A Goodness-of-Fit Test for Protein Conformational Sampling Models

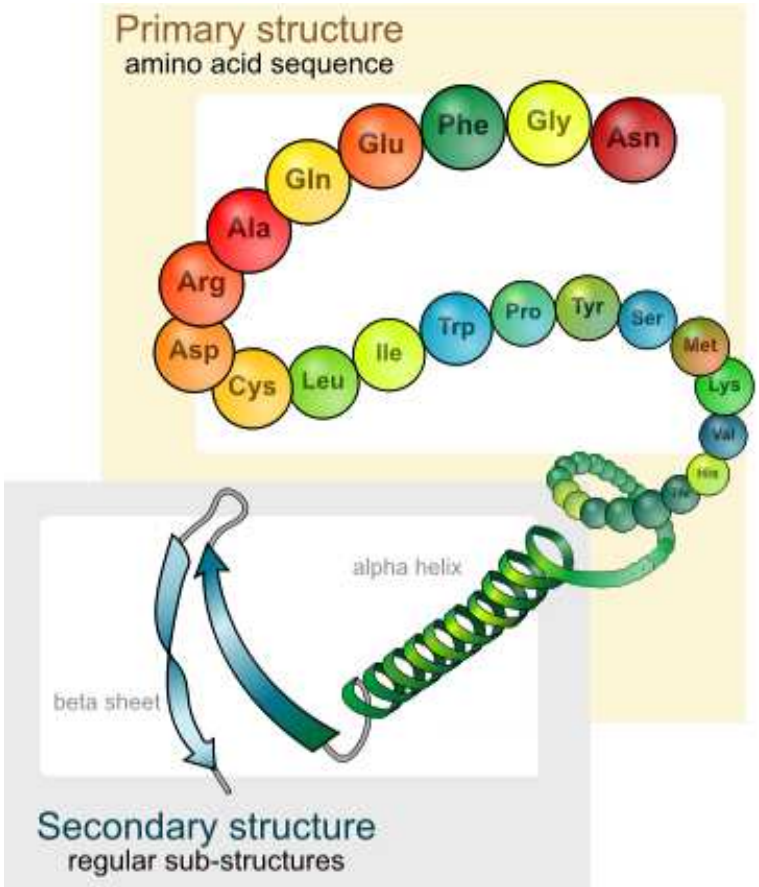
Mehdi Maadooliat
Department of Mathematics, Statistics and Computer Science
Marquette University

Xin Gao
King Abdullah University of Science and Technology

Jianhua Z. Huang
Texas A&M University

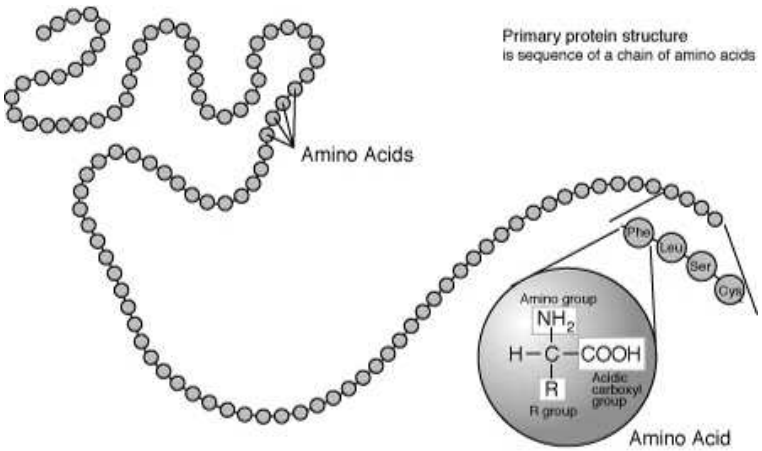
October 28, 2013

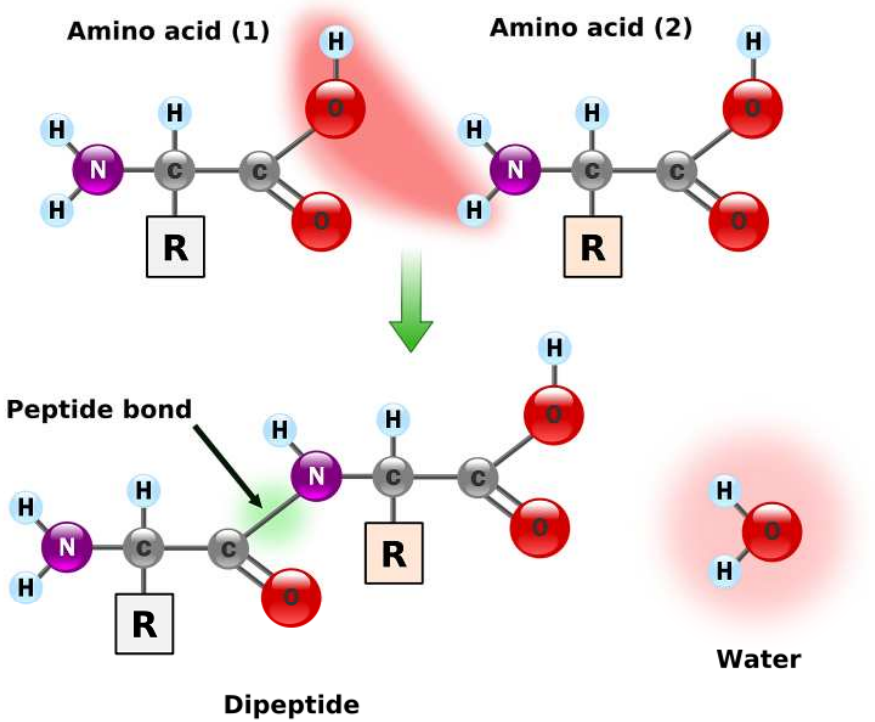
What is Protein? (Protein Structure)



