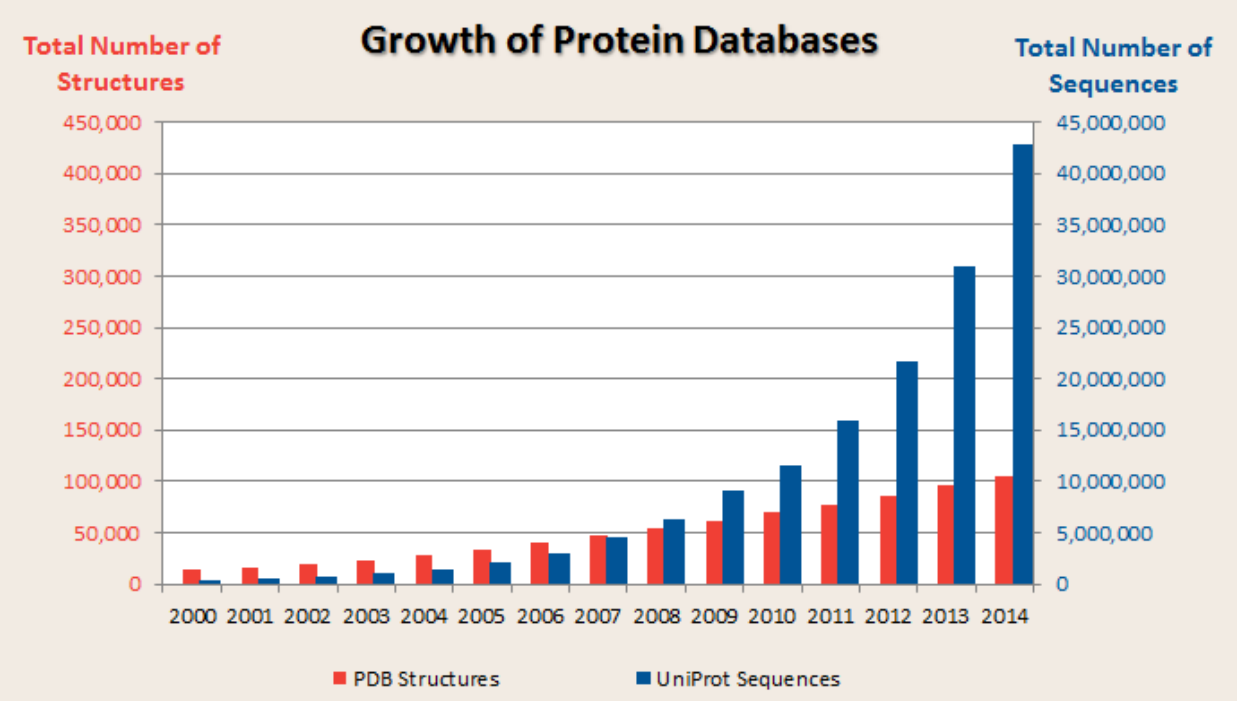
Potentials of mean force for protein structure prediction: from hack to math

Bayes@Lund 2017, April 20th

Thomas Hamelryck

Statistical Structural Biology Group
Bioinformatics center
Departments of Biology / Computer science
University of Copenhagen
Denmark

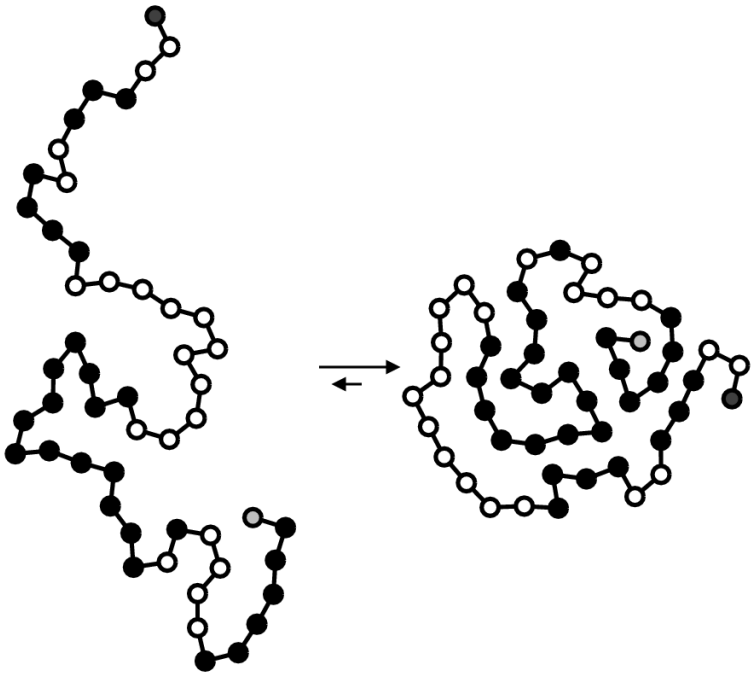
Proteins in the post-genomic wild west



- Sequences spectacularly outnumbers structures[1].
- The quest for methods to routinely predict, simulate or design protein **structure**, **dynamics** and **interactions** continues.

