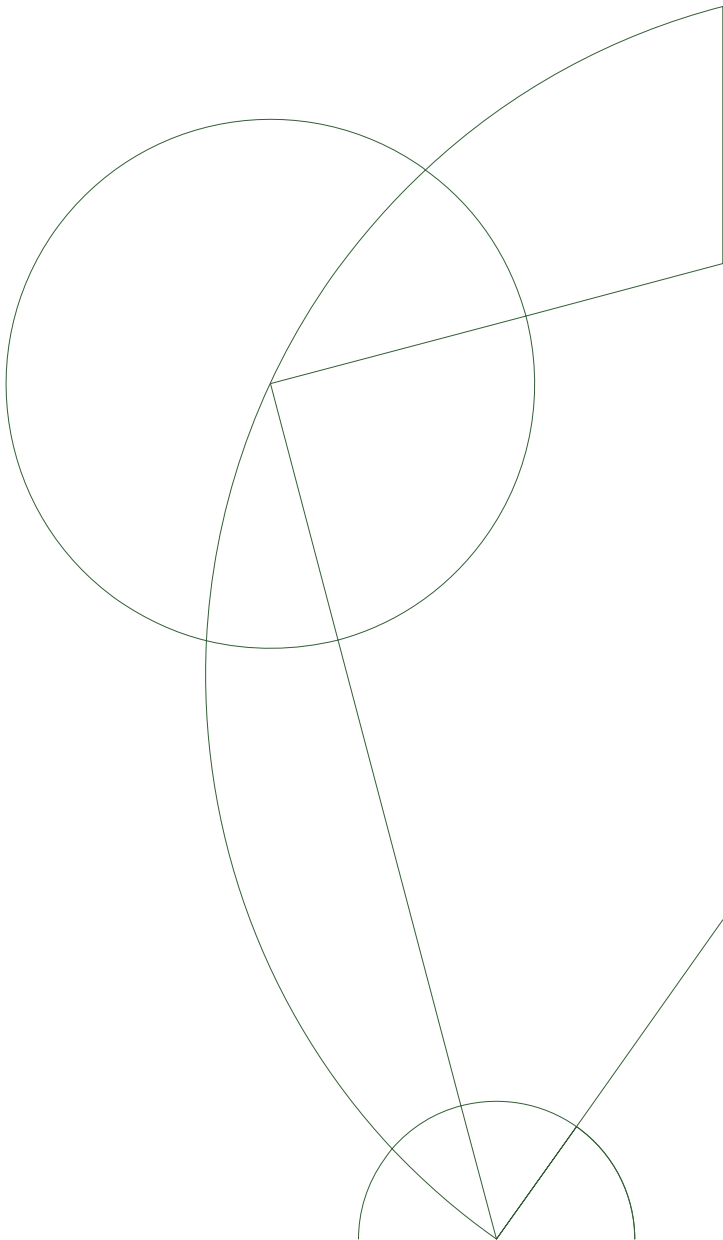
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PhD Thesis

Anders S. Christensen

Protein Structure Determination Using Chemical Shifts



Academic supervisor: Jan H. Jensen

March 10, 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 *ShareLatex* (unpublished).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`

6. PHAISTOS GUI (BSD license) `https://github.com/andersx/guistos`

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

Nuclear magnetic resonance (NMR) spectra of proteins are increasingly used in protein chemistry to obtain knowledge about both structure and function of proteins.

Conventionally, chemical shifts are measured and assigned in order to get the valuable nuclear Overhausen effect (NOE) and residual dipolar couplings (RDC) restraints used to determine a protein structure. NOE and RDC values relate directly to distances and relative orientations in the protein structure, and thus serve as a very powerful tool to determine the right structure.

The connection between chemical shifts and the protein structure is less straight-forward. The chemical shift depends on the shielding of an external magnetic field by the electron density around a nucleus. In other words, the wave function (and its derivatives with respect to the induced current and the nuclear magnetic moment) for a given protein structure must be known in order to know the related set of chemical shifts. Fortunately, a multitude of approximations exists, which allow the chemical shifts of a protein to be calculated with high accuracy on the time scale of milliseconds (compared several days for a gas-phase quantum mechanical calculation).

The fact that there is no clear geometric interpretation of chemical shifts makes it difficult to use as restraints to determine a protein structure. It is, however, well-known, that chemical shifts correlate with both local secondary structure as well as non-local structure. For instance, H^{α} chemical shifts are typically larger in an α-helices and smaller in β-sheets, and H^{N} that engage in short hydrogen bonds typically have large chemical shifts than if they were in a longer hydrogen bond.

Typically chemical shifts are used in protein structure determination in two different ways. (1) via an energy function that scores the agreement between predicted chemical shifts and experimental structures or (2) via a biased introduced at the conformation sampling stage.

Usually these approaches are employ Monte Carlo sampling, although Vendruscolo and co-workers have explored a chemical shift biased molecular dynamics approach.