source: Wikipedia

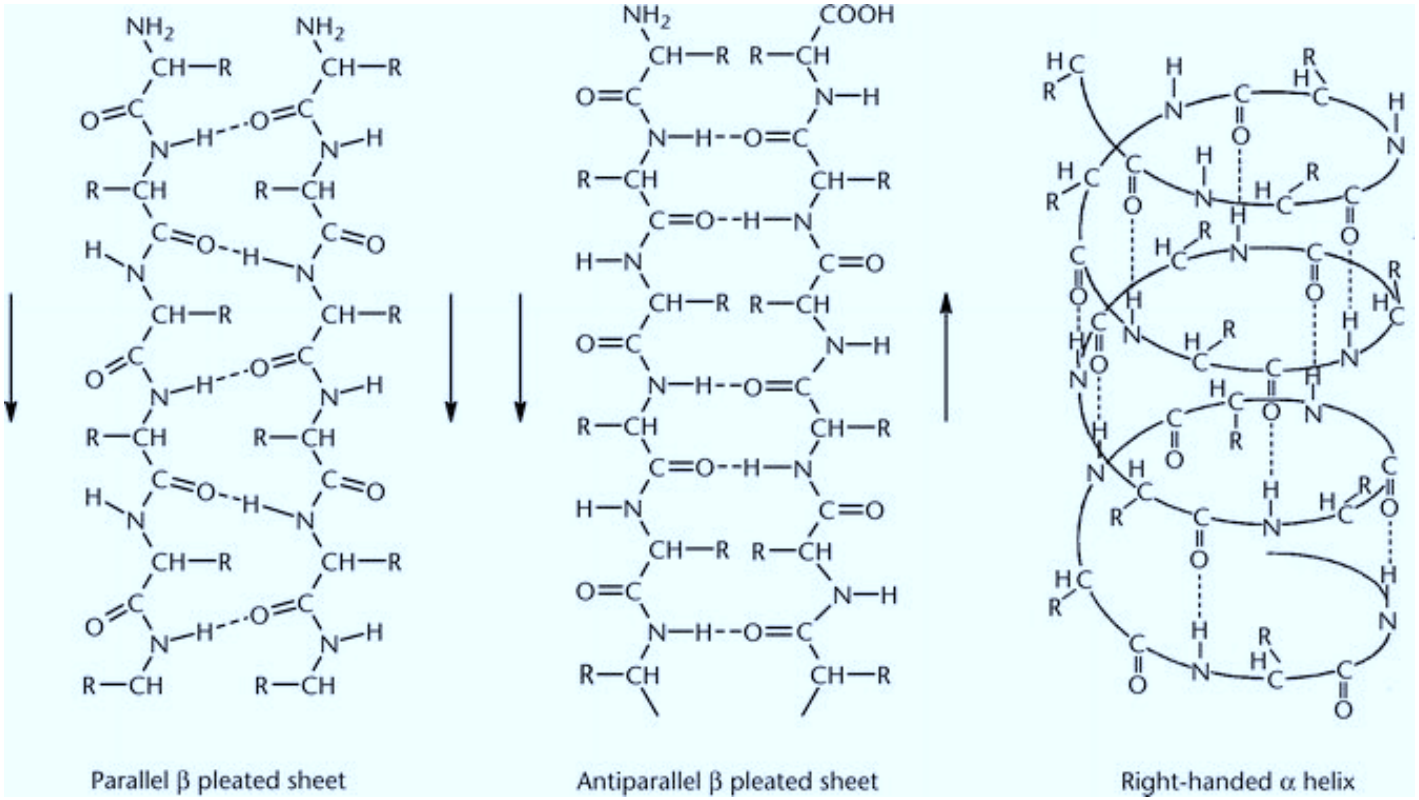
What is Protein? (Primary Structure)





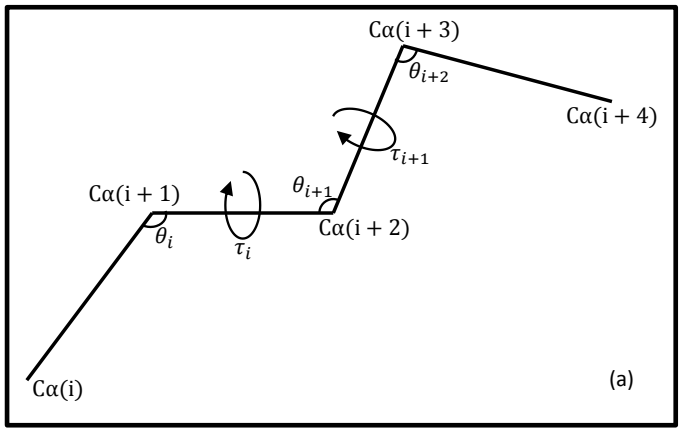
source: Wikipedia

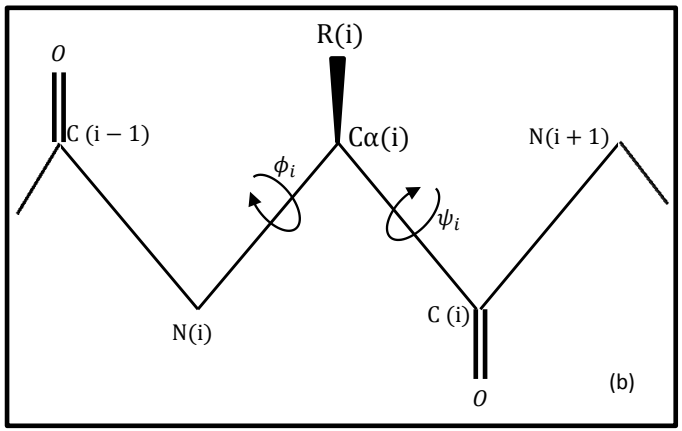
What is Protein? (Secondary Structure)

