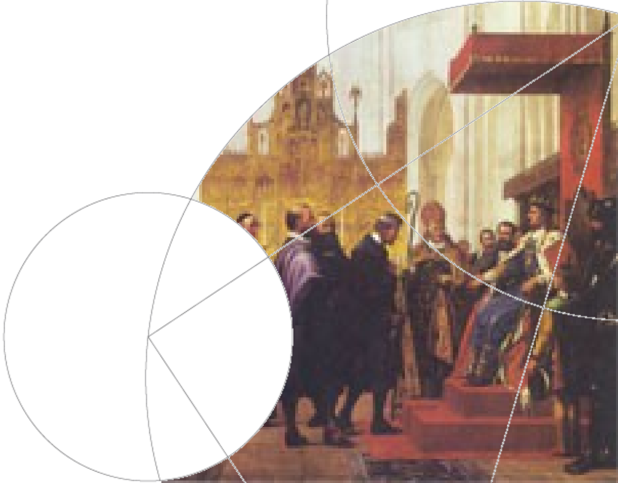
Probabilistic models of protein structure

Thomas Hamelryck
Structural Bioinformatics group
Bioinformatics Centre

Associate professor
University of Copenhagen, Denmark

Visiting professor
University of Leeds, UK



The protein folding problem

Central problem in science

- Biology, physics....and statistics
- Biotechnology
  - Enzyme design, new chemistry
  - New materials (f.ex. spider silk)
- Medicine
  - Drugs, vaccines

Proteins are linear polymers of amino acids

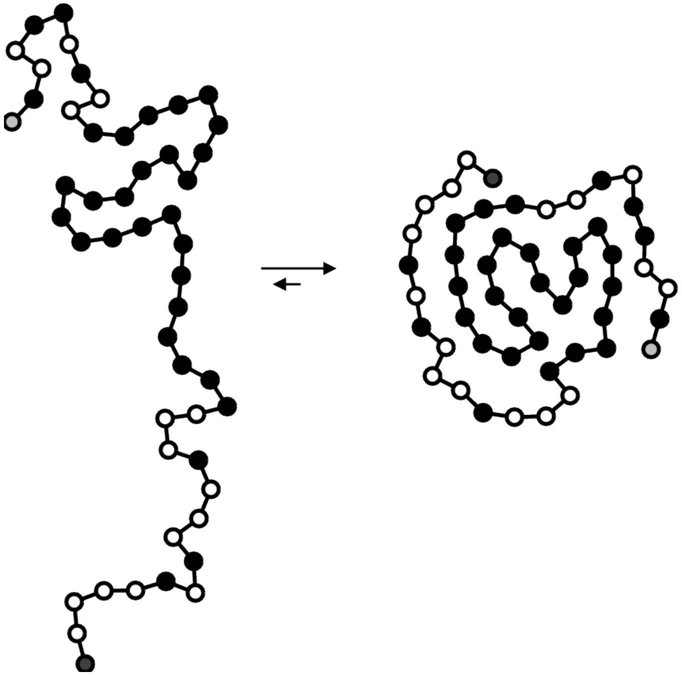
- 20 different amino acids
- Hydrophobic amino acids on the inside
- Hydrophylic amino acids on the outside

Sequence encodes a compact 3D shape

- Protein fold

Predicting structure from sequence

- One of the main open problems in biology
- Our goal is to formulate a probabilistic model of protein structure, and apply it to inference, prediction and design



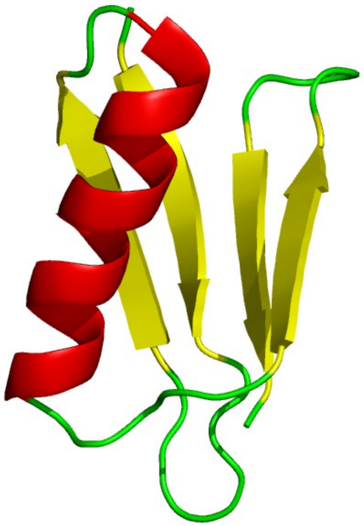
How can we formulate a probabilistic model of protein structure?

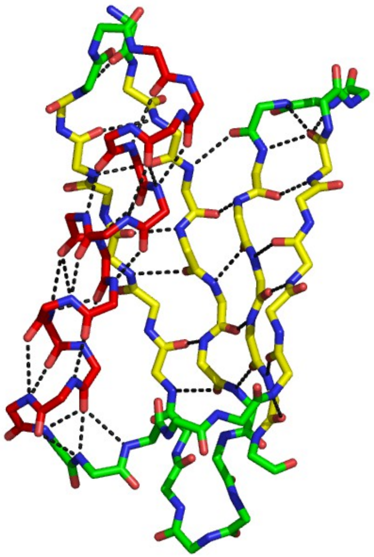
Local structure

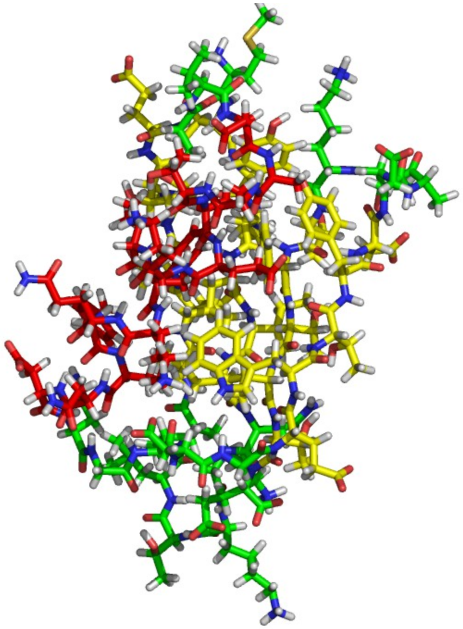
- Shape of the protein on a local length scale
  - Helices, strands, coils...
- Can we develop an efficient local model that allows sampling?

Nonlocal structure

- Interactions between residues far apart in sequence
- Which model and how to combine it with the local model?







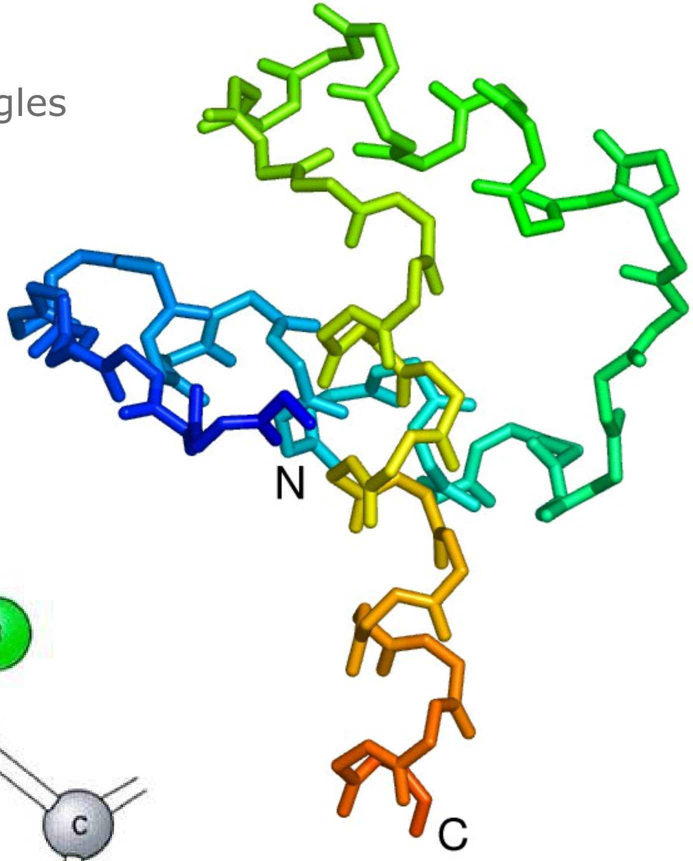
Parameterization of a protein's structure