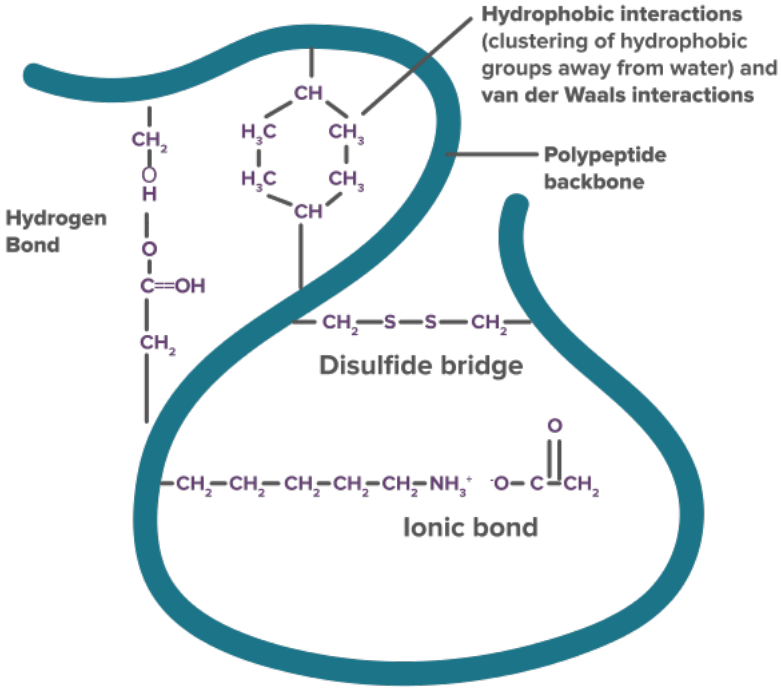
[1] Picture: https://www.dnastar.com

Today's menu - some protein with maths



This talk concerns *"the* ***extraction of a force field*** *from a data base of known 3D structures, which reasonably models the protein-solvent system"* (Sippl, 1993). Such force fields or energy functions are also important for modelling protein interactions.

Protein folding and its forces

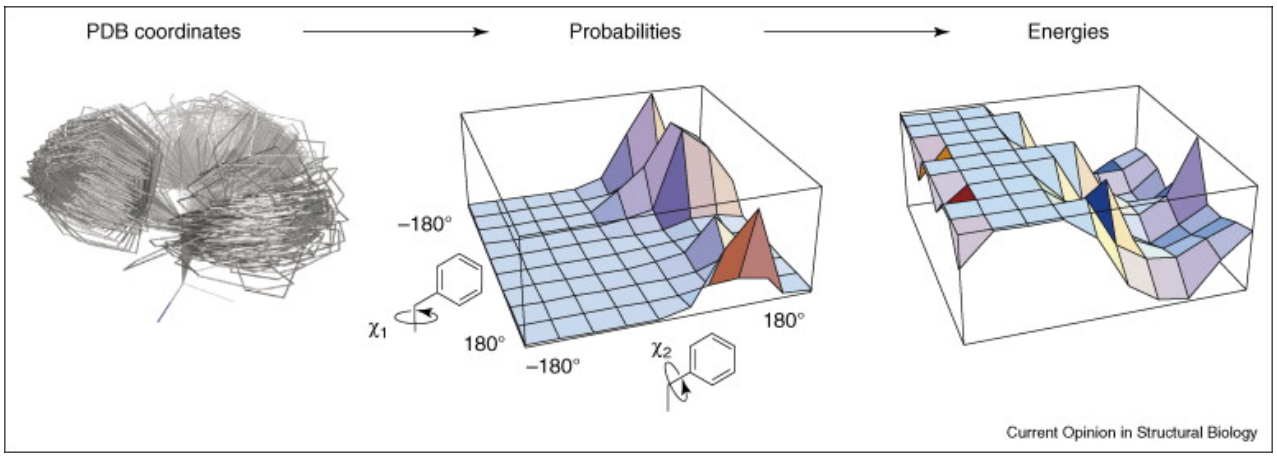


- Electrostatic, hydrogen bonding, hydrophobic, van der Waals and repulsive forces shape proteins into their 3-D folds[2].
- How can we derive information on these energies from the set of known protein structures in a well-defined way?

[2] Picture: https://www.khanacademy.org

Knowledge-based energy functions

- Energy functions range from quantum mechanics over Newtonian physics to statistical approaches[3] – called **knowledge based energies**.



- Knowledge based energies are attractive because they can be **efficiently applied to simplified representations of proteins**.
- They aim to approximate the **free energy**.

[3] Picture: Boas & Harbury, Curr. Opinion Struct. Biol., 2007

Turning probabilities into energies

- **Boltzmann's law** turns the energy e_x of a microstate x into its probability p_x,

$$p_x = \frac{1}{Z} \exp\left(\frac{-e_x}{kT}\right),$$

with k the Boltzmann's constant, T the absolute temperature and Z a normalization factor (*Zustandssumme*).

- Hence, the **inverse of Boltzmann's law** turns probabilities into energies,

$$e_x = -kT \log(p_x) - kT \log(Z).$$

- The main problem is that Z cannot be calculated for most practical cases.