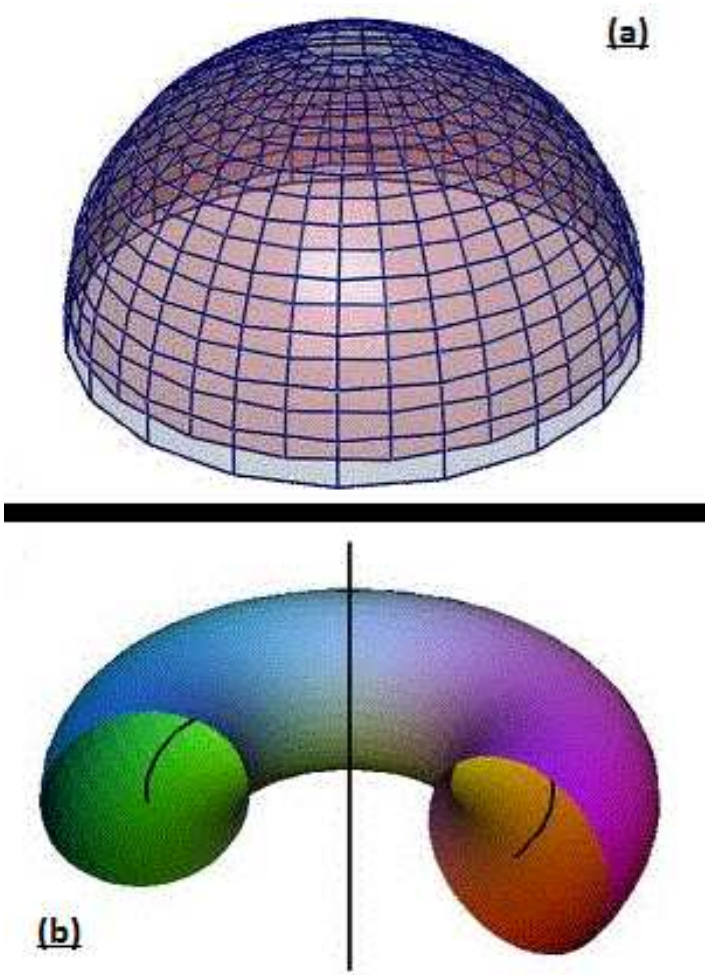


source: Encyclopedia of Life Sciences

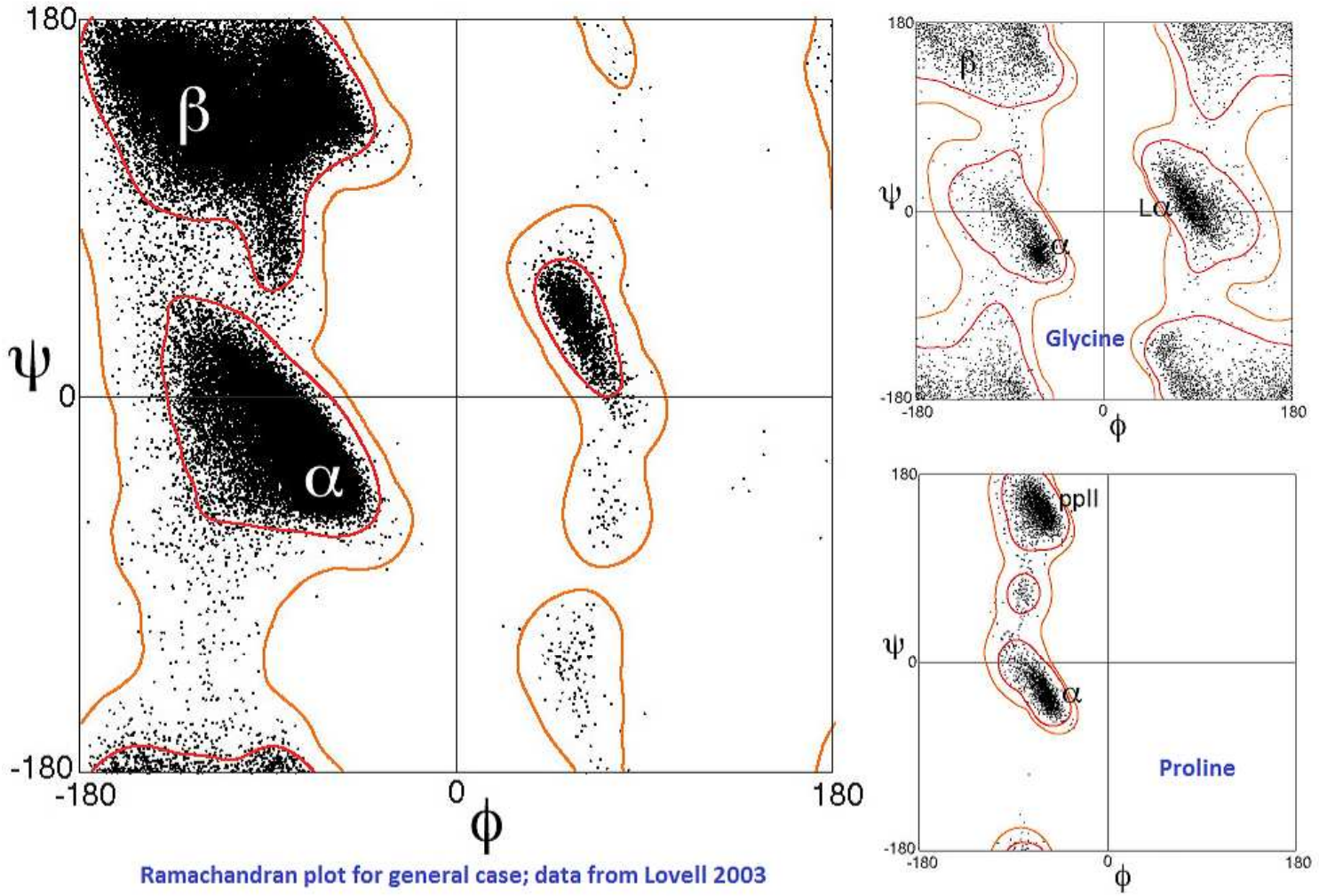
What is Protein? cont.





