



# 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. *Share-Latex* (unpublished).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

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 Overhauser 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. 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,  $H^\alpha$  chemical shifts are typically larger in an  $\alpha$ -helices and smaller in  $\beta$ -sheets.

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).

# Contents

<b>Acknowledgement</b>	<b>i</b>
<b>Publication list</b>	<b>ii</b>
<b>Preface</b>	<b>iv</b>
<b>1 Chemical shifts in a probabilistic framework</b>	<b>2</b>
1.1 Defining an energy function from Bayes' theorem . . . . .	2
1.1.1 Jeffreys' prior (general, one-parameter case) . . . . .	4
1.1.2 Jeffreys' prior (Gaussian and Cauchy distributions) . . . . .	4
1.1.3 Sampling of nuisance parameters . . . . .	5
1.2 Prior distributions for protein structure . . . . .	5
1.2.1 Molecular mechanics force field . . . . .	5
1.2.2 Generative probabilistic models . . . . .	6
<b>2 Exiting Methods</b>	<b>8</b>
2.1 CS-ROSETTA . . . . .	8
2.2 CHESHIRE . . . . .	8
2.3 Vendruscolo CS-MD . . . . .	8
2.4 Meiler and Baker . . . . .	8
2.5 Ad Bax . . . . .	8
2.6 Conventional // CYANA, etc . . . . .	8
<b>3 Structure Based Prediction of Protein Chemical Shifts</b>	<b>9</b>
<b>4 Mass-spectroscopy in protein structure determination</b>	<b>10</b>
4.1 Cross-linking mass spectroscopy (XLMS) . . . . .	10
4.2 Hydrogen exchange mass spectroscopy (HXMS) . . . . .	10
4.2.1 Hydrogen bond criterion . . . . .	10
4.2.2 Beta-binomial model . . . . .	10
4.2.3 HXMS likelihood function . . . . .	12

# 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

$$E_{\text{hybrid}} = w_{\text{data}} E_{\text{data}} + E_{\text{physical}}. \quad (1.1)$$

This approach, however, does not uniquely define neither shape nor weight of  $E_{\text{data}}$ . 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  $E_{\text{data}}$  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:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \quad (1.2)$$

Now consider a set of chemical shifts  $\{\delta_i\}$ , and the uncertainty to which these can be predicted  $\{\sigma_i\}$  from a structure,  $\mathbf{X}$  (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,  $\mathbf{X}$ , and optimal choice of  $\{\sigma_i\}$  is found by maximizing (via Bayes' theorem)

$$p(\mathbf{X}, \{\sigma_i\} | \{\delta_i\}) = \frac{p(\{\delta_i\} | \mathbf{X}, \{\sigma_i\}) p(\mathbf{X}, \{\sigma_i\})}{p(\{\delta_i\})}. \quad (1.3)$$

Here, *marginal distribution* of  $p(\{\delta_i\})$  merely serves as a normalizing factor and the *likelihood* of  $p(\{\delta_i\} | \mathbf{X}, \{\sigma_i\})$ , 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.  $\text{C}^\alpha$ ,  $\text{H}^\alpha$ , etc.) are assumed to carry the same uncertainty denoted by  $\sigma_j$ :

$$p(\{\delta_i\} | \mathbf{X}, \{\sigma_i\}) \simeq \prod_{i=0}^n p(\Delta\delta_i | \mathbf{X}, \sigma_i) \quad (1.4)$$

$$= \prod_{j=0}^m \prod_{i_j=0}^{n_j} p(\Delta\delta_{i_j} | \mathbf{X}, \sigma_j) \quad (1.5)$$

$$= \prod_{j=0}^m \prod_{i_j=0}^{n_j} \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) \quad (1.6)$$

$$= \prod_{j=0}^m \left(\frac{1}{\sigma_j \sqrt{2\pi}}\right)^{n_j} \exp\left(\sum_{i_j=0}^{n_j} -\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) \quad (1.7)$$

$$= \prod_{j=0}^m \left(\frac{1}{\sigma_j \sqrt{2\pi}}\right)^{n_j} \exp\left(\frac{-\chi_j^2(\mathbf{X})}{2\sigma_j^2}\right) \quad (1.8)$$