Manfred Sippl's bright idea (1990)

- Sippl starts with the inverse of Boltzmann's law,

$$e_x = -kT \log(p_x) - kT \log(Z)$$

and subtracts a so-called **reference energy** $\mathcal{E}_x$,

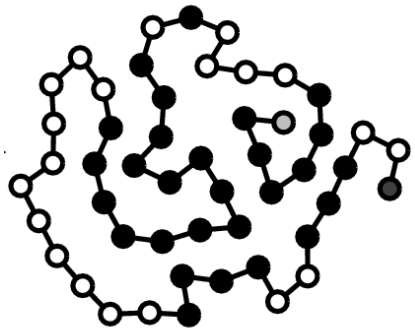
$$\mathcal{E}_x = -kT \log(\mathcal{P}_x) - kT \log(\mathcal{Z}),$$

where $\mathcal{P}_x$ is the probability of microstate x and $\mathcal{Z}$ is the normalisation factor according to a certain **reference state**,

$$\Delta e_x = e_x - \mathcal{E}_x = -kT \log\left(\frac{p_x}{\mathcal{P}_x}\right) - kT \log\left(\frac{Z}{\mathcal{Z}}\right).$$

- Now, if we assume that $\frac{Z}{\mathcal{Z}} \approx 1$, the second term disappears.

Sippl's knowledge based energy for proteins



- Following Sippl, the total energy of a protein conformation is,

$$\Delta e = \sum_x -kT \log \left(\frac{p_x}{\mathcal{P}_x} \right),$$

  where the sum runs over all **pairwise distances**.
- The reference state is some **random packing** of amino acids.
- For the 2D toy protein, $p_x = p(\text{distance } r \mid \text{colors involved})$.

Justification by fuzzy analogy

- Sippl's approach works, but why does it work? Where does the reference energy come from? Why is the subtraction needed?

Reversible Work Theorem for liquids

- The reversible work required to bring two liquid particles from infinite separation to a distance r from each other is

$$W_r = -kT \log \left(\frac{p_r}{\mathcal{P}_r} \right)$$

  where p_r and $\mathcal{P}_r$ are the probabilities of finding two particles at a distance r in the liquid and the reference state.
- The reference state is precisely defined as the **ideal gas state**, consisting of non-interacting particles.
- W_r is a so-called **potential of mean force**.

Sippl in trouble

- Sippl's idea is based on some shaky assumptions.

1. Applying **Boltzmann's law** is unwarranted.
   - The probablities do not come from a single protein's Boltzmann distribution, but from **structures of many different proteins** from the Protein Data Bank (PDB).
2. Calling these energies **potentials of mean force** based on a vague analogy with some physics of liquids is unwarrenated.
   - It's not clear at all what to use as **reference state**.
   - People hack around and use what works.
3. Even if Sippl's energy was a true potential of mean force, it would not be the desired **free energy**.

- So why does Sippl's **hack** actually work, and is it optimal? To answer that question, we will need to turn to the world of **Bayesian probability**.

Bayesian probability – a primer

- Bayesian reasoning consists of **updating a current belief** on parameter θ in the light of new data d.
- The current belief is quantified as the **prior distribution** $\pi(\theta)$.
- The data is quantified as the **likelihood** $p(d \mid \theta)$.
- The updated belief is quantified as the **posterior distribution**

$$p(\theta \mid d) \propto p(d \mid \theta) \times \pi(\theta)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Bayes in trouble

- Sometimes, classic Bayesian updating is not applicable, notably if we do not obtain data that can be related to a parameter, but data on the **changed probability** of that parameter.

Example

- Suppose we obtain data **d** concerning a parameter θ. Using the standard Bayesian calculus, we update the prior over θ by multiplication with the likelihood of **d**,

$$p(\theta \mid \mathbf{d}) \propto p(\mathbf{d} \mid \theta) \times \pi(\theta).$$

- Suppose we are given information $\mathcal{I} \equiv p(\theta > 0) = \varepsilon$ instead.
  - How do we update $\pi(\theta)$?

$$p(\theta \mid \mathcal{I}) \propto \ ? \times \pi(\theta).$$

Whitworth's horses (1901)

- Suppose four horses A, B, C, D have equal probability of winning in a race ($p = 0.25$).
- However, the probability of A winning is updated to 0.4.
- How can the individual probabilities of B, C or D winning be updated?
  - The probability of B, C **or** D winning is $1 - 0.4 = 0.6$
  - Assuming they still have equal probability of winning, we obtain

$$p(\mathrm{B\ wins}) = p(\mathrm{C\ wins}) = p(\mathrm{D\ wins}) = 0.6/3 = 0.2$$

Whitworth's horses and partitions

- Whitworth's problem requires updating of a prior distribution based on new information **on a partition** with two elements:
  1. {A wins}
  2. {B wins, C wins, D wins} $\equiv$ A loses.
  - The probabilities changed from $(0.25, 0.75)$ to $(0.4, 0.6)$.
- We assume that the **conditional probabilities** of B, C, D winning remain the same

$$
\begin{aligned}
p(\text{A loses}) &\rightarrow \text{changes from } 0.75 \text{ to } 0.6 \\
p(\text{B wins} \mid \text{A loses}) &\rightarrow \text{remains the same at } 1/3 \\
\Rightarrow p(\text{B wins} \mid \text{A loses}) p(\text{A loses}) &= \frac{0.6}{3} = 0.2
\end{aligned}
$$

  - In other words, the relative probabilities of the elements **within the partitions** remain the same.