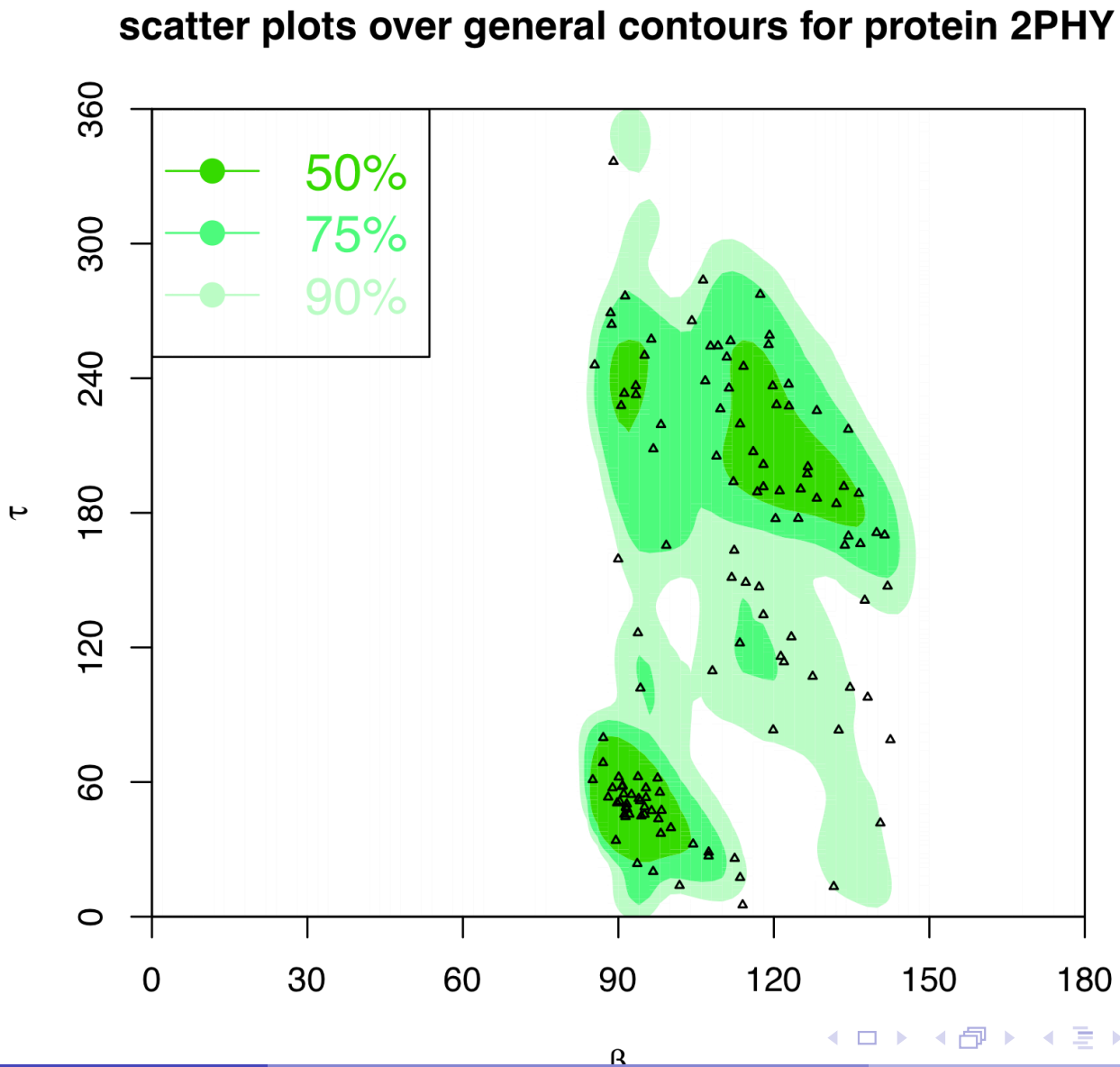


Ramachandran plot

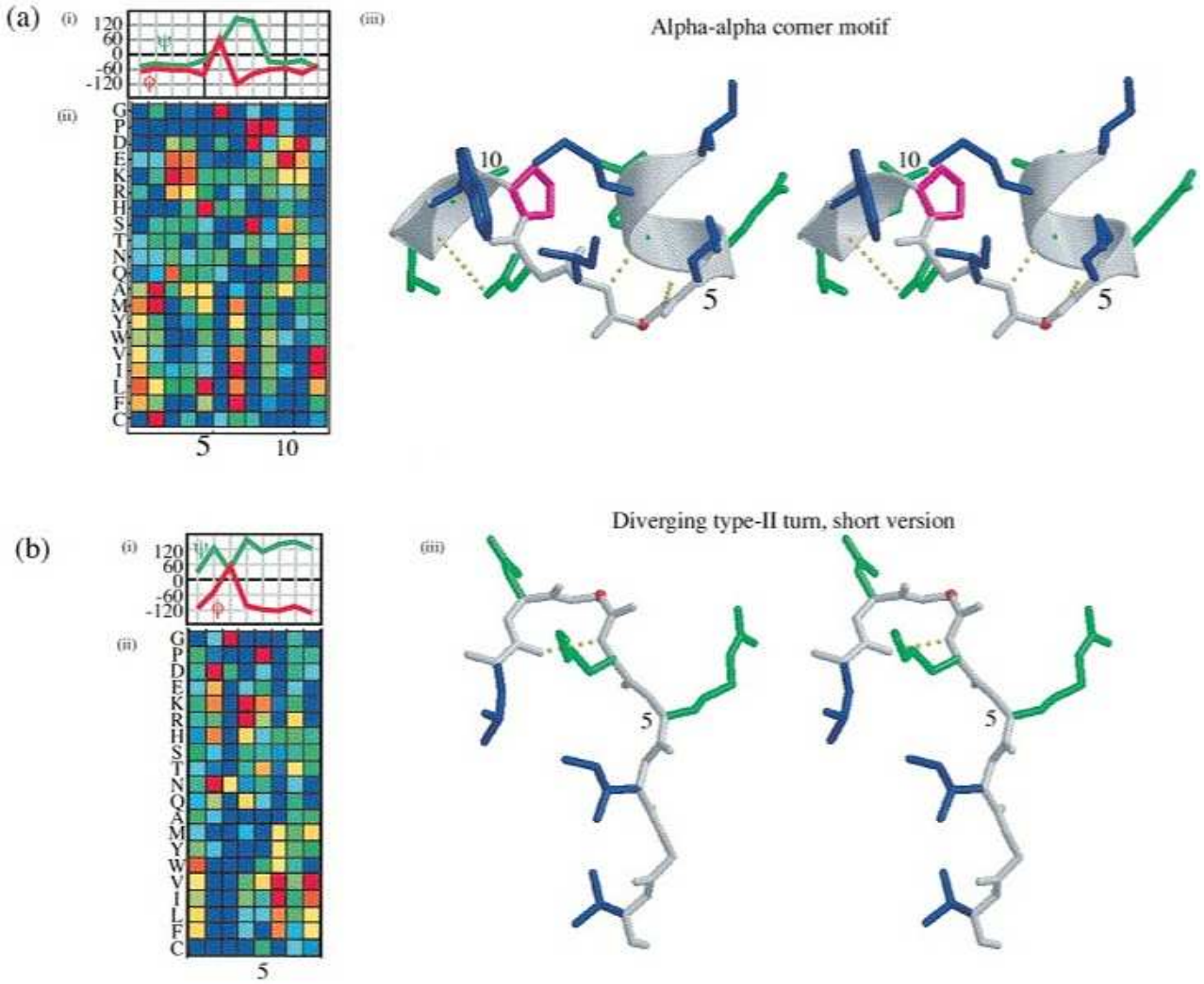


source: Wikipedia

Ramachandran Plot cont.

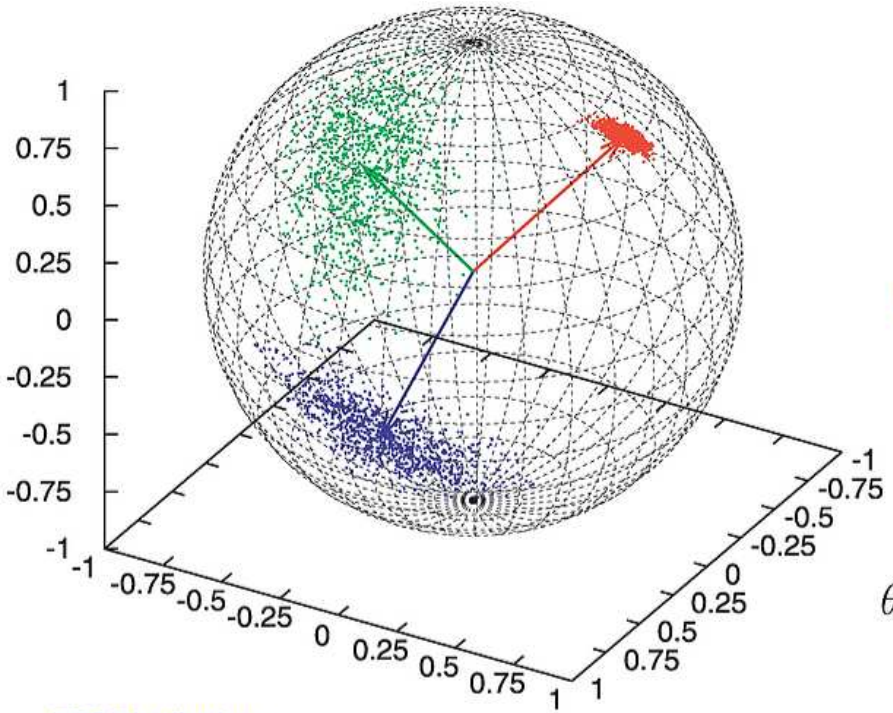


What is HMMSTR? An HMM for local sequence-structure



source: Bystroff *et al.* (2000)

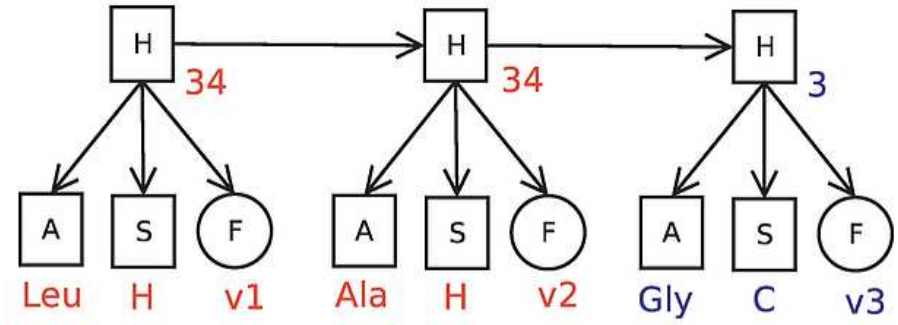
What is FB5-HMM? (Kent-HMM)