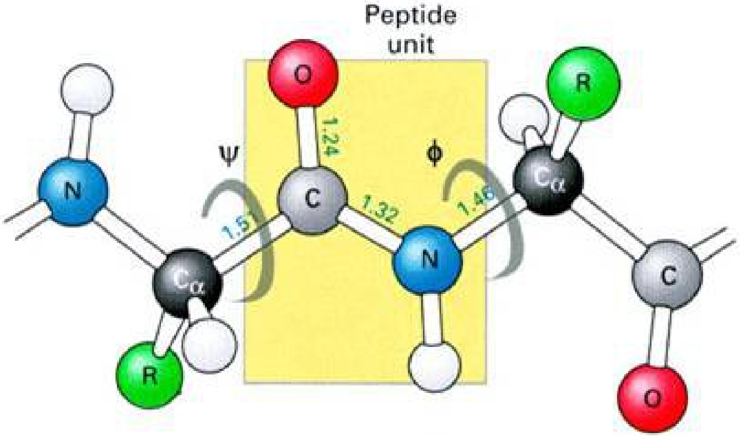
One amino acid=3 points

- N, Cα and C atoms
- We assume ideal bond distances and angles
- We leave out the side chains for now

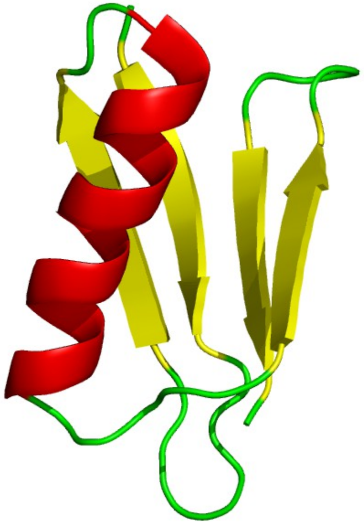
Parameterization?

- Sequence of n-2...
  - Dihedral angle pairs (ϕ,ψ)
  - Angles in [-π, π)
  - Points on the torus T^2





A probabilistic model of local structure

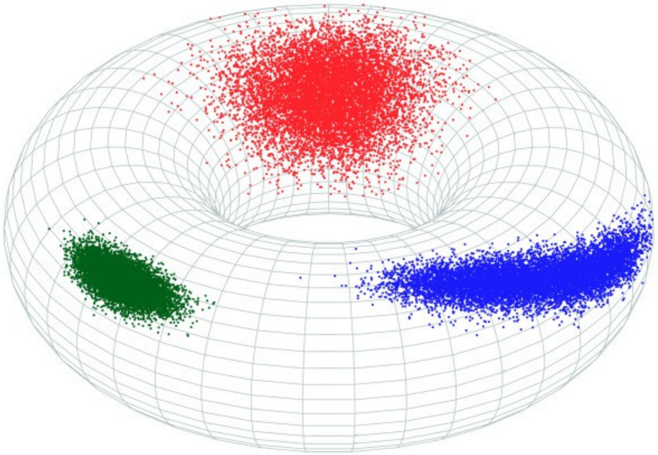


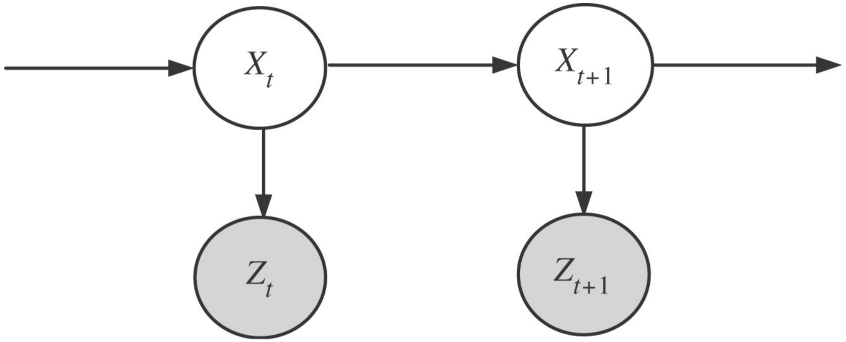
Goal: a probabilistic model for backbone angles

- Generative (allows sampling), continuous
- Sequence, angles, secondary structure

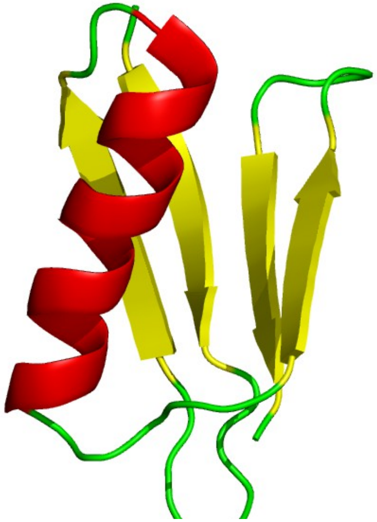
Two problems:

- Angles: Directional statistics
  - Bivariate von Mises distribution on the torus
    - (Mardia, Taylor & Subramaniam, 2007)
- Sequential nature: dynamic Bayesian network (DBN)
  - Hierarchical model, essentially a hidden Markov model



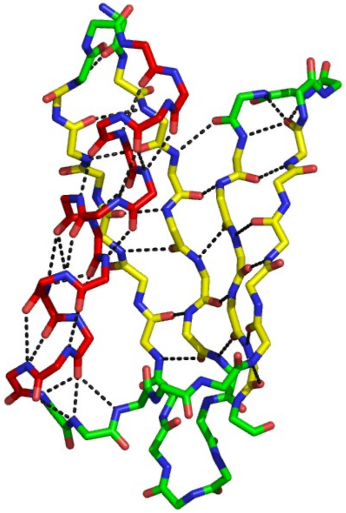


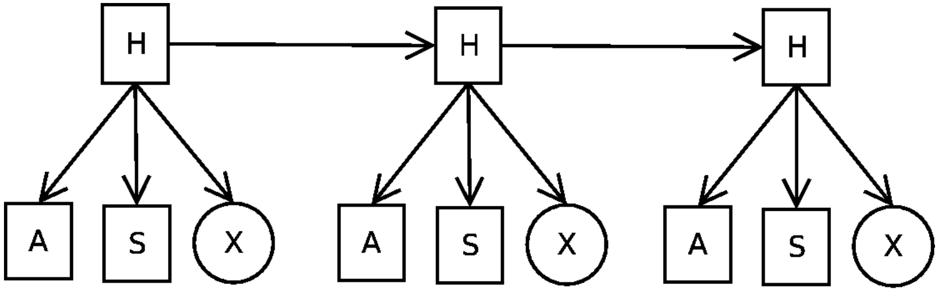
TORUSDBN: a model of local structure



Dynamic Bayesian network for local protein structure

- (Boomsma et al., PNAS, 2008)
- Probabilistic model with a graph representation
  - Nodes are variables, edges encode independencies
- Amino acid symbols (A)
- Secondary structure labels (S)
  - Angle pairs ϕ,ψ of the amino acids (X)
  - Points on the 2D torus
- Markov chain of hidden nodes (H)
  - Nuisance variable, statistical magic
- Trained using 1500 proteins