Jeffrey's conditioning

- **Jeffrey's conditioning** or **probability kinematics** allows Bayesian updating of a prior $\pi(\theta)$ given new information on a partition $E = \{E_1, \ldots, E_n\}$ of the support Ω_θ.

Jeffrey's conditioning

- We have a prior distribution $\pi(\theta)$ with matching $\pi(E_i)$.
- We obtain new information on θ in the form of updated probabilities $p(E_i)$. How do we update $\pi(\theta)$ to $p(\theta)$?
- If we assume the conditional probabilities remain the same, that is $\pi(\theta \mid E_i) = p(\theta \mid E_i)$ for all (i, θ), then

$$\begin{aligned} \pi(\theta) &= \pi(\theta \mid E_\theta)\pi(E_\theta) \\ \Rightarrow p(\theta) &= \pi(\theta \mid E_\theta)p(E_\theta) \end{aligned}$$

with E_θ being the partition that contains θ.

Jeffrey reframed - The reference ratio

- Jeffrey's conditioning can be reformulated by a simple application of Bayes's theorem[4]. This is convenient if $p(\theta \mid E_\theta)$ is not available. It will also shed light on Sippl's energy.

The reference ratio formulation

- We start with the usual formulation of Jeffrey's conditioning

$$p(\theta) = \pi(\theta \mid E_\theta) p(E_\theta)$$

and apply Bayes' theorem to the first factor

$$\begin{aligned} p(\theta) &= \frac{\pi(E_\theta \mid \theta)\pi(\theta)}{\pi(E_\theta)} p(E_\theta) \\ \Rightarrow p(\theta) &= \frac{p(E_\theta)}{\pi(E_\theta)} \pi(\theta) \end{aligned}$$

[4] Hamelryck *et al.*, PLoS ONE, 2010

Back to the races

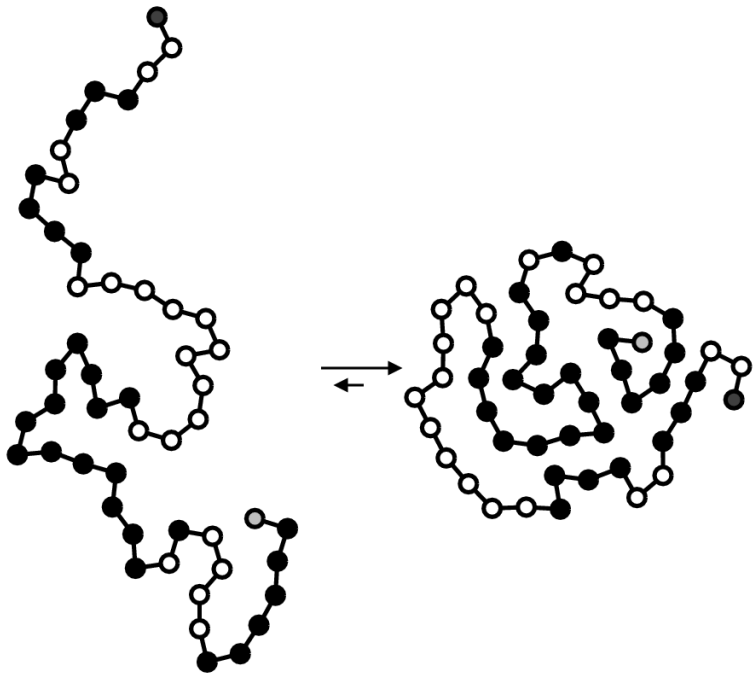
- The reference ratio formulation of Jeffrey's conditioning is

$$p(\theta) = \frac{p(E_\theta)}{\pi(E_\theta)}\pi(\theta).$$

- Applied to Whitworth's horses, this becomes

$$\begin{aligned} p(\text{B wins}) &= \frac{p(\text{A loses})}{\pi(\text{A loses})}\pi(\text{B wins}) \\ &= \frac{0.6}{0.75} \times 0.25 = 0.2 \end{aligned}$$

A local model of protein structure as prior



- Suppose we have a prior distribution $\pi(\mathbf{x})$ over the **backbone angles x** of a protein, and that this distribution is only valid on a **local length scale**.
  - Baker's ROSETTA program pioneered the use of a **fragment library** as $\pi(\mathbf{x})$.

Adding a global model using Jeffrey's trick

- The **local** model could be **salvaged** by adding a second model with **global** information
  - This amounts to a **multiscale modelling** approach.
- As global model, we use a probability distribution $p(\mathbf{d})$ over the **pairwise distances d**.
  - We can assume $\mathbf{d} = f(\mathbf{x})$, that is, if we know the angles we can calculate the pairwise distances.
- $\mathbf{d} = f(\mathbf{x})$ is many-to-one, and thus $\mathbf{d}$ induces a **partition** on $\Omega_{\mathbf{x}}$.
- Thus, we can combine the local with the global model using Jeffrey's conditioning,

$$p(\mathbf{x}) = \frac{p(\mathbf{d})}{\pi(\mathbf{d})}\pi(\mathbf{x}),$$

following the **reference ratio formulation**.