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

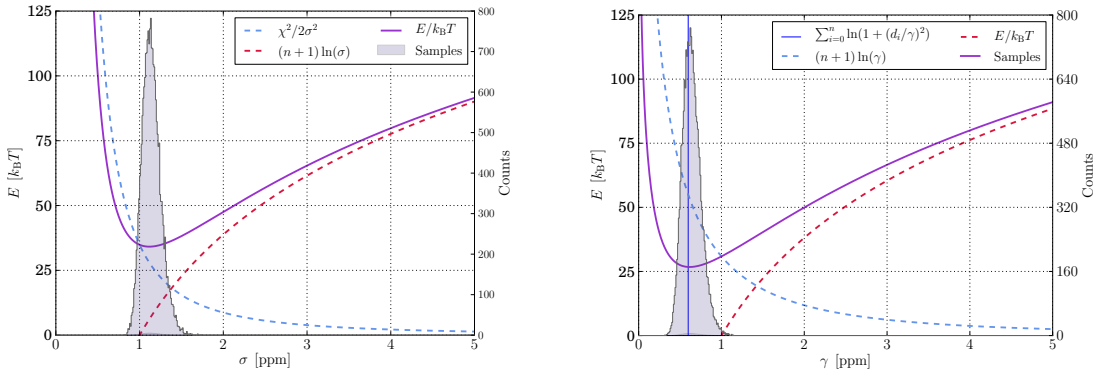
1.2 Prior distributions for protein structure

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 and TorusDBN-CS

BASILISK

Similarly to biased sampling in TorusDBN, it is possible to sample side-chain angles via BASILISK. There is currently no chemical shift dependent equivalent to BASILISK, but such a module is currently in our plans.

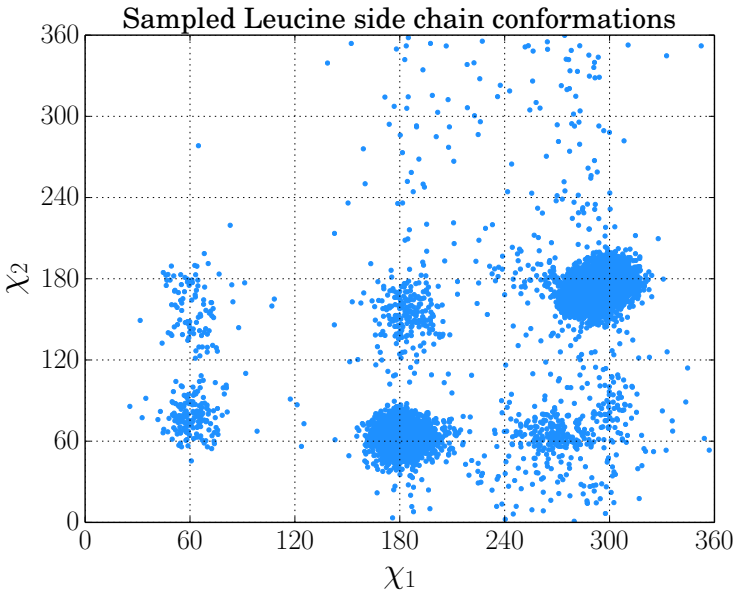


Figure 1.2: Example of Leucine rotamer conformations sampled from BASILISK. Values were sampled using the backbone conformation dependent option via the FragBuilder Python API.

Chapter 2

Graphical User Interface for PHAISTOS

Setting up simulations in PHAISTOS requires expert knowledge about the program. Firstly, while all modules and settings have reasonable default settings, there are still many things that cannot be specified via default alone, and secondly, the complete list of settings in PHAISTOS is around 2500 options that must be set or taken as default values.

In order to make PHAISTOS more available to new users, I wrote a GUI can set up most simulations for most of the simulations covered by this thesis. The GUI for PHAISTOS is aptly named Guistos. It is written in Python 2.x and

Using the GUI the user is only presented with the three most basic choices for setting up the simulation. These are (1) choice of energy terms, (2) type of Monte Carlo simulation and finally (3) a selection of Monte Carlo moves. Setting up these via Guistos is discussed next.

Energy Options

Firstly, the Energy Options section allows the user to select the molecular mechanics force field. Currently two force fields are supported in PHAISTOS, which are the OPLS-AA/L force field with a GB/SA solvent model, and the PROFASI coarse grained force field. Use of the PROFASI force field requires the Monte Carlo moves to restraint the bond angle and lengths in the protein to Engh-Huber standard values. This is automatically done if the PROFASI force field is selected. Conversely, the OPLS-AA/L force field includes energy terms for bond angles and lengths and these are degrees of freedom in the simulation if the OPLS-AA/L force field is selected.

Additionally, the Energy Options section allows the user to add restrains from one type spectroscopic data. Currently energy terms based on CamShift 1.35 and ProCS are supported. These options requires a NMR-STAR formatted file containing experimental chemical shifts.