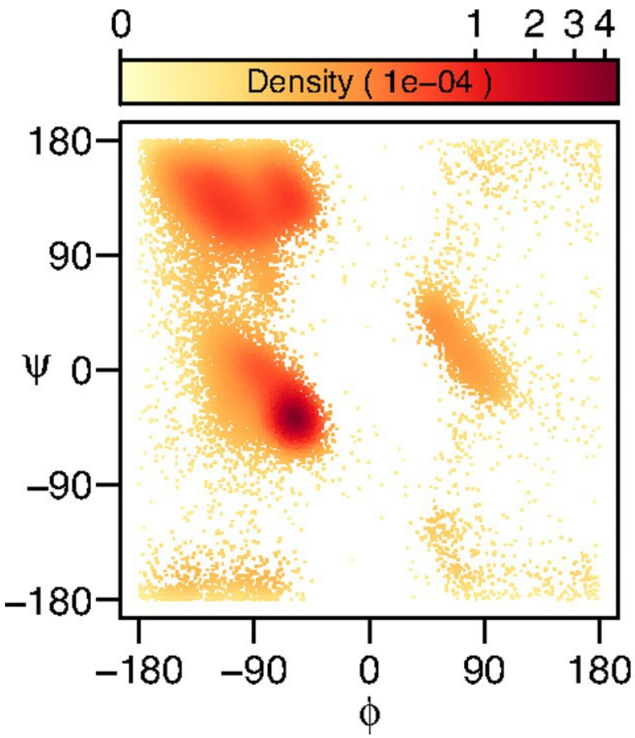




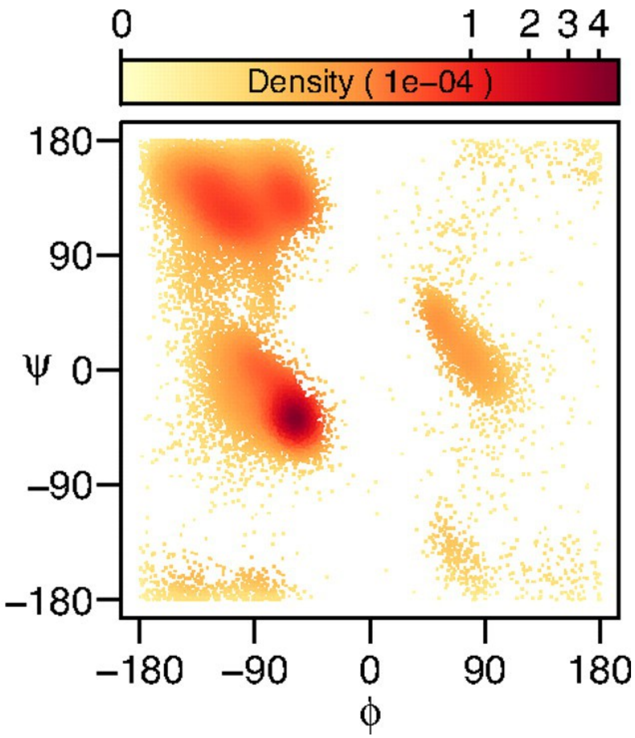
$$P(A,S,X)=\sum_H P(A|H)P(S|H)P(X|H)P(H)$$

Example I: Ramachandran plot

50K samples, protein test set versus TORUSDBN
Lengths of secondary structures are also reproduced



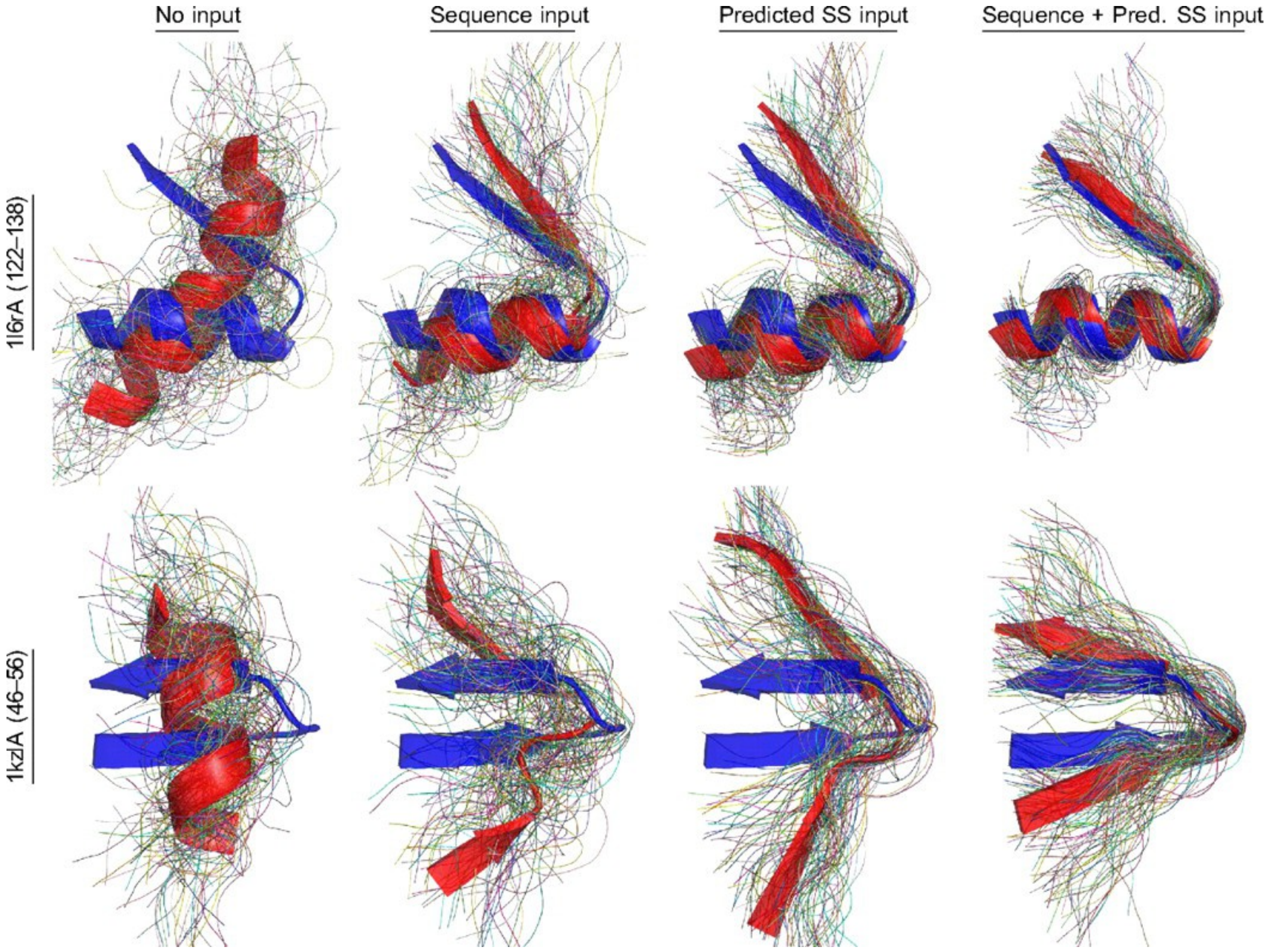
Test set



TORUSDBN

Example II: Sampling motifs

Sampling motifs (α-turn-β and β-turn-β)



BASILISK: A model of side chain structure

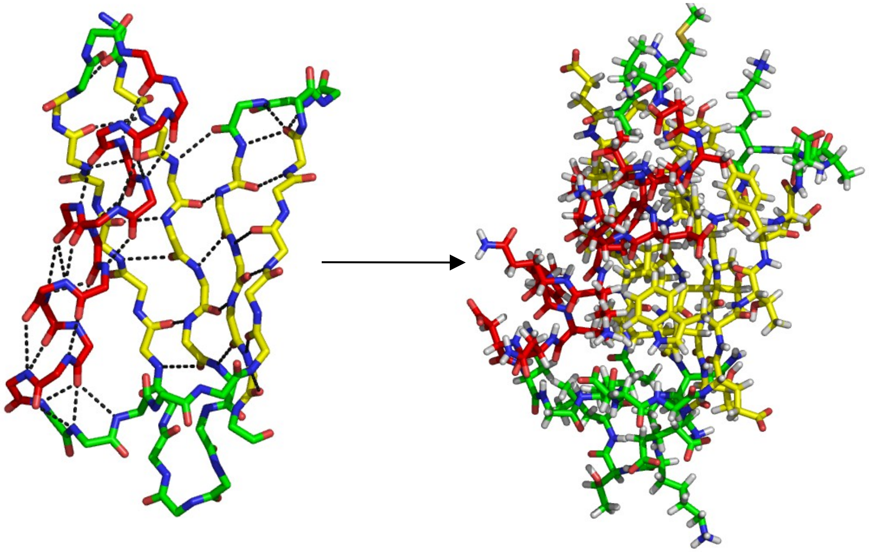
TORUSDBN does not include the side chains