Sippl's energy explained

- The probabilistic model of protein structure we obtained is

$$p(\mathbf{x}) = \frac{p(\mathbf{d})}{\pi(\mathbf{d})}\pi(\mathbf{x}).$$

- If we formulate this model **in terms of energies**, using a minus-log transformation, we get

$$e(\mathbf{x}) = -kT \log\left(\frac{p(\mathbf{d})}{\pi(\mathbf{d})}\right) - kT \log(\pi(\mathbf{x}))$$

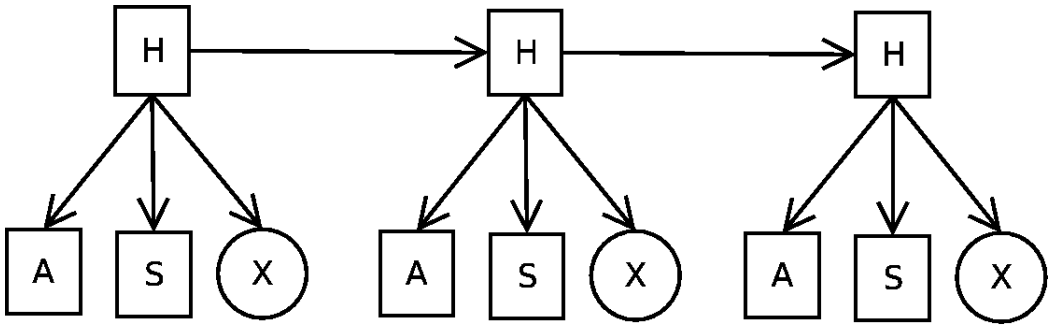
  which leads us to **Sippl's "potential of mean force"**.
  - The **reference distribution** is defined by the local model $\pi(\mathbf{x})$.
  - The last term is usually "invisible" because it is brought in by sampling, ie. using a fragment library.
- Thus, Sippl's "potentials of mean force" can be understood as an approximation of Jeffrey's conditioning.

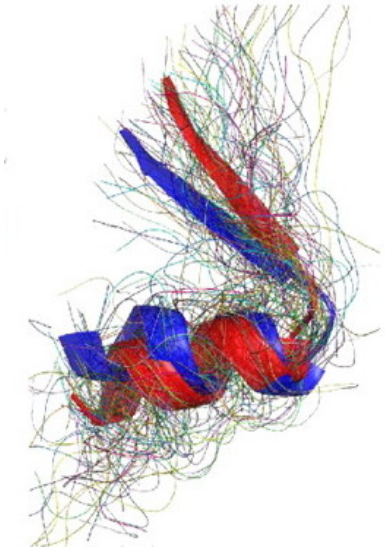
A Bayesian model of protein structure

- Probability kinematics allows to formulate an efficient **divide-and-conquer strategy** for Bayesian protein structure prediction, as follows:
- Estimate a model $\pi(\mathbf{x} \mid \mathbf{a})$ that covers local protein structure
  - $\mathbf{x}$=sequence of dihedral angles, $\mathbf{a}$=amino acid sequence
  - This model is **high-dimensional and detailed** but not accurate on the global scale.
- Estimate a model $p(\mathbf{y} \mid \mathbf{a})$ that covers nonlocal protein structure
  - With $\mathbf{y} = f(\mathbf{x})$, that is, $\mathbf{y}$ is a **low-dimensional, many-to-one** deterministic function of $\mathbf{x}$
  - This model is accurate on the global scale but without detail.
- Tie the two together using probability kinematics

$$p(\mathbf{x} \mid \mathbf{a}) = \frac{p(\mathbf{y} \mid \mathbf{a})}{\pi(\mathbf{y} \mid \mathbf{a})} \pi(\mathbf{x} \mid \mathbf{a})$$

Model of local protein structure

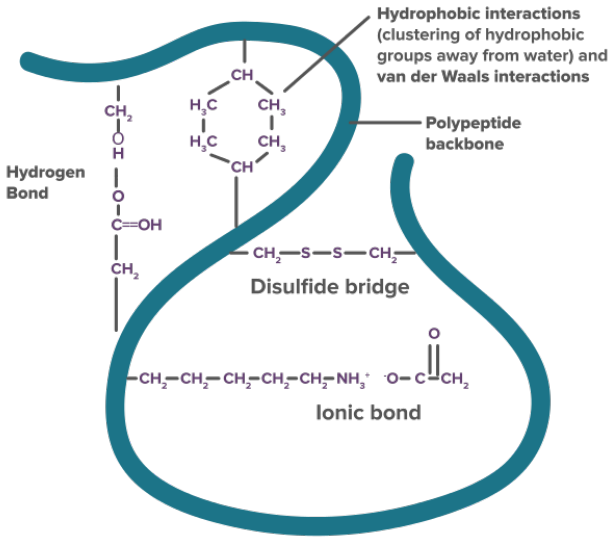




- Typically, the local model consists of a **fragment library** that models the backbone angles (ϕ, ψ).
- We formulated a probabilistic model, based on a hidden Markov model, that relates the amino acid sequence **a** to the dihedral angles sequence $\mathbf{x} = (\phi, \psi)$, using a bivariate von Mises distribution on the torus[5].

[5] Boomsma *et al.*, PNAS, 2008 & 2014.

Model of nonlocal protein structure



- As a statistical **descriptor** of nonlocal structure, $\mathbf{y} = f(\mathbf{x})$, we use a vector of five physical energies.
  - Hydrogen bond energy in helices, strands and coils
  - Hydrophobic energy and electrostatic energy (ionic bonds)
- $p(\mathbf{y} \,|\mathbf{a})$ is a multivariate Gaussian distribution (obtained using Bayesian Deep Learning[6]).