Furthermore,  $p(\mathbf{X}, \{\sigma_j\})$  can be simplified as

$$p(\mathbf{X}, \{\sigma_j\}) \propto p(\{\sigma_j\} | \mathbf{X}) p(\mathbf{X}) \quad (1.9)$$

$$= p(\{\sigma_j\}) p(\mathbf{X}), \quad (1.10)$$

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).  $p(\mathbf{X})$  of the molecular protein structure is here simply the Boltzmann distribution, i.e.

$$p(\mathbf{X}) = \frac{1}{Z} \exp\left(-\frac{E(\mathbf{X})}{k_B T}\right) \quad (1.11)$$

where  $E(\mathbf{X})$  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  $p(\mathbf{X})$  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).



## 1.1. DEFINING AN ENERGY FUNCTION FROM BAYES' THEOREM

---

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

$$p(\mathbf{X}, \{\sigma_i\} | \{\delta_i\}) \propto p(\{\delta_i\} | \mathbf{X}, \{\sigma_i\}) p(\mathbf{X}) p(\{\sigma_i\}) \quad (1.12)$$

$$\propto \prod_{j=0}^m \left( \frac{1}{\sigma_j \sqrt{2\pi}} \right)^{n_j} \exp \left( -\frac{1}{2\sigma_j^2} \chi_j^2 \right) \exp \left( -\frac{E(\mathbf{X})}{k_B T} \right) p(\{\sigma_j\}) \quad (1.13)$$

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

$$E_{\text{hybrid}}(\mathbf{X}) = -k_B T \ln \left( p(\mathbf{X}, \{\sigma_i\} | \{\delta_i\}) \right) \quad (1.14)$$

$$= E(\mathbf{X}) - k_B T \sum_{j=0}^{n_j} n_j \ln \left( \frac{1}{\sigma_j \sqrt{2\pi}} \right) + \frac{\chi_j^2}{2\sigma_j^2} \quad (1.15)$$

### 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

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

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

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

### 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

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

This immediately gives us the Jeffreys' prior:

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

Similarly for the  $\gamma$  parameter of the Cauchy distribution of the form

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

we obtain the following Jeffreys' prior:

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

### 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:

$$p(\theta \rightarrow \theta') = p(\theta' \rightarrow \theta) \quad (1.22)$$

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,  $\gamma$  and  $\sigma$ , 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  $\gamma$  and  $\sigma$  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.:

$$p(\mathbf{X}) \propto \exp\left(\frac{-E}{k_B T}\right) \quad (1.23)$$

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.

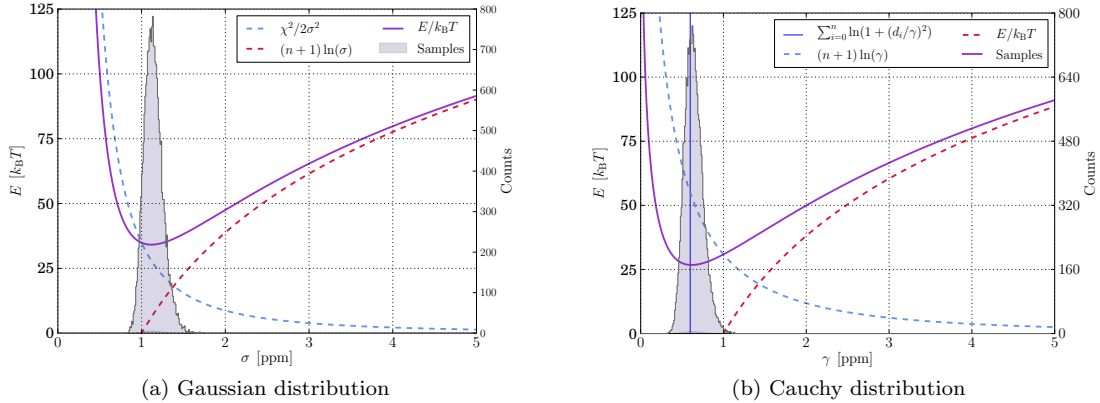


Figure 1.1: Sampling of  $\sigma$  and  $\gamma$  for 2OED for Ca-chemical shifts.  $n = 55$  and  $\chi^2 = 69.7$ .