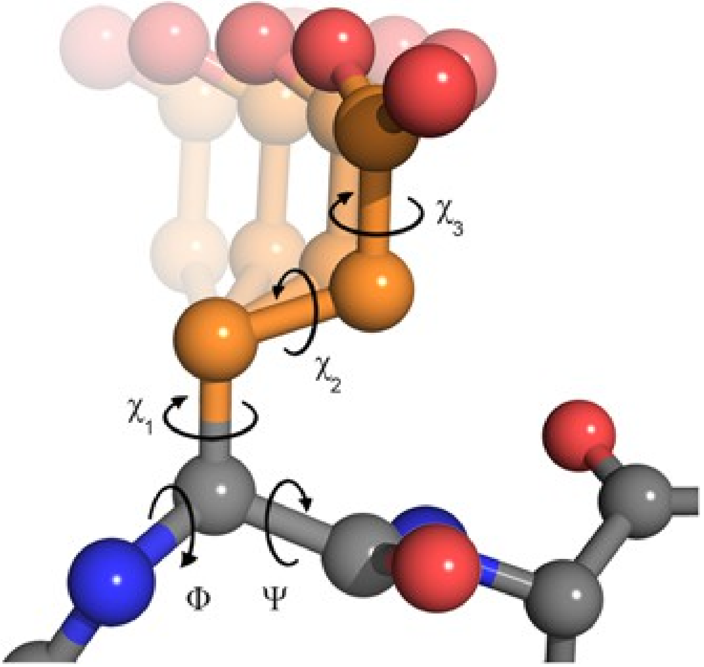
Side chains are also parametrized using dihedral angles (chi, χ)

- Again, assuming ideal bond angles and lengths
- From zero to four angles

BASILISK complements TORUSDBN

- (Harder et al., BMC Bioinformatics, 2010)



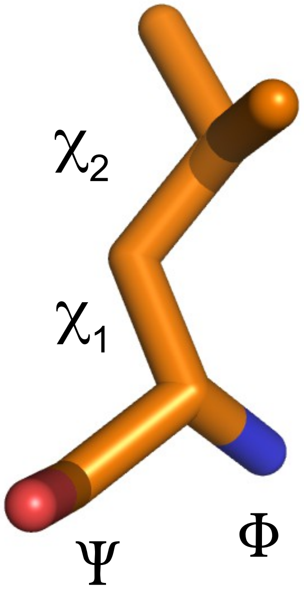


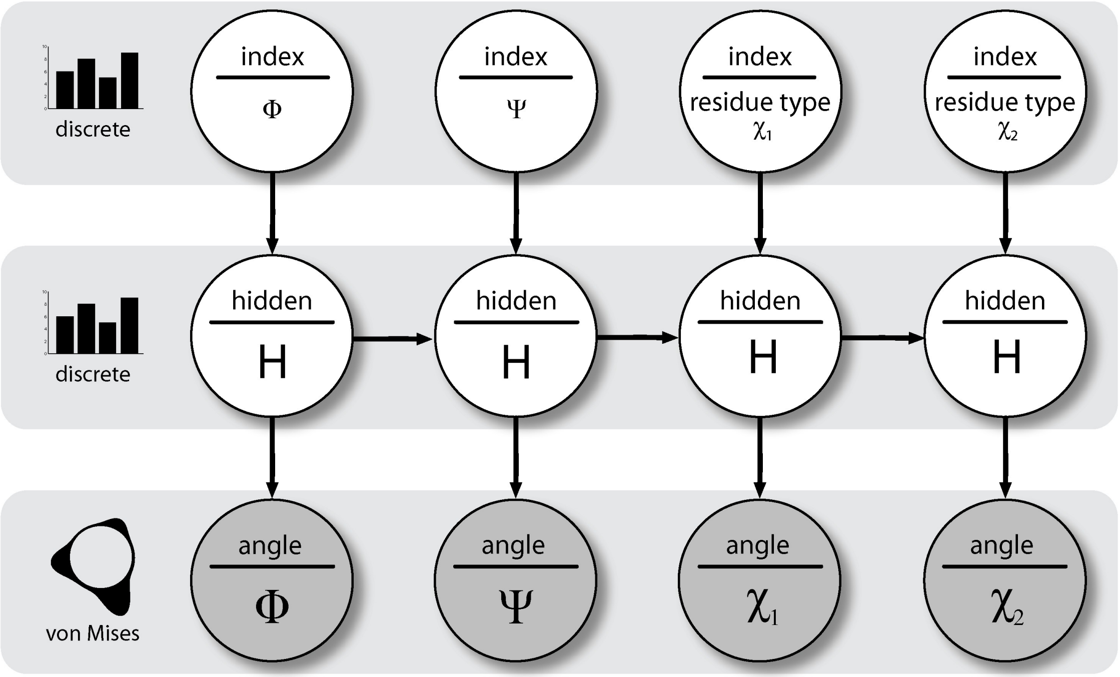
Glutamate

BASILISK: probabilistic model for side chains

A dynamic Bayesian network that represents side chains

- All relevant amino acids in one model (*transfer learning*)
- Generative, continuous
- Includes the backbone angles
- (Harder et al., BMC Bioinformatics, 2010)





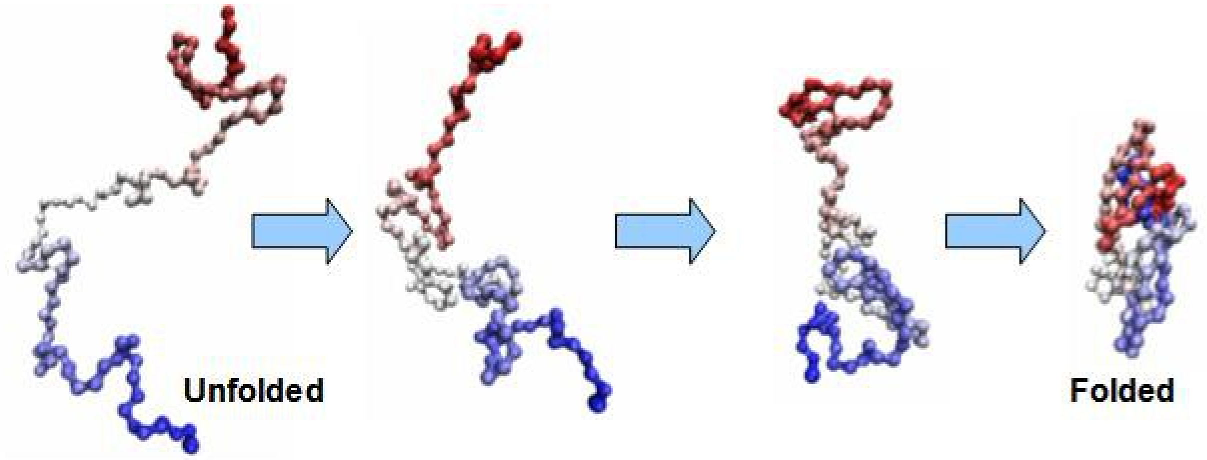
A probabilistic model in atomic detail, but...

TORUSDBN and BASILISK constitute the first probabilistic model of protein structure with atomic detail

- Hurray, the problem is solved, and we can go to the beach?

Problem: this model works on a local length scale

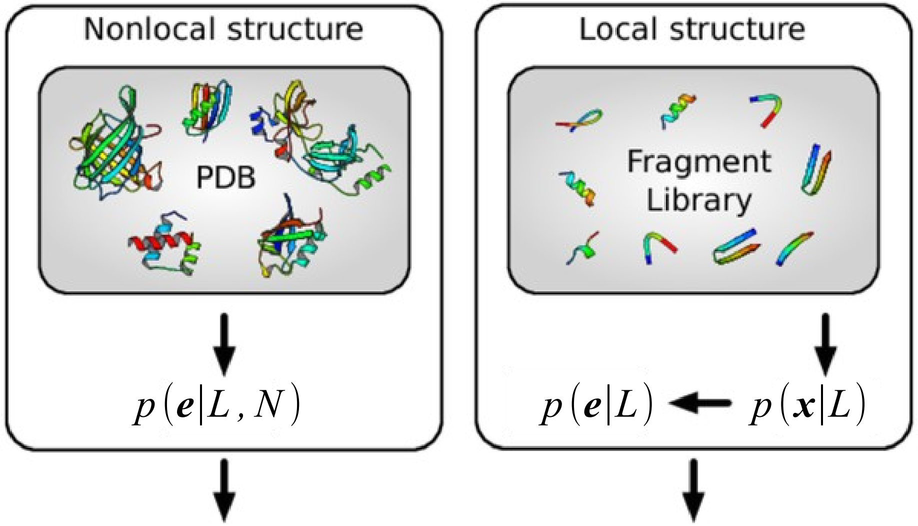
- We used a Markov chain of hidden nodes
- Nonlocal features are missing: hydrophobic effect, long range hydrogen bonding, electrostatic interactions...