Monte Carlo Options

This section allows the user to select the four types of Monte Carlo simulation offered by PHAISTOS and the only the most basic options to set up that particular simulation: Metropolis-Hastings offers the choice of a constant temperature (in Kelvin). Muninn and Simulated Annealing offer the choice of a temperature range (in Kelvin), and additionally Muninn offers the choice between multicanonical or $1/k$ sampling. Greedy Optimization does not offer any customizable option.

Monte Carlo Move Sets

Selecting a good mix of the different Monte Carlo moves offered by PHAISTOS can significantly speed up convergence of a simulation, compared to using an inferior move set. Choosing a good

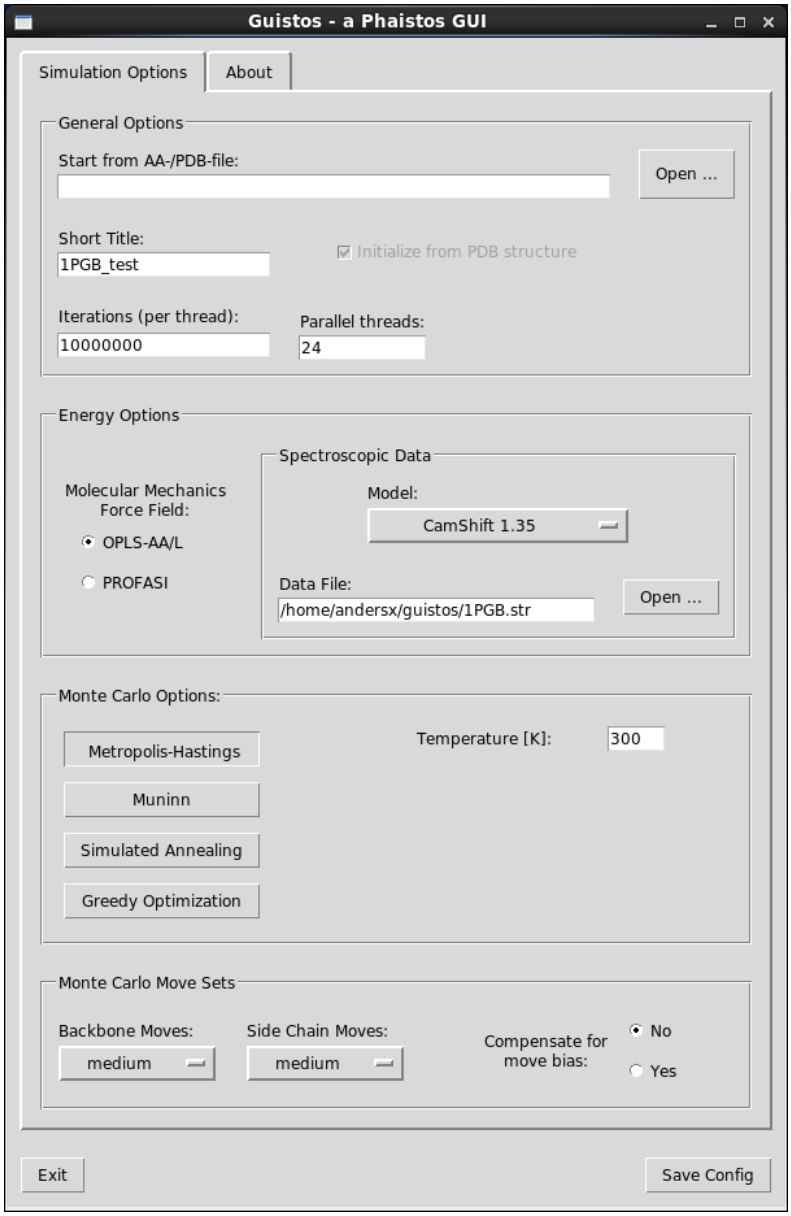


Figure 2.1: Screenshot of Guistos

set of moves is in the opinion of this author currently somewhere in between black art and sheer luck, and requires a good deal of experience with simulations in PHAISTOS.

To make it easier for new users, three move sets have been predefined using the experience of this author. These are named "small", "medium" and "large". The "small" move set is intended for uses such as refinement or sampling around a compact native state, while the "medium" move set is intended for folding simulations that start from extended, but are expected to also sample a native state, and finally the "large" move set is intended for sampling conformational space quickly, but will have problems with sampling compact structures. All move sets sample from TorusDBN (backbone angles) and BASILISK (side chain angles), and an option to remove this bias is also present.

Using Guistos

Guistos is freely released under the open source two-clause BSD-license, and can be downloaded from `https://github.com/andersx/guistos`. A screenshot of Guistos can be seen in Fig. 2.1. After specifying all relevant settings in the Guistos window, a configuration-file is saved by pressing the "Save Config" button. A simulation in PHAISTOS can the be executed via the following command:

```
1 ./phaistos --config-file my_simulation.config
```

Chapter 3

Problems

NMR based protein structure prediction has several obstacles that prevent NMR from being the go-to method in many situations.