### 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  $(\phi, \psi)$  backbone angles uniformly (i.e. in the range  $[-180^\circ, 180^\circ]$  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  $(\phi, \psi)$  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  $(\phi, \psi)$  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.

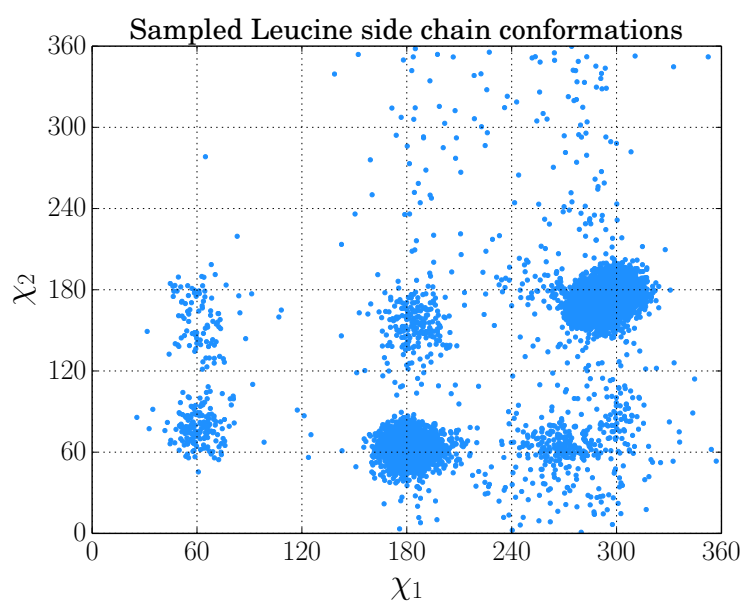


Figure 1.2: Example of Leucine rotamer conformations sampled from BASILISK. Values were sampled using the backbone conformation dependent option via the FragBuilder Python API.

## Chapter 2

# Exiting Methods

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

### 2.1 CS-ROSETTA

The CS-ROSETTA is so far the most successful ...

### 2.2 CHESHIRE

### 2.3 Vendruscolo CS-MD

### 2.4 Meiler and Baker

### 2.5 Ad Bax

### 2.6 Conventional // CYANA, etc

## Chapter 3

# Structure Based Prediction of Protein Chemical Shifts

## Chapter 4

# 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

### 4.1 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.

### 4.2 Hydrogen exchange mass spectroscopy (HXMS)

#### 4.2.1 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

$$E_{\text{HB}} = \left( \frac{1}{r_{\text{ON}}} + \frac{1}{r_{\text{OH}}} - \frac{1}{r_{\text{CH}}} - \frac{1}{r_{\text{CN}}} \right) \cdot 27.89 \text{ kcal/mol} \quad (4.1)$$

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

#### 4.2.2 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.

$$P(N_{\text{HB}} = k | \alpha, \beta) = \binom{n}{k} \frac{B(k + \alpha, n - k + \beta)}{B(\alpha, \beta)} \quad (4.2)$$