Towards a complete model of protein structure

Augment TORUSDBN, p(**x**|L), with nonlocal information

- Add a probability distribution on some nonlocal feature vector **e**
  - **e**=f(**x**)
  - For example, radius of gyration
- Can be done with the reference ratio method (PLoS ONE, 2010)
- Used for 20 years in protein structure prediction as potentials of mean force, without having a clue why it works



$$p(\boldsymbol{x}|L,N)=\frac{p(\boldsymbol{e}|L,N)}{p(\boldsymbol{e}|L)}\times p(\boldsymbol{x}|L)$$

Proof of reference ratio method

Required: $p(\mathbf{x} \mid L, N)$, with L=local, N=nonlocal structure
Given: $p(\mathbf{x} \mid L)$, $p(\mathbf{e} \mid L)$, $p(\mathbf{e} \mid L, N)$ with $\mathbf{e} = F(\mathbf{x})$
Solution:
First, we note that

$$\begin{aligned} p(\mathbf{x} \mid L) = p(\mathbf{x}, \mathbf{e} \mid L) &= p(\mathbf{x} \mid \mathbf{e}, L) p(\mathbf{e} \mid L) \\ \Rightarrow & \quad p(\mathbf{x} \mid \mathbf{e}, L) = \frac{p(\mathbf{x} \mid L)}{p(\mathbf{e} \mid L)} \qquad (1) \end{aligned}$$

In addition

$$\begin{aligned} p(\mathbf{x} \mid L, N) = p(\mathbf{x}, \mathbf{e} \mid L, N) &= p(\mathbf{x} \mid \mathbf{e}, L, N) p(\mathbf{e} \mid L, N) \\ &= p(\mathbf{x} \mid \mathbf{e}, L) p(\mathbf{e} \mid L, N) \qquad (2) \end{aligned}$$

Putting (1) in (2) results in

$$p(\mathbf{x} \mid L, N) = \frac{p(\mathbf{e} \mid L, N)}{p(\mathbf{e} \mid L)} p(\mathbf{x} \mid L)$$

Probability kinematics, Jeffrey's conditioning

Introduced by Richard C. Jeffrey in the 50ies

- Philosopher of probability, Princeton
- (“The logic of decision”, 1965)
- (Diaconis & Zabell, JASA, 1982)

Of general interest for multi-scale problems

- Reference ratio method
  - Estimate local model
  - Estimate nonlocal model from local model
  - Estimate nonlocal model from data
  - Glue everything together with PK
  - Explains “potentials of mean force”
    - (Hamelryck et al., PLoS ONE, 2010)
- Speech signals, images, movements,...
- Azzalini's skew distributions

Richard C. Jeffrey
(1926-2002)

$$p\left(\boldsymbol{x}\middle|L,N\right)=\frac{p\left(\boldsymbol{e}\middle|L,N\right)}{p\left(\boldsymbol{e}\middle|L\right)}\times p\left(\boldsymbol{x}\middle|L\right)$$

An energy vector provides nonlocal information

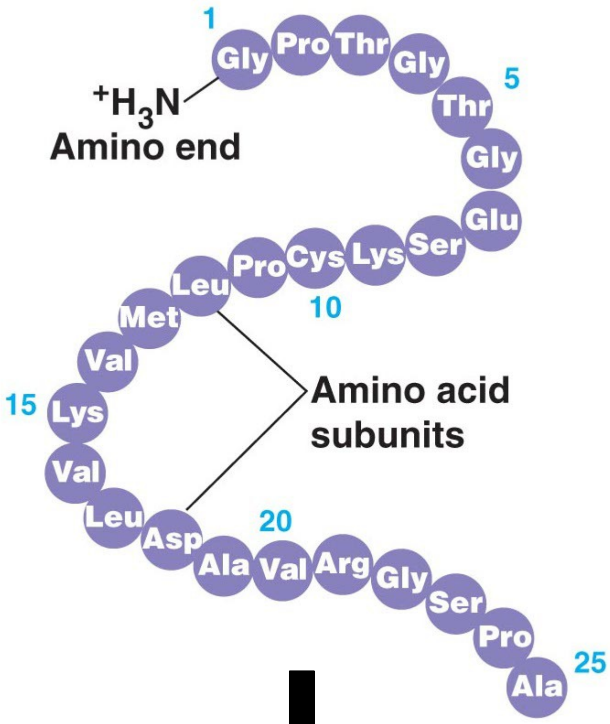
Radius is not enough; we need more detail

An energy vector describes global structure

- p(**e**|**a**) is a simple multivariate Gaussian
- Inferred for a given sequence **a**
- PROFASI force field

Five energies

- Electrostatic interactions e_1
- Hydrophobic interactions e_2
- Hydrogen bonds $e_{3\text{-}5}$
  - Helices, sheets, other cases
  - Information on secondary structure



$$p\left(\boldsymbol{x}|\boldsymbol{a},L,N\right)=\frac{p\left(\boldsymbol{e}|\boldsymbol{a},L,N\right)}{p\left(\boldsymbol{e}|\boldsymbol{a},L\right)}\times p\left(\boldsymbol{x}|\boldsymbol{a},L\right)$$

$$\text{with}\,\boldsymbol{e}=f\left(\boldsymbol{x}\right)$$

$$\downarrow$$

$$\boldsymbol{e}=\left\{e_1,\ldots,e_5\right\}$$

Proof-of-concept: Results for Top7

Proof-of-concept (Valentin et al., Proteins, 2013)

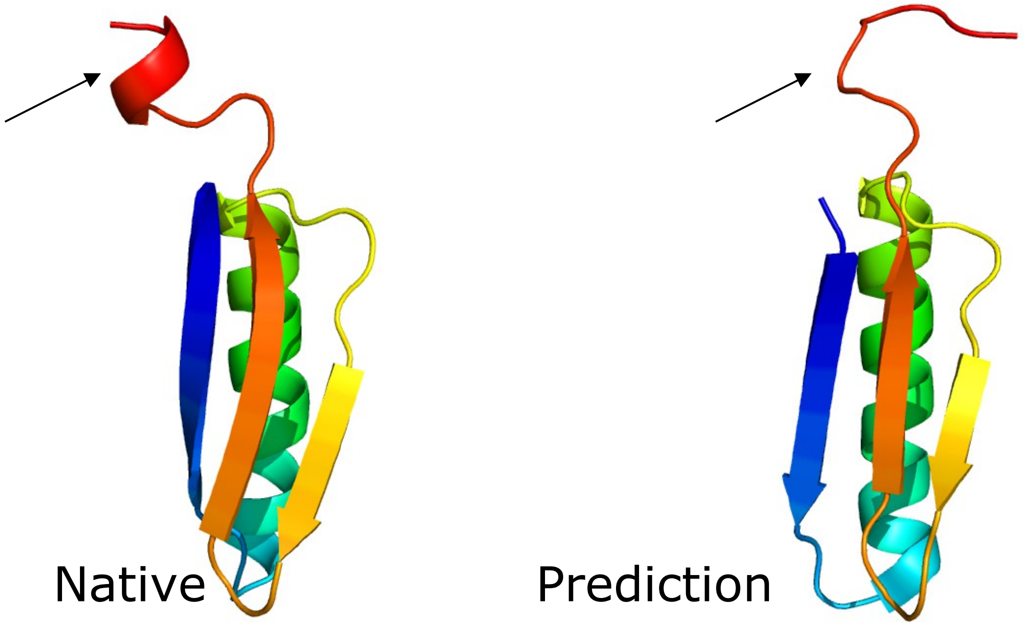
- Energy vector from native structure (noisy)

Tested and works for four proteins, up to 60 residues

- Prediction=centroid of largest cluster
- Note: PROFASI does not fold these proteins correctly
- Can handle disordered regions