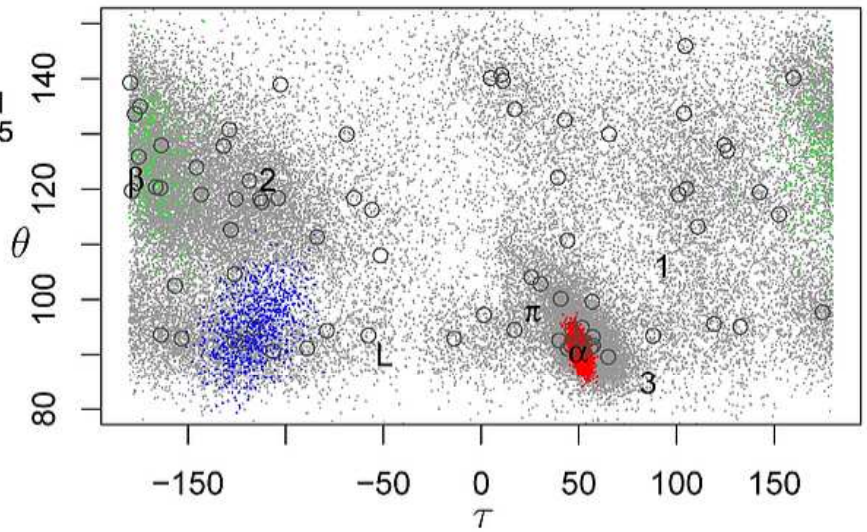


FB5 Distribution:

$$f(\mathbf{x}) = \frac{1}{c(\kappa,\beta)} \exp\{\kappa\gamma_1 \cdot \mathbf{x} + \beta[(\gamma_2 \cdot \mathbf{x})^2 - (\gamma_3 \cdot \mathbf{x})^2]\}$$

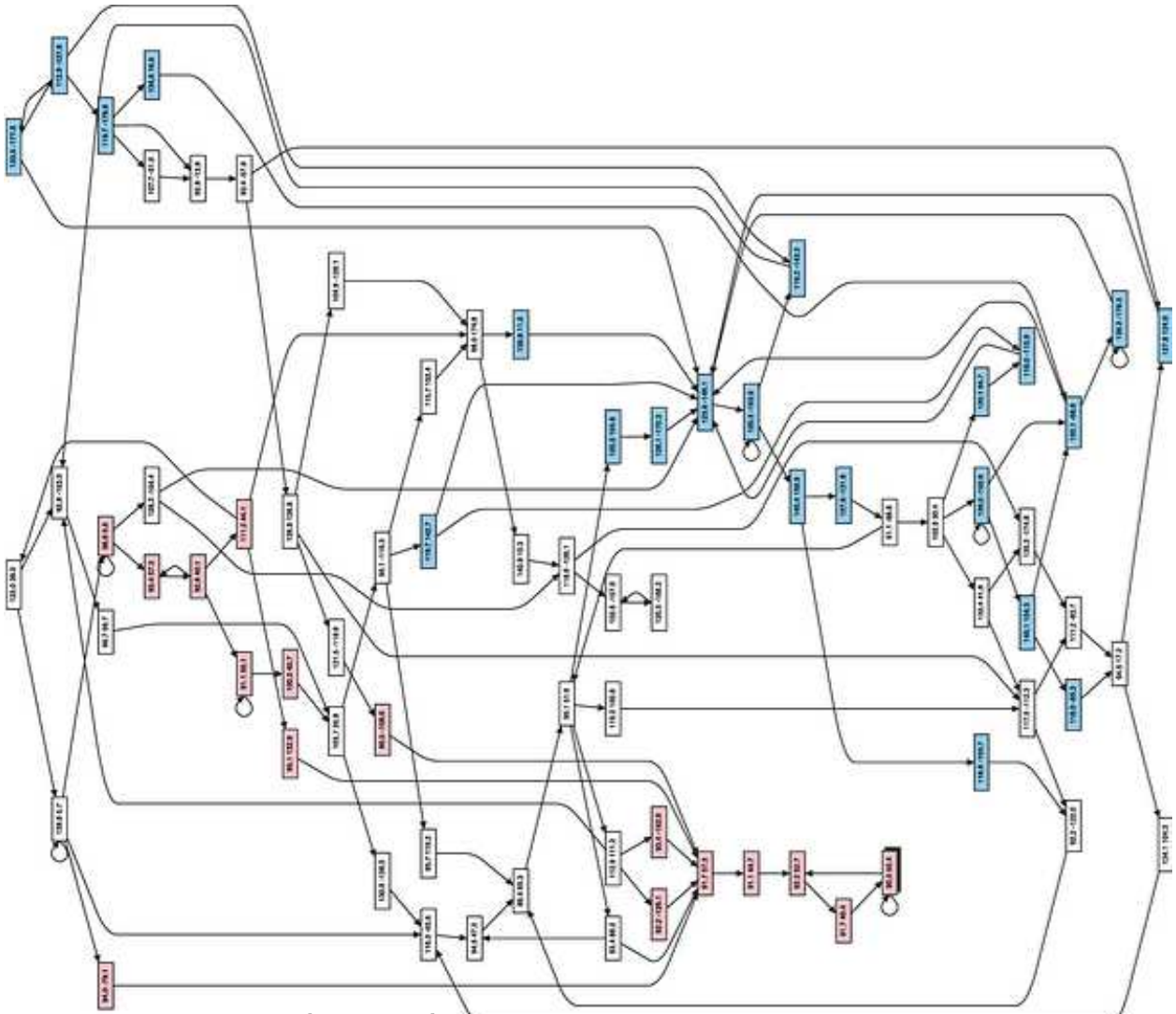
$$x = \cos(\theta)$$
$$y = \sin(\theta)\cos(\tau)$$
$$z = \sin(\theta)\sin(\tau)$$