[6] Work in progress!

The resulting posterior - summary

- The resulting posterior is

$$p(\mathbf{x} \mid \mathbf{a}) = \frac{p(\mathbf{y} \mid \mathbf{a})}{\pi(\mathbf{y} \mid \mathbf{a})} \pi(\mathbf{x} \mid \mathbf{a})$$

  - The nonlocal model $p(\mathbf{y} \mid \mathbf{a})$ is a 5-dimensional Gaussian that models hydrogen bonding, electrostatic interactions and the hydrophobic effect.
  - The local model $\pi(\mathbf{x} \mid \mathbf{a})$ is a hidden Markov model that models the (ϕ, ψ) angles.
  - These models are easy to estimate and computationally efficient.
  - The posterior is valid on both the local and nonlocal scale.
- However, often the estimation of $\pi(\mathbf{y} \mid \mathbf{a})$ forms a serious bottleneck.

Variational reference ratio I

- In order to avoid the direct estimation of $\pi(\mathbf{y} \mid \mathbf{a})$, we use a Gaussian approximation

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{a}) &= \frac{p(\mathbf{y} \mid \mathbf{a})}{\pi(\mathbf{y} \mid \mathbf{a})} \pi(\mathbf{x} \mid \mathbf{a}) \\ &\approx \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}, \Sigma) \pi(\mathbf{x} \mid \mathbf{a}) \\ &= q(\mathbf{x} \mid \mathbf{a}, \boldsymbol{\mu}, \Sigma) \end{aligned}$$

- The parameters of the Gaussian are estimated by minimizing the following Kullback-Leibler divergence

$$\arg \min_{\boldsymbol{\mu}, \Sigma} D_{\mathrm{KL}} \left(q(\mathbf{y} \mid \mathbf{a}, \boldsymbol{\mu}, \Sigma) \parallel p(\mathbf{y} \mid \mathbf{a}) \right)$$

- This reminds of Variational Bayes estimation.

Variational reference ratio II

VPK algorithm

- Generate n samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from $\pi(\mathbf{x})$.
- Choose an initial value $(\boldsymbol{\mu}_0, \Sigma_0)$
- Assign weights to the samples equal to $\mathcal{N}(\mathbf{y}_i \mid \boldsymbol{\mu}_0, \Sigma_0)$
  - with $\mathbf{y}_i = f(\mathbf{x}_i)$
- Starting from $(\boldsymbol{\mu}_0, \Sigma_0)$, use the downhill simplex method to find

$$(\boldsymbol{\mu}_1, \Sigma_1) = \arg \min_{(\boldsymbol{\mu}, \Sigma)} D_{\mathrm{KL}} \left(q(\mathbf{y} \mid \mathbf{a}) \parallel p(\mathbf{y} \mid \mathbf{a}) \right)$$

- We need a method to compute the KL divergence between a Gaussian and a set of weighted samples.

Variational reference ratio III

- Now we address the problem of calculating the KL divergence between a Gaussian and a set of **weighted samples**.
  - The weights w_i are given by the approximation of the ratio, $\mathcal{N}(\mathbf{y}_i \mid \boldsymbol{\mu}, \Sigma)$.
  - The KL divergence is minimized in function of $\boldsymbol{\mu}, \Sigma$.
- We write $D_{\mathrm{KL}}(p \parallel q)$ in terms of the cross- and differential-entropy.

$$D_{\mathrm{KL}}(q \parallel p) \quad = \quad S_C(q \parallel p) - S_D(q)$$

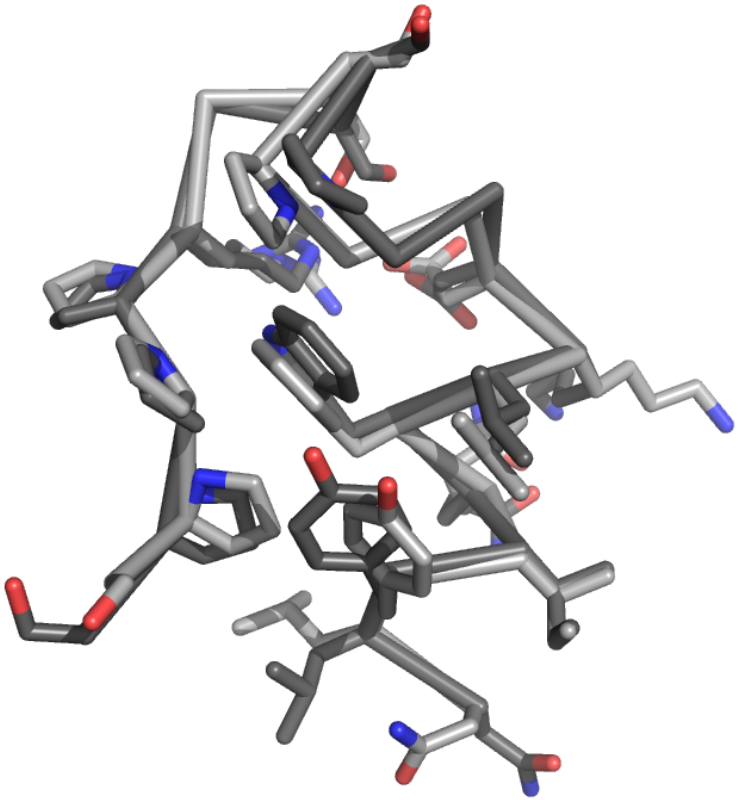
- The **cross entropy** C can be readily calculated as