PHAISTOS

www.phaistos.org



Conclusions & acknowledgments

Probabilistic model of protein structure

- Local model: graphical models, directional statistics
- Nonlocal information using probability kinematics
- Powerful, general approach to multi-scale problems

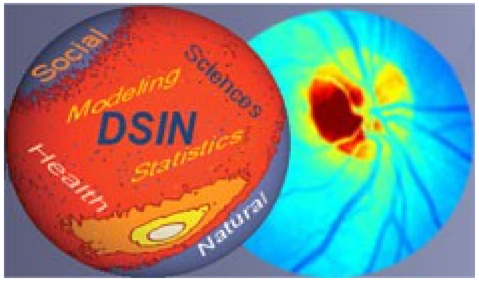
Acknowledgments

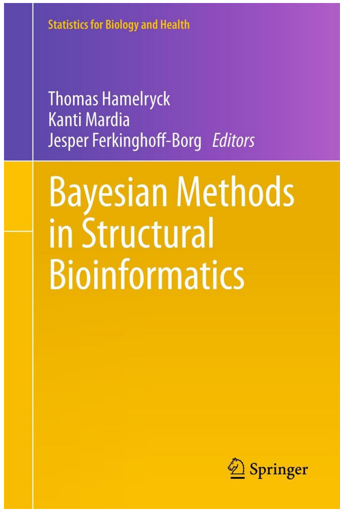
- Wouter Boomsma (TORUSDBN, PHAISTOS, CRISP)
- Jan Valentin (reference ratio *de novo* prediction)
- Sandro Bottaro (CRISP local move)
- Jes Frellsen (MUNINN, reference ratio)
- Simon Olsson (TYPHON, ensembles)
- Tim Harder (BASILISK, TYPHON)
- Pengfei Tian (PROFASI implementation)

Collaborators

- Kanti Mardia, John T. Kent, Leeds, UK
- Jesper Ferkinghoff-Borg, DTU, Denmark

Ministry of Science, Innovation and Higher Education




Statistics for Biology and Health

Thomas Hamelryck
Kanti Mardia
Jesper Ferkinghoff-Borg *Editors*

Bayesian Methods in Structural Bioinformatics

Springer

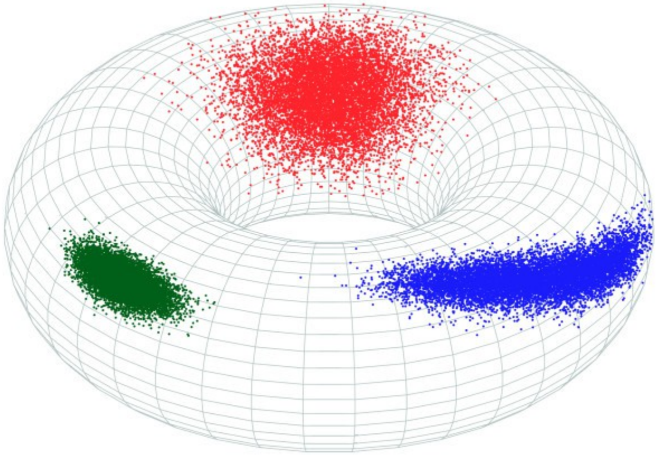
http://www.binf.ku.dk

PhD position (deadline 15/06)

Bivariate von Mises distribution

Mardia, Taylor & Subramaniam, (2007) *Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data.* **Biometrics** 63:505–512

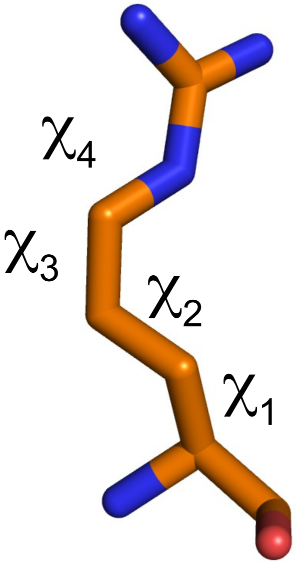
$$f(\phi, \psi) = c(\kappa_1,\kappa_2,\kappa_3)\exp(\kappa_1\cos(\phi - \mu) + \kappa_2\cos(\psi - \nu) - \kappa_3\cos(\phi - \mu - \psi + \nu))$$

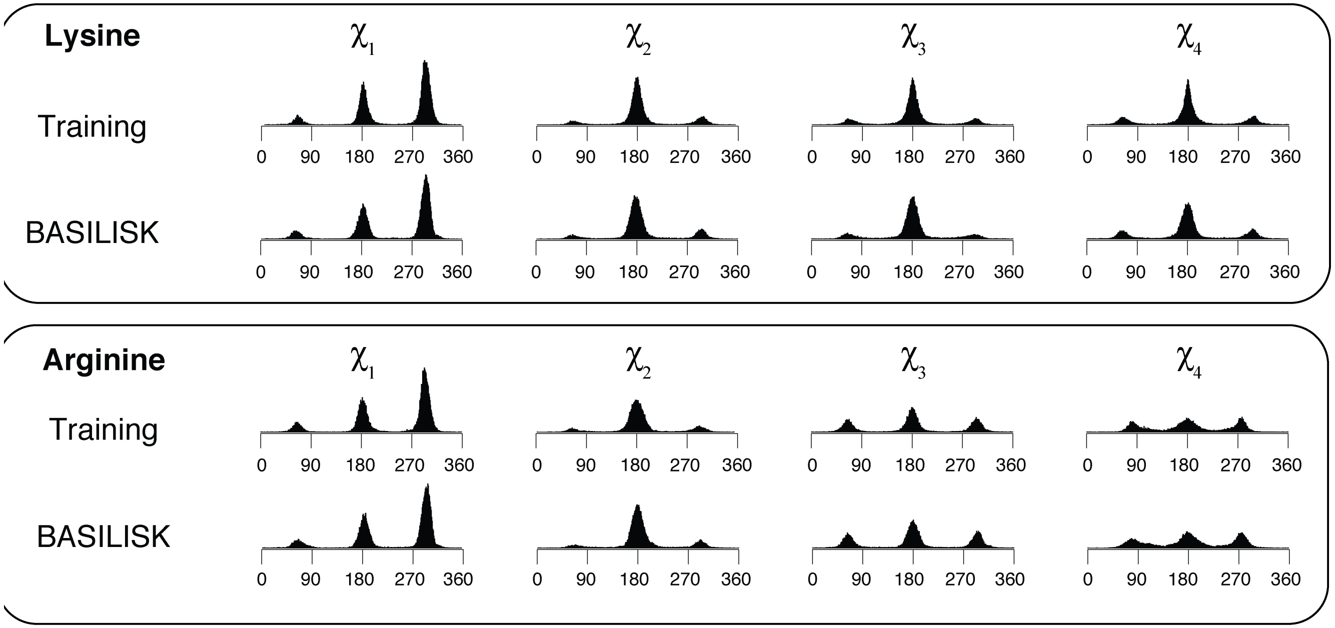


Example: lysine and arginine

Lysine and arginine have large side chains

- 4 χ angles, plus two backbone angles
- Challenging to capture in a probabilistic model