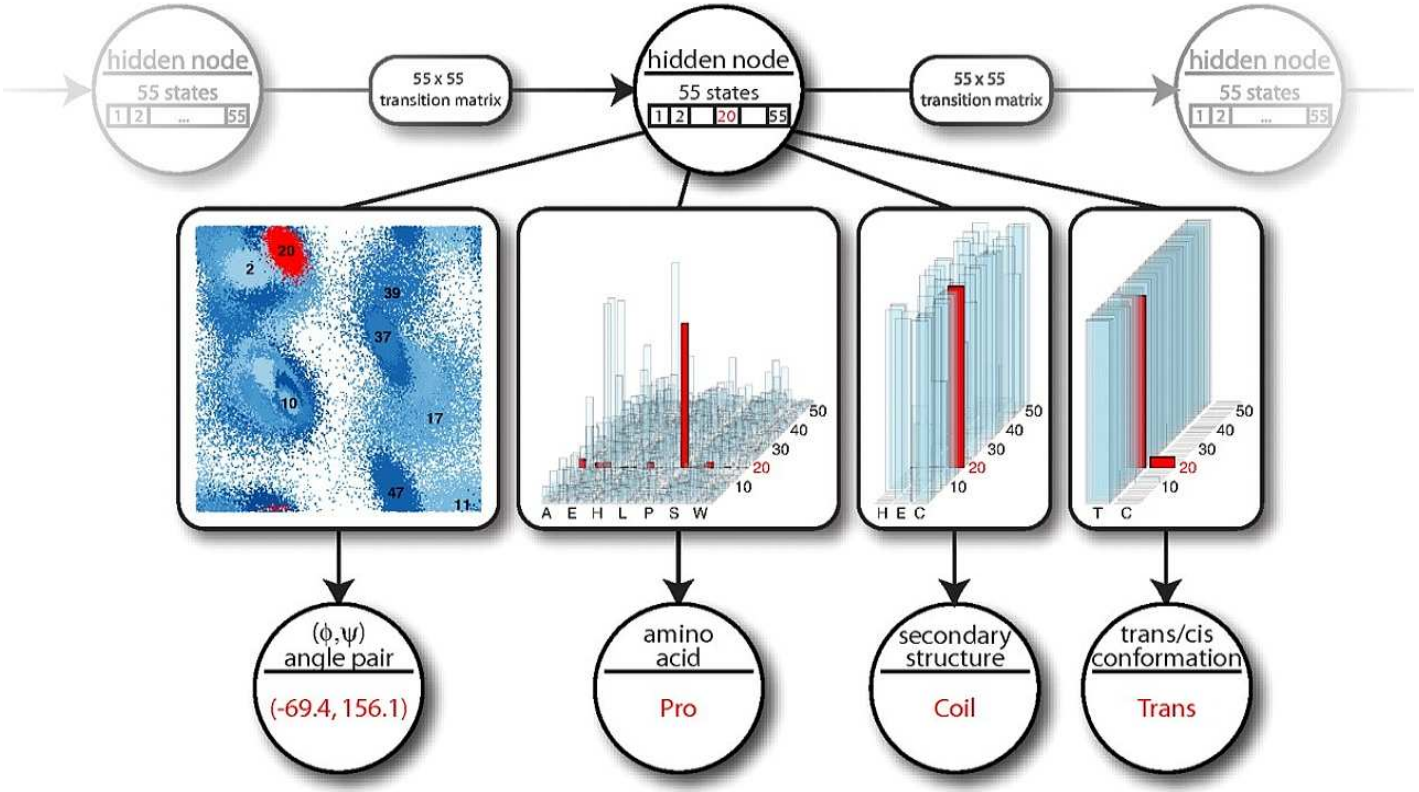
source: Hamelryck *et al.* (2006)

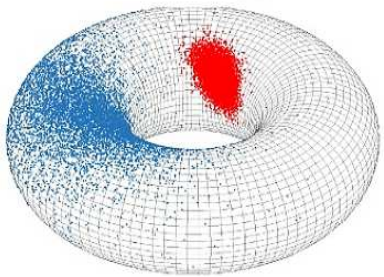
Why FB5-HMM?



source: Hamelryck *et al.* (2006)

What is BVM-HMM? (Bivariate von Mises-HMM)





General Form:

$$f(\phi,\psi) \propto \exp[\kappa_1 \cos(\phi-\mu)+\kappa_2 \cos(\psi-\nu) + (\cos(\phi-\mu), \sin(\phi-\mu))\mathbf{A}(\cos(\psi-\nu), \sin(\psi-\nu))^T],$$

Cosine Form:

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

source: Boomsma *et al.* (2008)

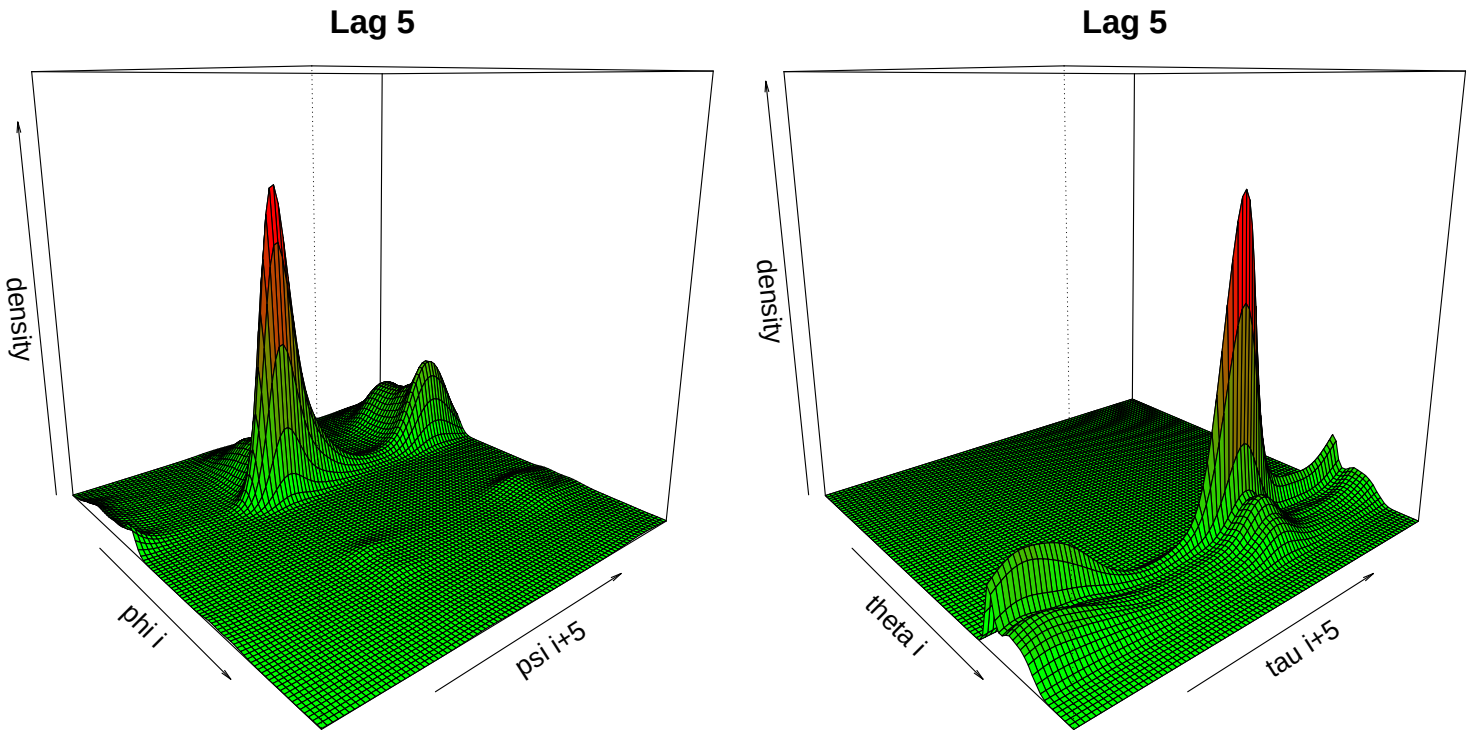
What do we mean by Lag?