Chapter 4

Exiting Methods

A number of methods to use chemical shifts in protein folding have been proposed. This section describes some of the most notable approaches.

4.1 ROSETTA

The ROSETTA methodology is (currently) arguably the most successful method to determine a protein structure computationally. The basis of ROSETTA is an energy function that has been demonstrated to work remarkably. Briefly described, the all-atom ROSETTA energy function consists of several additive terms such as Lennard-Jones potentials, terms for solvent exposure, hydrogen bonding, electrostatic pair-interactions and dispersion interactions, and finally torsional potentials for backbone and side chain angles. The weights between the terms are empirically optimized.

The strength of the ROSETTA energy function is that in nearly all reported cases, the experimental X-ray structure has a lower energy than any other proposed structure. The demonstrated accuracy of the energy function does come at the cost of computational speed and incomplete conformational sampling seems to be the prohibitive for further success for ROSETTA. Consequently, most publication using ROSETTA employ hundreds to thousands of cores running for several days in order to determine one structure.

Recently, the ROSETTA method has been extended to include various forms of NMR data. Chemical shifts are used to bias the fragments from which all-atom protein structures are constructed, which is the minimized through one of ROSETTA's protocols.

The largest structures determined by ROSETTA are summed up in Tab

4.2 CHESHIRE

4.3 Vendruscolo CS-MD

4.4 Meiler and Baker

4.5 Ad Bax

4.6 Conventional // CYANA, etc

4.7 *Ab initio* methods

Chapter 5

Structure Based Prediction of Protein Chemical Shifts

Chapter 6

Mass-spectroscopy in protein structure determination

This section introduces the two experimental methods, hydrogen exchange mass spectroscopy (HXMS) and cross-correlation mass spectroscopy (XCMS), and possible application in protein folding.

These experimental methods are attractive, since the experiments are relatively easy to carry out.

mass spectroscopy

6.1 Cross-linking mass spectroscopy (XLMS)

XLMS has previously been used to

We considered several linkers of different lengths. The *de facto* standard linkers, DSS and DST which measure 11 ångströ m and 6.4 ångströ m, respectively.

6.2 Hydrogen exchange mass spectroscopy (HXMS)

6.2.1 Hydrogen bond criterion

The interresidue hydrogen bond criterion of the DSSP program[Kabsch and Sander, 1983] is used to identify hydrogen bonds. The DSSP program uses an electrostatic model, assuming partial charges of -0.42 e and +0.20 e to the carbonyl oxygen and amide hydrogen respectively, and -0.42 e and +0.20 e to the carbonyl carbon and amide nitrogen, respectively. A hydrogen bond is empirically defined as having an interaction energy given as

$$E_{\mathrm{HB}} = \left(\frac{1}{r_{\mathrm{ON}}} + \frac{1}{r_{\mathrm{OH}}} - \frac{1}{r_{\mathrm{CH}}} - \frac{1}{r_{\mathrm{CN}}} \right) \cdot \ 27.89 \text{ kcal/mol} \tag{6.1}$$

stronger than a cut-off of -0.5 kcal/mol.

6.2.2 Beta-binomial model

The likelihood model that correlates a measured deuterium uptake to an integer number of hydrogen bonds in a strand is a simple model based on a beta-binomial distribution.

$$P(N_{\mathrm{HB}} = k|\alpha, \beta) = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha, \beta)} \tag{6.2}$$