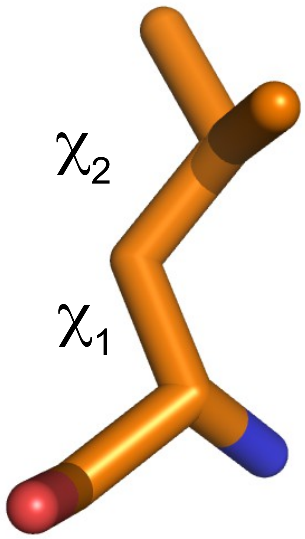


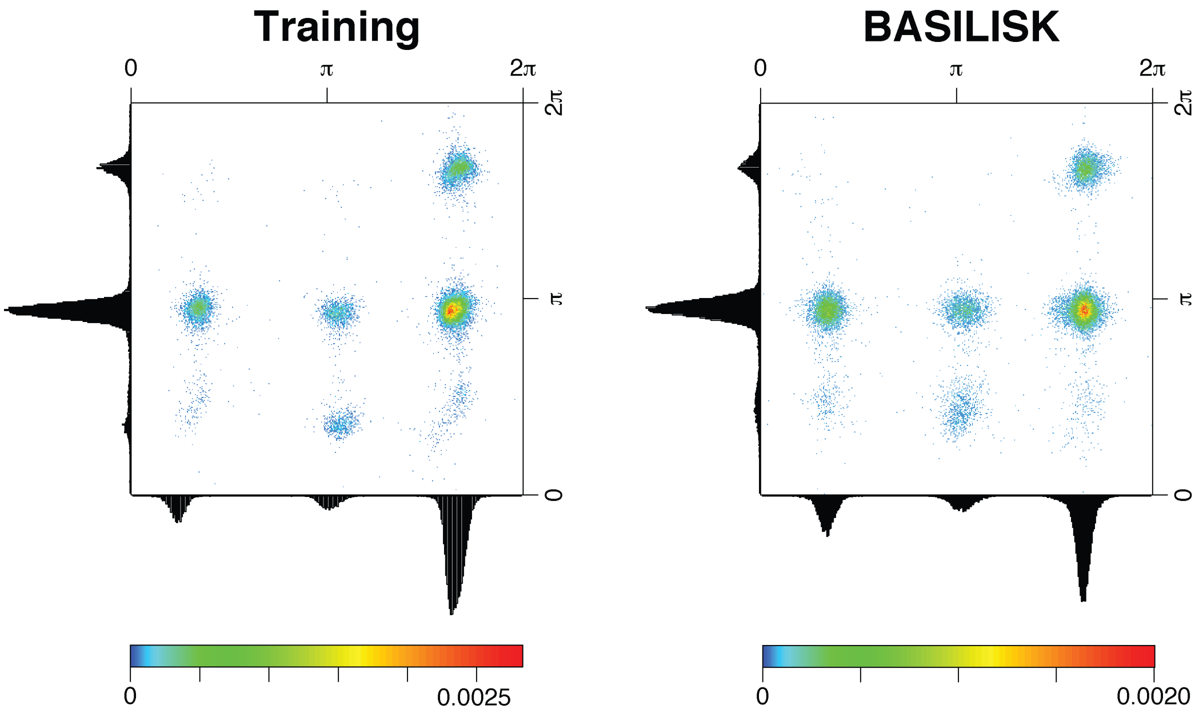


Example: leucine

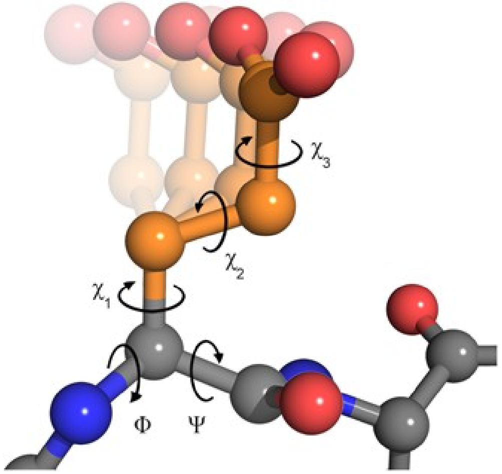
The joint probability distributions are also well captured

- Leucine: two χ angles



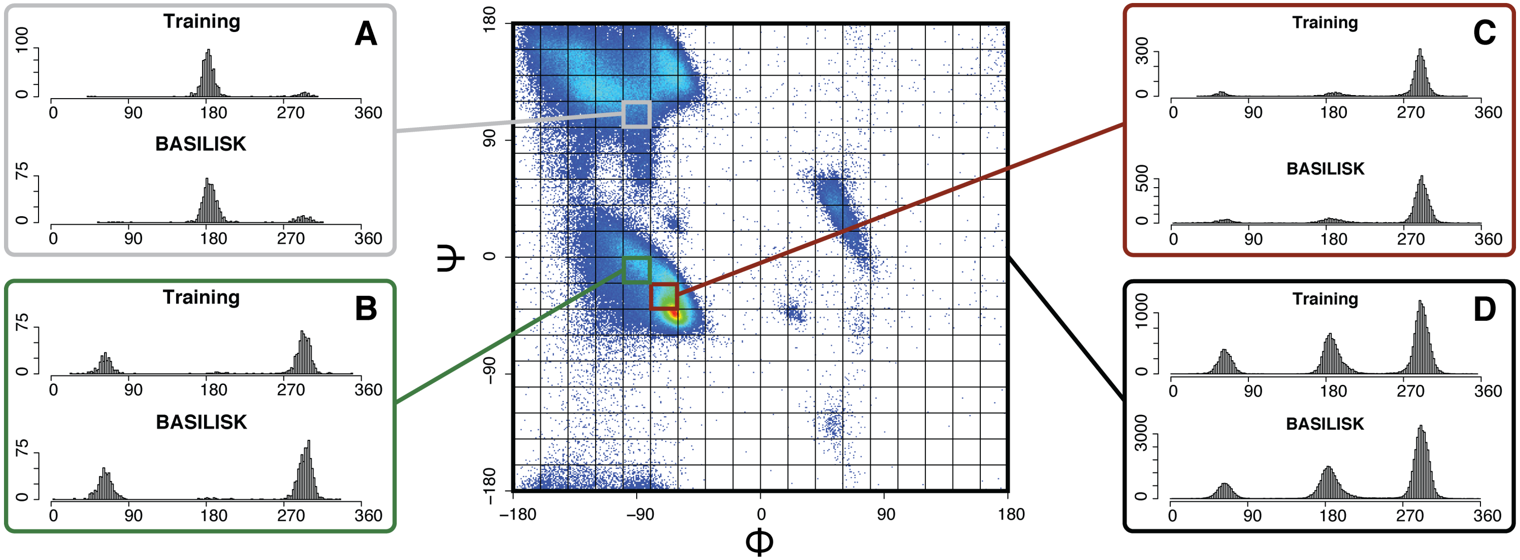


Backbone dependence



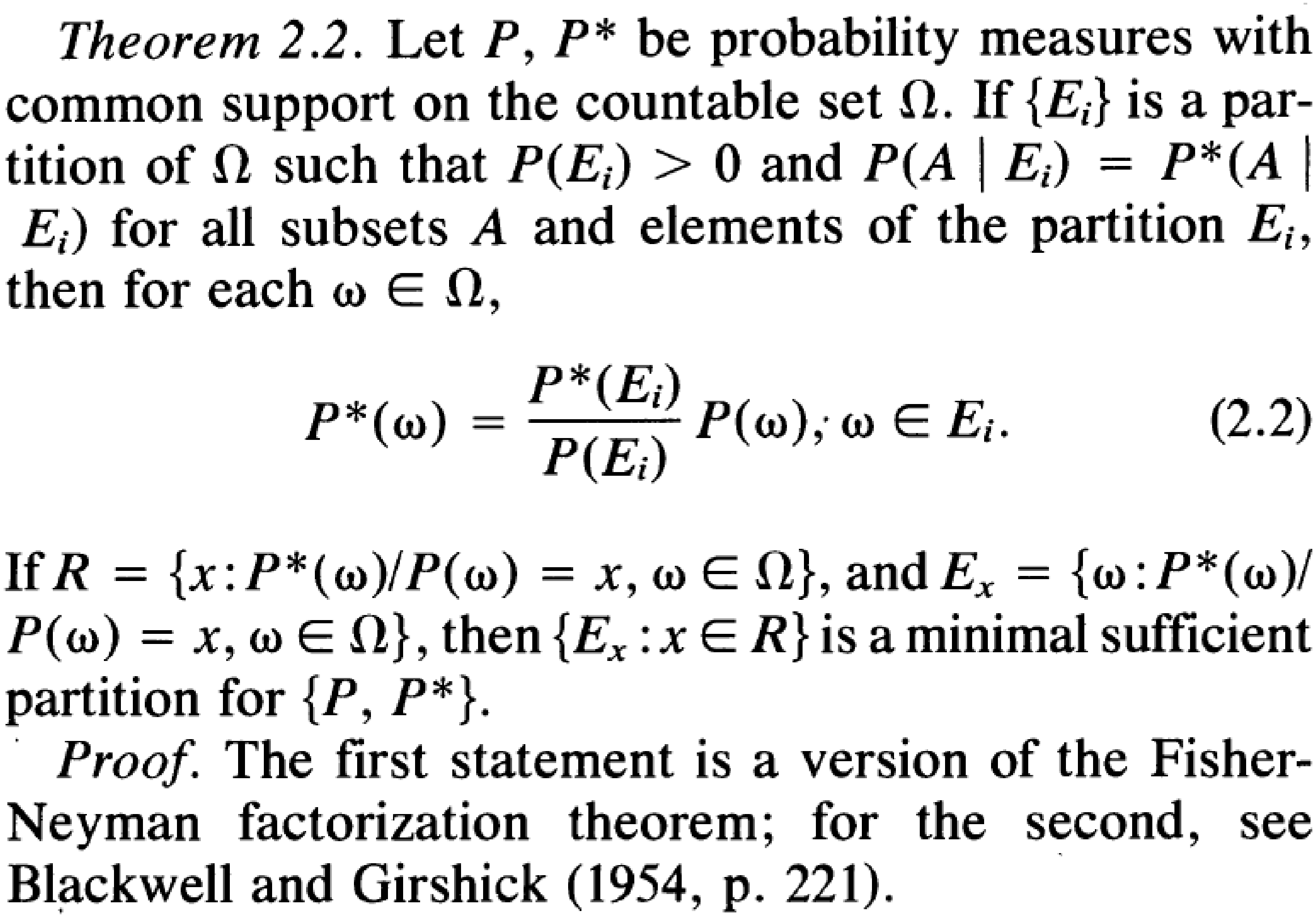
The side chain depends strongly on the backbone