- HMM is essentially lag-one modeling
- The goal is to go beyond lag-one dependence ($\ell = 1, \ldots, L$)
- For protein j, the collection of pair of backbone angles

$$\{(\eta_1, \zeta_{1+\ell})^\top, (\eta_2, \zeta_{2+\ell})^\top, \cdots, (\eta_{n_j-\ell}, \zeta_{n_j})^\top\}$$

  can be viewed as a random draw from the lag-ℓ marginal bivariate density
- Denote that density as $f^{(\ell)}(\eta, \zeta)$

Bivariate Kernel Density Estimate of the Lag-Distributions



- Using bivariate kernel density estimator:

$$\hat{f}^{(\ell)}_{\mathbf{H}}(\eta,\zeta) = \frac{\sum_{j=1}^{N}\sum_{i=1}^{n_j-\ell}\varphi\Big(\frac{\eta \ominus \eta_{j,i}}{h_\eta}\Big)\varphi\Big(\frac{\zeta \ominus \zeta_{j,i+\ell}}{h_\zeta}\Big)}{\{\sum_{j=1}^{N}(n_j-\ell)\}h_\eta h_\zeta}, \qquad (1)$$

Bivariate Kernel Density Estimates (different lags)

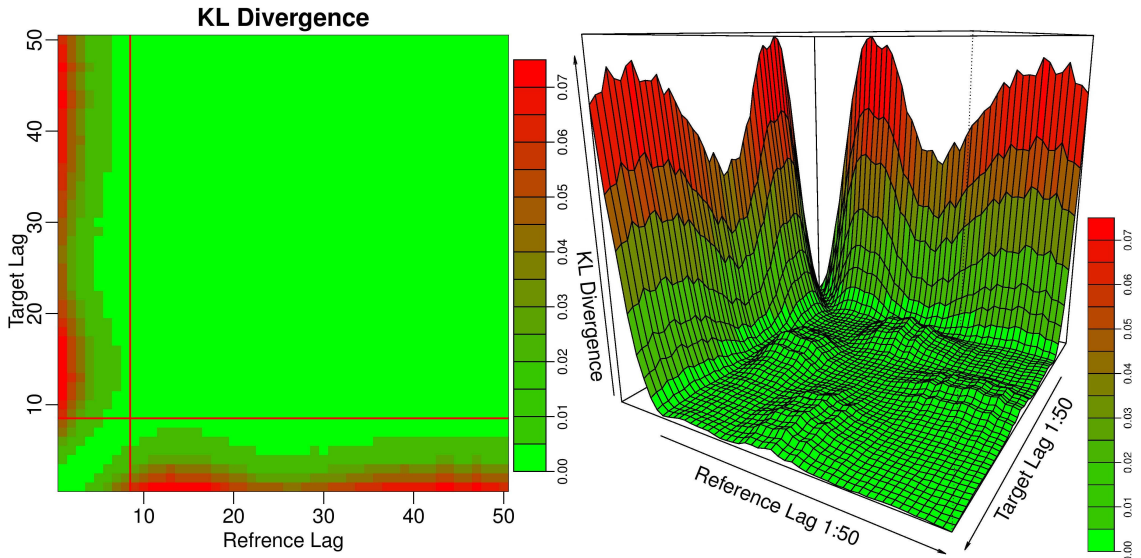
Kullback-Leibler Divergence of the Lag-Distribution

- We denote the KL Divergence between $\hat{f}^{(\ell)}$ and $\hat{f}^{(\ell')}$ as

$$D_{KL}(\hat{f}^{(\ell)}||\hat{f}^{(\ell')}) = \iint \hat{f}^{(\ell)}(\eta,\zeta) \ln \frac{\hat{f}^{(\ell)}(\eta,\zeta)}{\hat{f}^{(\ell')}(\eta,\zeta)} \mathrm{d}\eta\mathrm{d}\zeta$$

- We symmetrised the divergence to obtain the following:

$$D_{KL}(\hat{f}^{(\ell)}, \hat{f}^{(\ell')}) = D_{KL}(\hat{f}^{(\ell)}||\hat{f}^{(\ell')}) + D_{KL}(\hat{f}^{(\ell')}||\hat{f}^{(\ell)}) \qquad (2)$$



Singular Value Decomposition of the Lag-Distribution

- Let's factorize each bivariate distribution to sum of m multiplicative univariate densities

$$\hat{f}^{(\ell)}(\eta, \zeta) = \sum_{k=1}^{m} \sigma_k u_k^{(\ell)}(\eta) v_k^{(\ell)}(\zeta) + \epsilon$$

- Later, we focus on $(u_k^{(1)}, \cdots, u_k^{(L)})$, and $(v_k^{(1)}, \cdots, v_k^{(L)})$ for each k $(k = 1, \cdots, m)$

- Considering the density values in grids over the angular plane, we obtain a $d \times d$ matrix $\hat{F}^{(\ell)}$

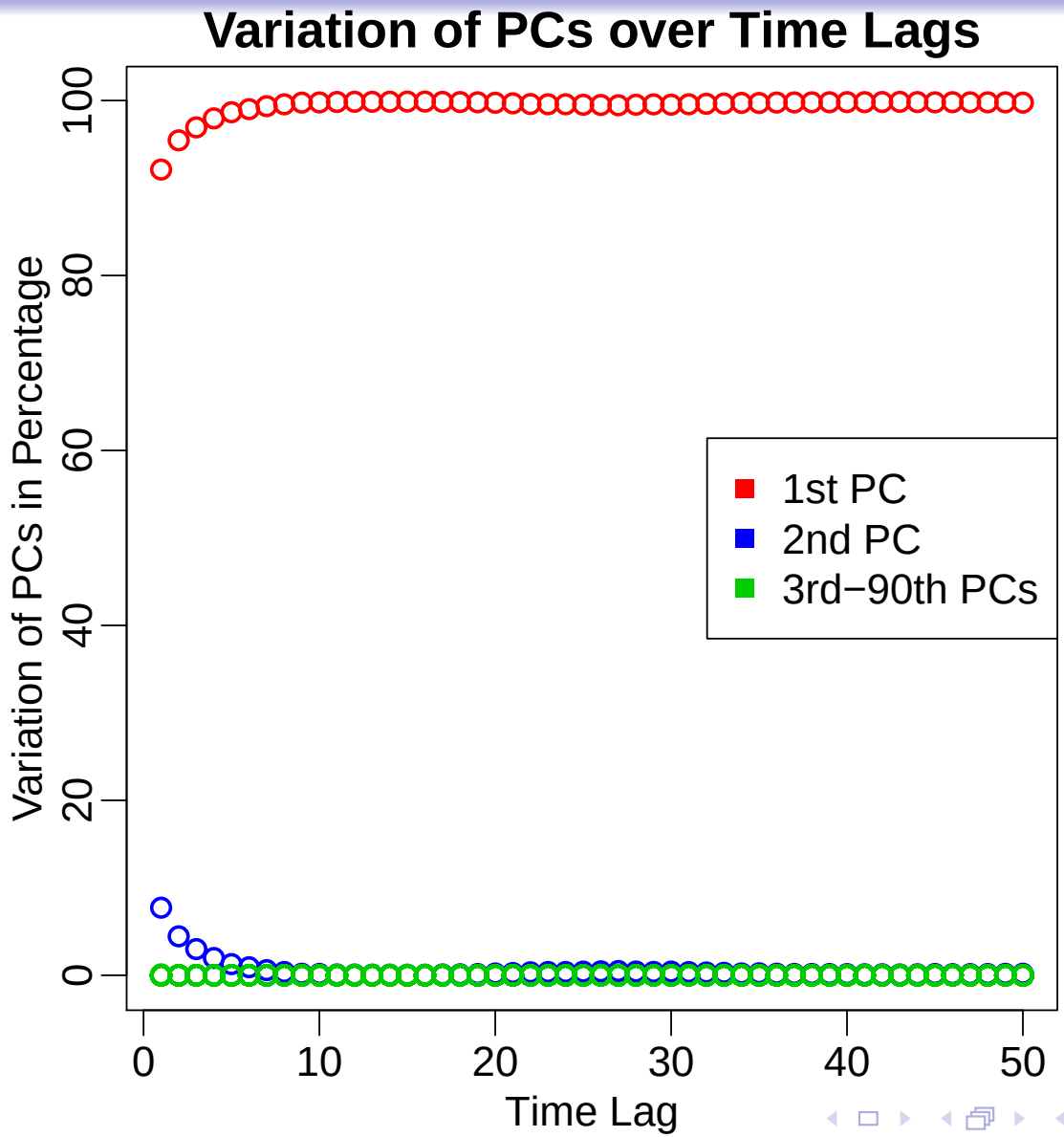
- We may use SVD to obtain the low-rank approximation of $\hat{F}^{(\ell)}$

