

It is thus clear from Eq XX, that both $\{\sigma_j\}$ and $\boldsymbol{X}$ (i.e. a set of 3D coordinates) are treated as variables. We are thus required to take MC steps in 3D space as well as σ_j space. Computationally this is implemented using a set of standard MC moves to navigate in 3D space, and introducing a so-called *none move* in which all 3D coordinates are kept, but a small change in variables space is proposed. In short all terms besides $p(\{\sigma_j\})$ are calculated every Monte Carlo step in which the structure is physically altered and one element in $\{\sigma_j\}$ is conversely altered whenever a none move is performed.

If the error in the prediction model is assumed to follow a Gaussian distribution, then the values of $\{\sigma_j\}$ must be assumed to follow a log-normal distribution (because σ_j can only be a positive value). Since the shape parameters

of these log-normal distributions have unknown shape parameter (here denoted by $\{\sigma_{\sigma_j}\}$), a step in σ_{σ_j} space is also taken each none move.

The Monte Carlo criterion in a standard Metropolis-Hastings scheme is thus

$$\mathrm{acc}\Big(\boldsymbol{X}_k, \{\sigma_i\}_k \rightarrow \boldsymbol{X}_{k+1}, \{\sigma_i\}_{k+1}\Big) = \min\left\{1, \frac{p\Big(\boldsymbol{X}_{k+1}, \{\sigma_i\}_{k+1}\Big)}{p\Big(\boldsymbol{X}_k, \{\sigma_i\}_k\Big)}\right\} \tag{2.16}$$

which is evaluated as

$$\begin{aligned} \frac{p\Big(\boldsymbol{X}_{k+1}, \{\sigma_i\}_{k+1}\Big)}{p\Big(\boldsymbol{X}_k, \{\sigma_i\}_k\Big)} &= \frac{p\Big(\boldsymbol{X}_{k+1}\Big)}{p\Big(\boldsymbol{X}_k\Big)} \frac{p\Big(\{\sigma_i\}_{k+1}\Big)}{p\Big(\{\sigma_i\}_k\Big)} & (2.17) \\ &= \frac{\exp\Big(-E_{\mathrm{hybrid}}\left(\boldsymbol{X}_{k+1}\right)/k_{\mathrm{B}}T\Big)}{\exp\Big(-E_{\mathrm{hybrid}}\left(\boldsymbol{X}_k\right)/k_{\mathrm{B}}T\Big)} \frac{p\Big(\{\sigma_i\}_{k+1}\Big)}{p\Big(\{\sigma_i\}_k\Big)} & (2.18) \end{aligned}$$

The ratio describing the changes, such as changes in $\{\sigma_i\}$ which are not evaluated as part of the hybrid energy function, is named *sampling bias*. The bias we must compensate for, is introduced when the values of σ_j are proposed according to a log-normal distribution with σ_{σ_j}

$$p(\sigma_j|\sigma_{\sigma_j}, \mu_j) = \frac{1}{\sigma_j\sqrt{2\pi\sigma_{\sigma_j}^2}} \exp\left\{\frac{(\ln\sigma_j - \mu_j)^2}{2\sigma_{\sigma_j}^2}\right\} \tag{2.19}$$

Here we set $\sigma_{\sigma_j} = \left(\frac{\chi^2}{\sigma^2} - n\right)^{-1}$ and $\mu_j = 0$. Consequently, sampling of σ_j is itself biased due to the introduction of these qualified (although arbitrary) choices. We thus have to correct the (acceptance) probability XX by multiplying by the ratio of the proposal densities of the step and the reverse step:

$$\begin{aligned} \frac{p\Big(\{\sigma_j\}_{k+1} \rightarrow \{\sigma_j\}_k\Big)}{p\Big(\{\sigma_j\}_k \rightarrow \{\sigma_j\}_{k+1}\Big)} &= \frac{\prod_{j=0}^{m} p(\sigma_{j,k}|\sigma_{\sigma_{j,k}}, \mu_{j,k})}{\prod_{j=0}^{m} p(\sigma_{j,k+1}|\sigma_{\sigma_{j,k+1}}, \mu_{j,k+1})} & (2.20) \\ &= \frac{p(\sigma_{a,k}|\sigma_{\sigma_{a,k}}, \mu_{a,k})}{p(\sigma_{a,k+1}|\sigma_{\sigma_a,k+1}\mu_{a,k+1})} & (2.21) \\ &= \frac{\sigma_{a,k}\sqrt{2\pi\sigma_{\sigma_{a,k}}}}{\sigma_{a,k+1}\sqrt{2\pi\sigma_{\sigma_{a,k+1}}}} \exp\left\{\frac{(\ln\sigma_j - \mu_j)^2}{2\sigma_{\sigma_j}} - \frac{(\ln\sigma_j - \mu_j)^2}{2\sigma_{\sigma_j}}\right\} & (2.22) \end{aligned}$$

where a is the index of the atom type in $\{j\}$ for which a change in σ_j is proposed.