- This is well captured by Basilisk
- Discretization implies an explosion of parameters



Glutamate

Probability kinematics II



Theorem 2.2. Let P, P^* be probability measures with common support on the countable set Ω. If $\{E_i\}$ is a partition of Ω such that $P(E_i) > 0$ and $P(A \mid E_i) = P^*(A \mid E_i)$ for all subsets A and elements of the partition E_i, then for each $\omega \in \Omega$,

$$P^*(\omega) = \frac{P^*(E_i)}{P(E_i)} P(\omega), \; \omega \in E_i. \tag{2.2}$$

If $R = \{x : P^*(\omega)/P(\omega) = x, \omega \in \Omega\}$, and $E_x = \{\omega : P^*(\omega)/P(\omega) = x, \omega \in \Omega\}$, then $\{E_x : x \in R\}$ is a minimal sufficient partition for $\{P, P^*\}$.

Proof. The first statement is a version of the Fisher-Neyman factorization theorem; for the second, see Blackwell and Girshick (1954, p. 221).

(Diaconis & Zabell, JASA, 1982)

Some features of the method

- Statistically well defined
- Pretty fast: under 5 days on 1 quad core CPU
- Link to physics
  - Better force fields means better performance
- Convergence can be easily evaluated
- Secondary structure can be explored freely
- Statistical uncertainty can be assessed
- Can handle disordered regions
- Unified approach to *de novo* and homology modelling
  - Protein design
- Open source implemented in PHAISTOS
  - (Boomsma et al., J. Chem. Theory Comput., 2013)
  - C++, available from sourceforge

http://www.phaistos.org

Probability kinematics, Jeffrey's conditioning

Introduced by Richard C. Jeffrey in the 50ies

- Philosopher of probability, Princeton
- ("The logic of decision", 1965)
- (Diaconis & Zabell, JASA, 1982)

We have Q(X)=Q(X,r)=Q(X|r)Q(r)

- Note that r is a deterministic function of X
- The model Q(X) is incorrect on a global scale
- That is, Q(X|r) is correct, but Q(r) is wrong

We want P(X)=P(X,r)=Q(X|r)P(r)

- P(r) is given and correct
- Problem: we have Q(r), Q(X), P(r) but not Q(X|r)

Solution is given by probability kinematics

- Follows from Q(X|r)=Q(X)/Q(r)
- Explains "potentials of mean force"
  - (Hamelryck et al., PLoS ONE, 2010)

Of general interest for multi-scale problems

- Speech signals, images, movements,...
- Azzalini's skew distributions

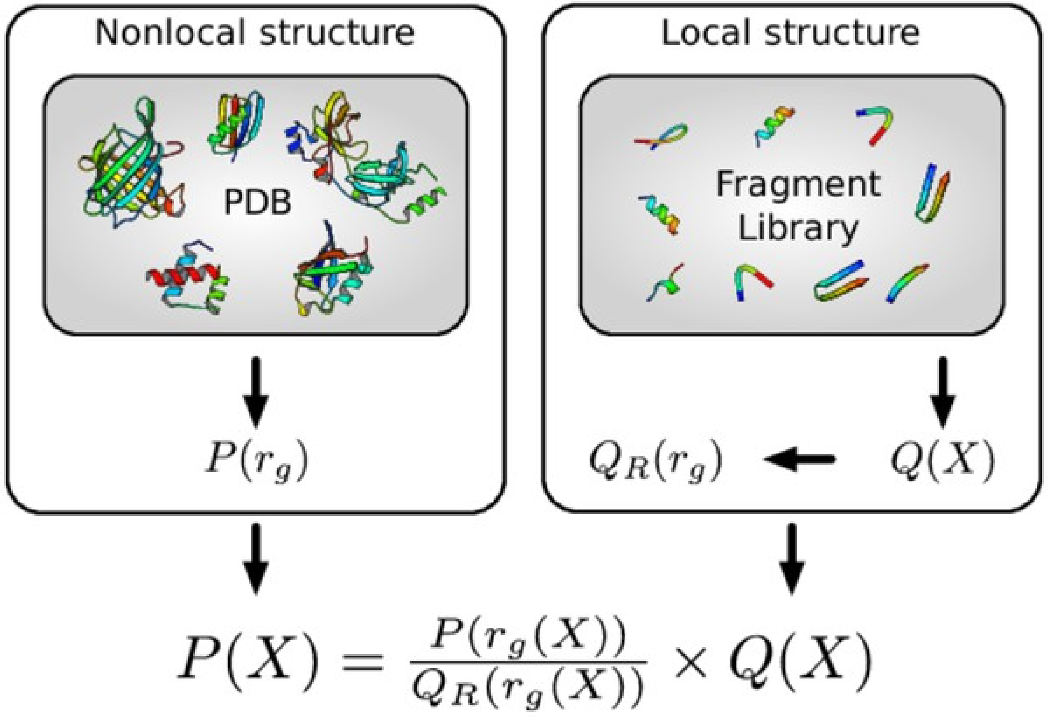
Richard C. Jeffrey
(1926-2002)

$$P(X)=\frac{P(r)}{Q(r)}\times Q(X)$$

Towards a complete model of protein structure

Augment TORUSDBN+BASILISK with nonlocal information

- Add a probability distribution on nonlocal features
    - For example, radius of gyration
- Can be done with the reference ratio method (PLoS ONE, 2010)
- Used for 20 years in protein structure prediction as potentials of mean force, without having a clue why it works



$$P(X) = \frac{P(r_g(X))}{Q_R(r_g(X))} \times Q(X)$$

Proof of reference ratio method bis

Thanks to Douglas Theobald, last Tuesday!

Since **e**=f(**x**), we know for the nonlocal model

$$p(\boldsymbol{x}|L,N)=p(\boldsymbol{e}|L,N)\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}\boldsymbol{x}} \qquad (1)$$

Similarly, for the local model

$$p(\boldsymbol{x}|L)=p(\boldsymbol{e}|L)\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}\boldsymbol{x}}$$

Thus, the Jakobian is

$$\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}\boldsymbol{x}}=\frac{p(\boldsymbol{x}|L)}{p(\boldsymbol{e}|L)} \qquad (2)$$

Putting the Jakobian (2) in (1), results in

$$p(\boldsymbol{x}|L,N)=\frac{p(\boldsymbol{e}|L,N)}{p(\boldsymbol{e}|L)}\times p(\boldsymbol{x}|L)$$