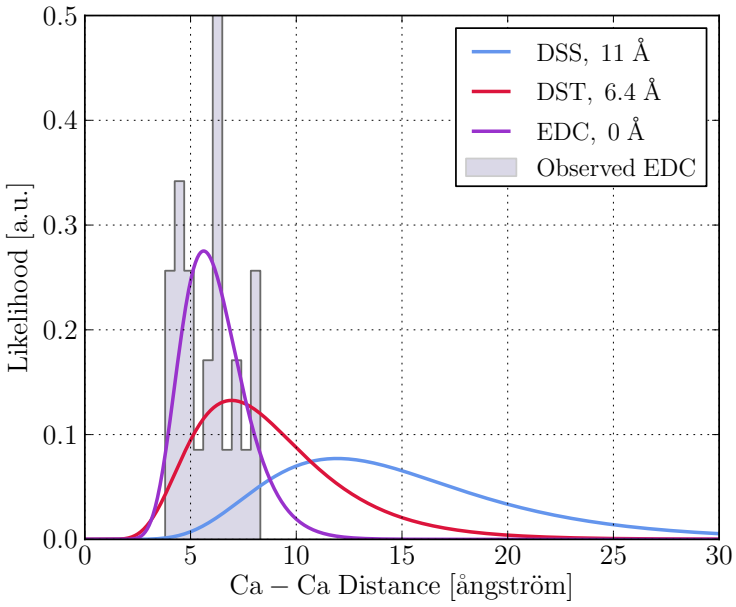


Figure 6.1: linkers

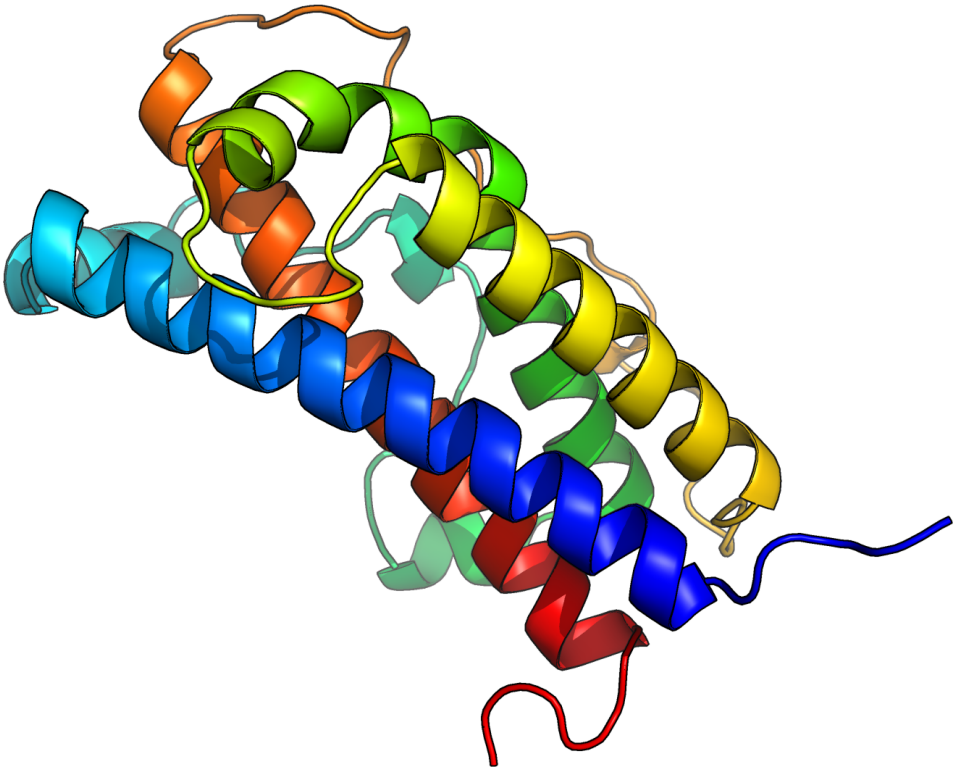


Figure 6.2: linkers

where n is the length of the strand (i.e. the maximum possible number of hydrogen bonds), N_{HB} is the number of observed hydrogen bonds and $k \in 0, ..., n$ are the possible values of N_{HB}. α and β are variables and $B(x, y)$ is the beta-function given as:

$$B(x, y) = \frac{(x-1)!(y-1)!}{(x+y-1)!} \tag{6.3}$$

The mean, μ, and variance, σ^2, and variance of the beta-binomial distribution is

$$\mu = \frac{n\alpha}{\alpha+\beta} \tag{6.4}$$

$$\sigma^2 = \frac{n\alpha\beta(n+\alpha+\beta)}{(\alpha+\beta)^2(1+\alpha+\beta)} \tag{6.5}$$

If μ and σ^2 are estimated, then α and β can then be derived for a strand of length n as:

$$\alpha = -\frac{\mu(\mu^2 - \mu n + \sigma^2)}{\sigma^2 n + \mu^2 - \mu n} \tag{6.6}$$

$$\beta = \frac{n\alpha}{\mu} - \alpha \tag{6.7}$$

6.2.3 HXMS likelihood function

The simplified forward-model that correlates the protein structure, $\mathbf{X}$ to the measured values of deuterium uptake, $\{D_i\}$ is obtained by constructing an empirical function $f(D_i) \mapsto \mu$

Chapter 7

Determined protein structures

This section describes all test-targets which I have attempted to fold using the methodologies presented in the previous chapters.

7.1 Barley Chymotrypsin Inhibitor II

An especially interesting target in this study is the barley chymotrypsin inhibitor II (CI-2). CI-2 is a 63 residue protein which consists of an α-helix which connects via a very flexible handle to a small β-sheet region.

The chemical shifts data was obtained from Kaare Theilum, who used CYANA to an automatic assignment algorithm to assign chemical shifts. This means that a very time-intensive step was skipped for this target.