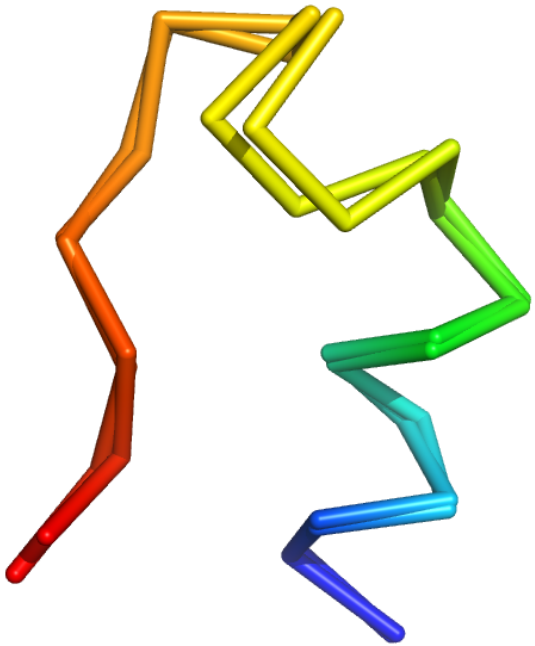
$$S_C(q \parallel p) \approx -\sum_i w_i \log p(\mathbf{y}_i \mid \mathbf{a})$$

- The **differential entropy** S_D can be approximated using a nonparametric nearest neighbor estimator[7].

[7] Ajgl and Šimandl (2011)