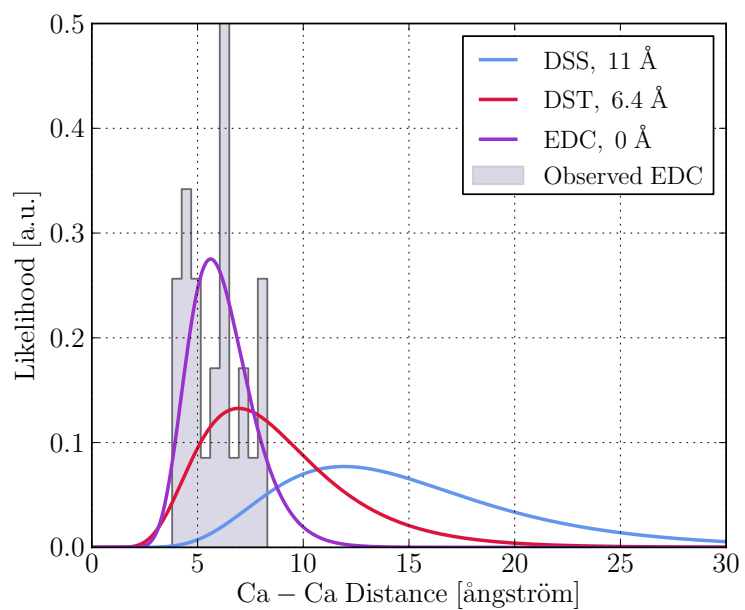


Figure 4.1: linkers

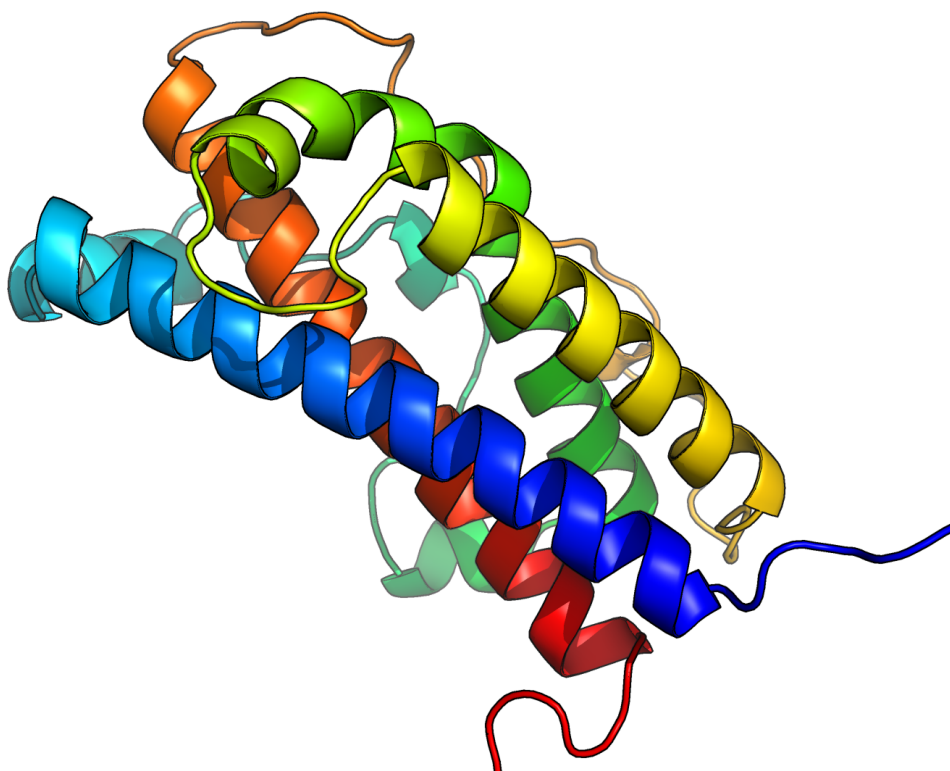


Figure 4.2: linkers



## 4.2. HYDROGEN EXCHANGE MASS SPECTROSCOPY (HXMS)

---

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

$$B(x, y) = \frac{(x-1)!(y-1)!}{(x+y-1)!} \quad (4.3)$$

The mean,  $\mu$ , and variance,  $\sigma^2$ , and variance of the beta-binomial distribution is

$$\mu = \frac{n\alpha}{\alpha + \beta} \quad (4.4)$$

$$\sigma^2 = \frac{n\alpha\beta(n + \alpha + \beta)}{(\alpha + \beta)^2(1 + \alpha + \beta)} \quad (4.5)$$

If  $\mu$  and  $\sigma^2$  are estimated, then  $\alpha$  and  $\beta$  can then be derived for a strand of length  $n$  as:

$$\alpha = -\frac{\mu(\mu^2 - \mu n + \sigma^2)}{\sigma^2 n + \mu^2 - \mu n} \quad (4.6)$$

$$\beta = \frac{n\alpha}{\mu} - \alpha \quad (4.7)$$

### 4.2.3 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  $f(D_i) \mapsto \mu$

# 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.