7.1.1 Computational methodology

Several folding protocols were tried for this protein. All runs were performed as 72 independent trajectories which ran for 50 mio MC steps (iterations). Sampling was carried out using either TorusDBN or TorusDBN-CS to bias the backbone moves and the PROFASI force field was used in all simulations. Three simulations used an energy function based on CamShift using a cauchy distribution with variable γ value as energy function. Additionally, three simulations used a potential on the radius of gyration to restrict the sampling to only compact structures. MUNINN was was set to multicanonical sampling and the thermodynamic beta-range was set to between 0.6 and 1.1, corresponding to a temperature range of 272K to 500K. The MC moves were set to 49% CRISP moves, 2% pivot moves and 49% uniform side chain moves.

7.1.2 Folding results

Three of the 7 attempted simulation types sample structures close to the experimental X-ray structure 1YPA (here loosely defined as a CA-RMSD < 5 Å for all CA atoms. Results are summarized n thable None of the simulations that sample from TorusDBN (not chemical shift biased) are able to sample the correct fold.

Furthermore, it was noted, that simulations that sample from either TorusDBN or TorusDBN-CS with only the PROFASI force field as energy function do not generate compact structures. To overcome this deficiency, additional simulations were carried out using a radius of gyration potential. In the case of sampling from TorusDBN-CS, the radius of gyration potential is enough to get a few samples with the correct fold. Here four of 72 threads would generate the correct fold, but unfortunately the lowest energy structures were found around 8-11 Å CA-RMSD. Evidently, the PROFASI force field alone is not accurate enough to describe the native CI-2 structure. Three

Table 7.1: Protocols used in the folding of the CI-2 protein and success rates.

Sampling	Force Field	CS Energy	Correct fold[a]	Iterations/day[b]
TORUS-CS + PP[c]	PROFASI	CamShift	13	10×10^6
TORUS-CS	PROFASI	CamShift	15	11×10^6
TORUS	PROFASI	CamShift	0	11×10^6
TORUS-CS + PP[c]	PROFASI	None	4[d]	49×10^6
TORUS-CS[e] + PP[c]	PROFASI	None	0	49×10^6
TORUS-CS	PROFASI	None	0	49×10^6
TORUS	PROFASI	None	0	49×10^6

[a] Number of threads with a CA-RMSD of < 5 Å (using all residues).
[b] Numbers are *per* thread.
[c] PP denote the use of a radius of gyration potential.
[d] Structures with the lowest energy did not correspond to the native structure in this run.
[e] This run was carried out using TorusDBN-CS trained using only high-quality X-ray structures.

simulations were performed with an energy term based on CamShift in addition the PROFASI force field. The increased accuracy in the energy function causes increased sampling around the native state.

Due to a very flexible region of CI-2 (residues 33 to 42), and somewhat flexible tails the residue range used to calculate CA-RMSD values is restricted to residue 12-32,43-52 in the following.

Computational efficiency

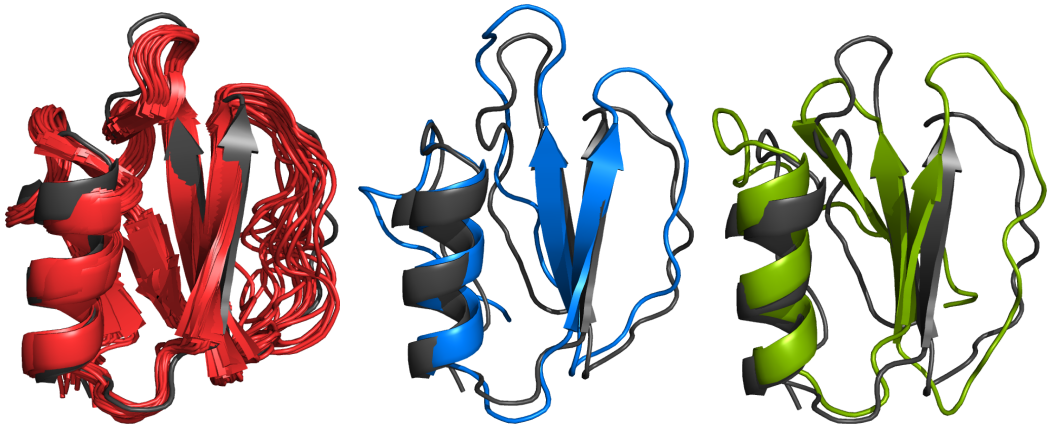
All runs were carried out on 3 24-core AMD Opteron 6172 servers running at 2.1 GHz. A run similar to the most successful was also run carried out on a faster a 12-core Intel X5675 node running at 3.07 GHz (using new random seeds). This simulation took two days, with a total of 2 out of 12 threads successfully identifying the native structure as having the lowest energy.

7.2 Folding of small proteins (<100 AA)

Table 7.2: Folded structure.

Name	Lengh	Type	PDB	RefDB	RMSD-range	Final RMSD
Protein G	56	a/b	2OED	2575	All	?
SMN Tudor Domain	59	a/b	1MHN	4899	5-54	?
Engrailed Homeodomain	61	B	1ENH	15536	8-53	?
CI-2	63	a/b	1YPA	N/A[a]	4-34,43-63	?
FF Domain	71	a/b	1UZC	5537	11-67	?
Ubiquitin	76	a/b	1UBI	17769	1-70	?