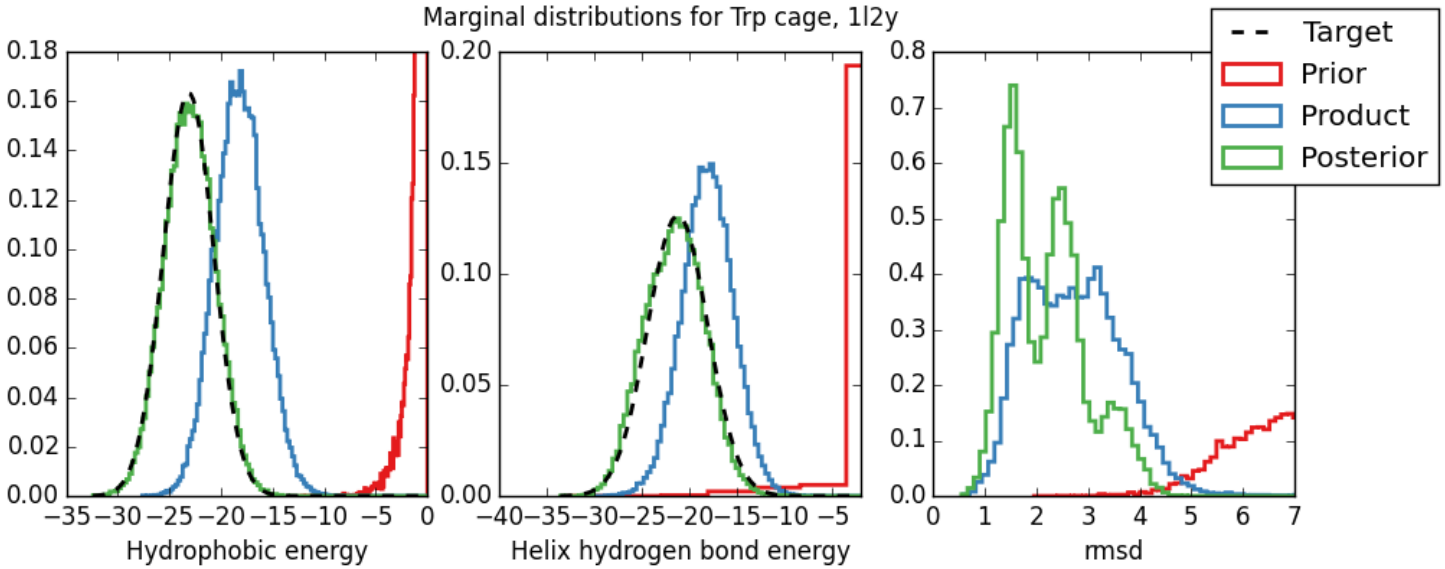
Example: Trp-Cage Miniprotein I





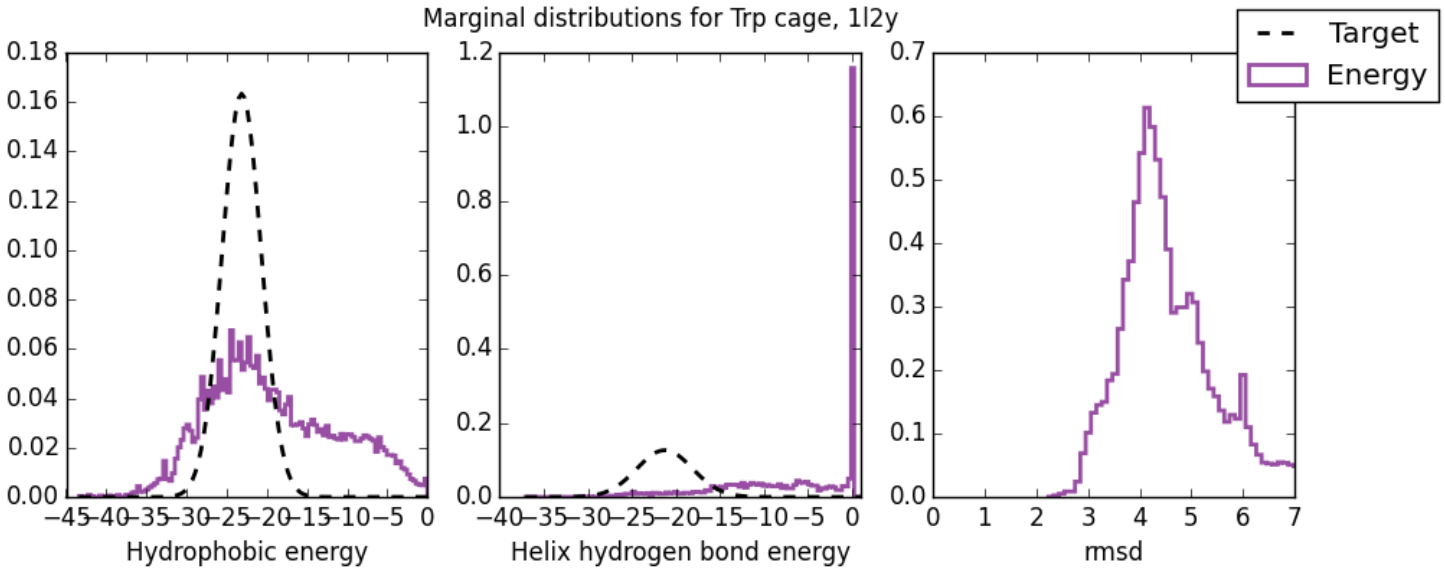
- 20 amino acids, Cα root mean square deviation=0.5 Å.
- Native structure in dark grey; prediction (using native energy) in light grey.

Example: Trp-Cage Miniprotein II



- Dashed black=target energy; Red=simulation using prior alone.
- Blue=$p(\mathbf{y} \mid \mathbf{a})\pi(\mathbf{x} \mid \mathbf{a})$, which amounts to assuming independence of $\mathbf{x}$ and $\mathbf{y}$.
- Green=VRR solution, $\mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}, \Sigma)\pi(\mathbf{x} \mid \mathbf{a})$.

Example: Trp-Cage Miniprotein III



- Dashed black=target energy; Purple=simulation using energy.
- Even though the energy is useful as a **descriptor** of nonlocal structure, the protein cannot be folded using the energy itself.

Implications and outlook

- Searching for "knowledge based potentials/proteins" in Google Scholar results in 3000+ hits.
- After more than 25 years of heated discussion about Sippl's potentials, we finally know why they work – **they approximate Jeffrey's conditioning**.
- The **reference state** is defined by the local model.
  - No more need to hack the reference state.
- These energies **generalize** beyond pairwise distances.
- Jeffrey's conditioning opens the way to **new, well-justified energy functions**.
- Jeffrey's conditioning allows us to formulate a **complete probabilistic model of proteins** in atomic detail, for the first time.

Thank you

DET FRIE FORSKNINGSRÅD
DANISH COUNCIL FOR INDEPENDENT RESEARCH

THE VELUX FOUNDATIONS
VILLUM FONDEN ✕ VELUX FONDEN

- Wouter Boomsma, KU
- Jesper Ferkinghoff-Borg, DTU
- Jesper Foldager, KU
- Jes Frellsen, KU⇒"**We rediscovered Jeffrey's conditioning**!"
- John Haslett, Trinity College, Dublin, Ireland
- John T. Kent, Kanti V. Mardia, Leeds, UK
- Douglas Theobald, Brandeis, USA
- *Dedicated to Richard Jeffrey* (1926 – 2002)

References

- Diaconis, P. and Zabell, S. (1983) *Some Alternatives to Bayes' rule*. **Technical Report-339** Stanford University, Department of Statistics.
- Hamelryck, T. *et al.* (2010) *Potentials of mean force for protein structure prediction vindicated, formalized and generalized*. **PLoS ONE** 5(11): e13714.
- Valentin, J. *et al.* (2013) *Formulation of probabilistic models of protein structure in atomic detail using the reference ratio method*. **Proteins** 82:288–299.
- Hamelryck, T. *et al.* (2015). *Proteins, physics and probability kinematics: a Bayesian formulation of the protein folding problem*. In **Geometry Driven Statistics**, Wiley.
- Boomsma, W., Mardia, KV., Taylor, CC., 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
- Boomsma, W., Tian, P., Ferkinghoff-Borg, J., Hamelryck, T., Lindorff-Larsen, K. , Vendruscolo, M. (2014) Equilibrium simulations of proteins using molecular fragment replacement and NMR chemical shifts. **Proc. Natl. Acad. Sci. USA** 111(38):13852-13857
- Ajgl, J., Šimandl, M. (2011) Differential entropy estimation by particles. **IFAC Proceedings** 44(1):11991-11996