[a] Using automatically assigned data obtained from Kaare Theilum (personal communitcation - see `https://github.com/andersx/cs-proteins`).



(a) NMR structure (red) (b) Lowest RMSD structure (blue) (c) Lowest energy structure (green)

Figure 7.1: Structures compared to the X-ray structure 1YPA. All structures are aligned using the residues 12-32,43-52. (a) shows the 3CI2 structure NNR structure. Note the flexible domain which is excluded from the fit-range. (b) Shows the lowest RMSD structure (1.113 ÅRMSD). (c) shows the lowest energy sample (2.76 ÅRMSD).

7.3 Folding of larger proteins (>100 AA)

It is worth to note, that using sparse NMR data, only three structures >200 residues have been determined: Alg13 (201 AA), Rhodopsin (225 AA) and MBP (376 AA) using the Rosetta program with the "resolution-adapted structural recombination" (RASREC) protocol

Alg13 was solved using backbone chemical shifts only 52 NOE, to an CA-RMSD of 4 Å to the experimental NMR structure (2jzc). Rhodopsin was folded to an CA-RMSD of 1.6 Å to the X-ray structure using 215 NOE restraints, backbone chemical shifts chemical shifts and RDCs. The MBP protein is a two-domain protein of 376 residues. MBP was folded to an RMSD of 3.6 Å using 1235 NOE restraints, backbone chemical shifts chemical shifts and RDCs. The NOEs corresponded to 55% yield of restraints, which, however mostly were not automatically assigned. An attempt to use only automatically assigned NOEs yielded 455 restraints, which corresponds to a yield of 20%. Using these, however, the MBP structure could only be determined to a total CA-RMSD of 12.3 Å. The N-terminal domain was converged to 2.7 Å, but the C-terminal domain and the angle between the two domains was incorrectly folded.

7.3.1 Folding protocol

This section presents folding results on a set of larger proteins (>100 AA) with known structures.

The protocol which folded Rhodopsin and Prolactin can be executed from PHAISTOS via the following commands. The third refinement stage is the same as the initial folding stage, only with 5,000,000 iterations and 8 threads.

Initial folding stage:

```
./phaistos --aa-file rhodopsin.aa \
  --iterations 50000000 \
  --threads 72 \
  --monte-carlo-muninn 1 \
  --monte-carlo-muninn-min-beta 0.6 \
  --monte-carlo-muninn-max-beta 1.1 \
```

```
  --monte-carlo-muninn-independent-threads 1 \
  --monte-carlo-muninn-weight-scheme multicanonical \
  --backbone-dbn-torus-cs 1 \
  --backbone-dbn-torus-cs-initial-nmr-star-filename \
                                      rhodopsin.str \
  --energy-profasi-cached 1 \
  --energy-isd-dist 1 \
  --energy-isd-dist-likelihood square_well \
  --energy-isd-dist-data-filename noe_ilv.txt \
  --energy-isd-dist-sample-gamme 0 \
  --energy-isd-dist-sample-sigma 0 \
  --energy-isd-dist-weight 0.0078125 \
  --move-backbone-dbn 1 \
  --move-backbone-dbn-weight 0.08 \
  --move-backbone-dbn-implicit-energy 1 \
  --move-crisp-dbn-eh 1 \
  --move-crisp-dbn-eh-weight 0.42 \
  --move-sidechain-uniform 1 \
  --move-sidechain-uniform-weight 0.5
```

Relaxation stage:

```
./phaistos --pdb-file rhodopsin_lowest_energy1.pdb \
  --init-from-pdb 1 \
  --iterations 2000000 \
  --threads 8 \
  --monte-carlo-metropolis-hastings 1 \
  --monte-carlo-metropolis-hastings-declash-on-reinitialize 0 \
  --backbone-dbn-torus-cs 1 \
  --backbone-dbn-torus-cs-initial-nmr-star-filename \
                                      rhodopsin.str \
  --energy-profasi-cached 1 \
  --energy-pp-compactness 1 \
  --move-backbone-dbn 1 \
  --move-backbone-dbn-weight 0.08 \
  --move-backbone-dbn-implicit-energy 1 \
  --move-crisp-dbn-eh 1 \
  --move-crisp-dbn-eh-weight 0.42 \
  --move-sidechain-uniform 1 \
  --move-sidechain-uniform-weight 0.5
```

7.3.2 Rhodopsin (225 residues)

Since the ILV(W) data set used by Rosetta was not available, simulated ILV restraints from the PDB structure 1h68 was used instead. A very conservative simulation of only 63 synthetic NOE restraints was used, which corresponds to around 4% assigned long-range contacts. The initial folding stage converged to around 6.8 as the lowest energy cluster - see Fig. 7.2). Two threads out of 72 converged to this native-like cluster.

The refinement stage with only 63 NOEs, however, did not converge to a lower CA-RMSD. The Rosetta data set includes 213 contacts, and a similar, less conservative synthetic NOE data was simulated from the PDB structure 1h68, in order to see, if an increased number of NOE restraints could drive the RMSD down. This resulted in 195 NOE restraints, which corresponds to 13% assigned NOEs. Using this, larger, set of restraints during the refinements gave a substantial decrease in CA-RMSD for the lowest energy sample to 3.6 Å.

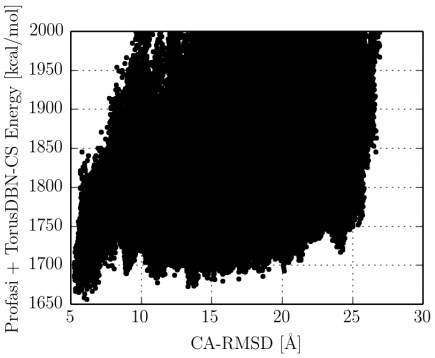
It must be noted, however, that in both the folding and refinment stages, samples are obtained with less than 2650 kcal/mol (calculated as the Profasi force field energy plus the likelihood from Torus-DBN-CS given the experimental chemical shifts), while the native energy is 2730 kcal/mol. This discrepancy is likely due the lack of accurate non-local information in the Profasi and Torus-DBN-CS models.

Further refinement was attempted with the CamShift energy-term instead of the Torus-DBN-CS model, but was found to be too slow to be practically feasible.

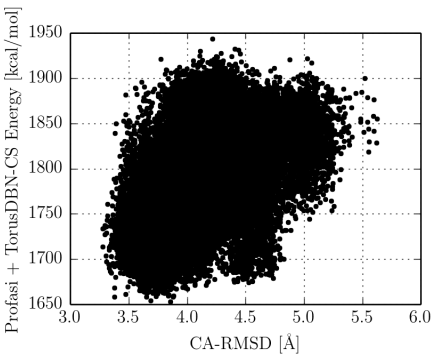
Table 7.3: Folded structure.

Name	Lengh	Type	PDB	BMRB	RMSD-range	RMSD
APO-LFABP	129	a/b	1LFO	15429[a]	All	
Prolactin	199	B	1RWS	5599	6-183	
Top7	120	a/b	2MBL	19404	5-104	
MSRB	151	a/b	3E0O	17008	36-105	
WR73	183	a/b	2LOY	16833	1-36,66-181	
HR4660B	174	a/b	2LMD	1870	16-162	
Rhodopsin	219	B	2KSY	16678	All	3.2

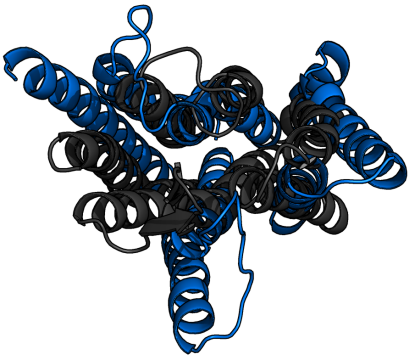
[a] Data was obtained from RefDB.



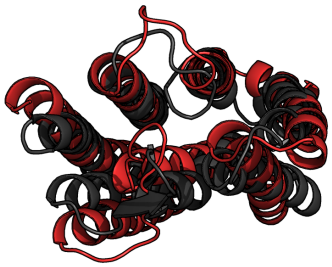
(a) Energy-scoring during folding stage.



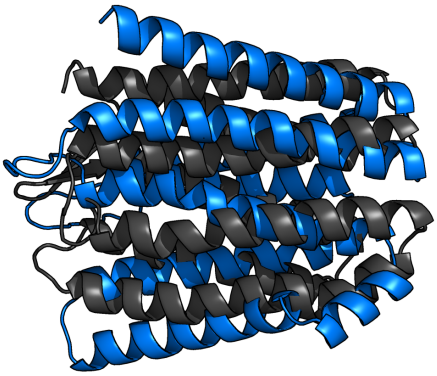
(b) Energy-scoring during refinement stage.



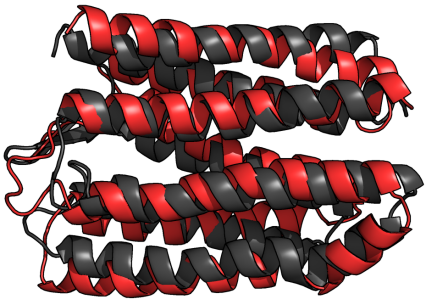
(c) Bottom view of folding stage lowest energy sample (blue).



(d) Bottom view of refinement stage lowest energy sample (red).



(e) Side view of folding stage lowest energy sample (blue).



(f) Side view of refinement stage lowest energy sample (red).

Figure 7.2: Some caption

Bibliography

[Kabsch and Sander, 1983] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.

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