$$\hat{F}^{(\ell)} = \mathbf{U}_\eta^{(\ell)} \mathbf{\Sigma}^{(\ell)} \mathbf{V}_\zeta^{(\ell)\top}$$

Density Estimates or Square Root of the BVKDE

Scaled SVD or non-Scaled SVD?

Scree Plot over Different Lags



How to Use LagSVD for Model Assessment?

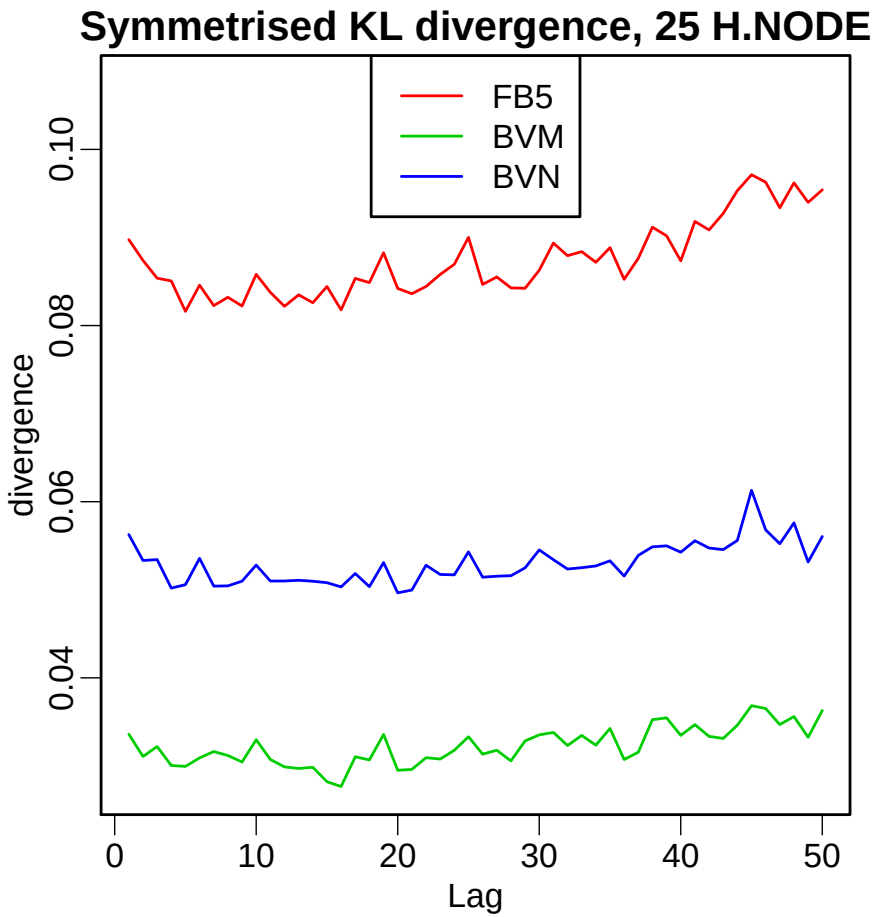
- To evaluate the performance of M different models $(m = 1, \cdots, M)$
- After learning the M models, we simulate protein structures from each
- Assessment measures can be motivated by comparing the output of the LagSVD of each simulation versus the LagSVD of the real protein from PDB
- A simple measurement for assessing model m in lag ℓ, Symmetrised KLD:

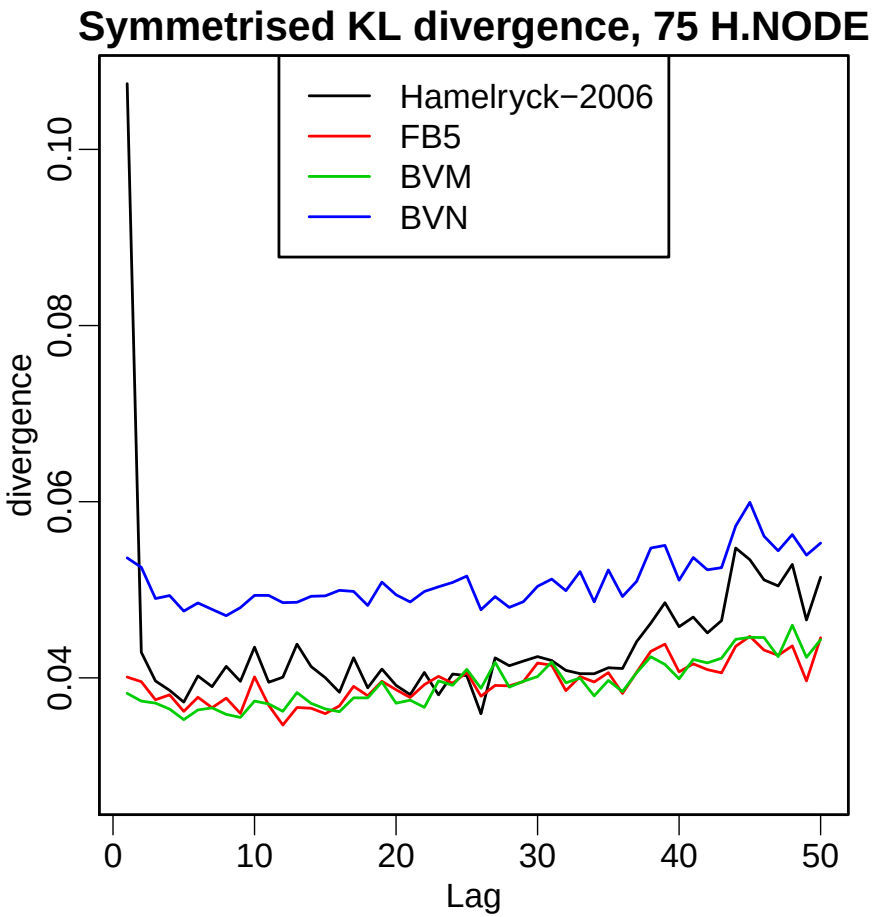
$$\mathrm{SKLD}_m^{(\ell)} = D_{KL}(\hat{f}_r^{(\ell)}, \hat{f}_m^{(\ell)})$$

How to do Model Assessment? cont.

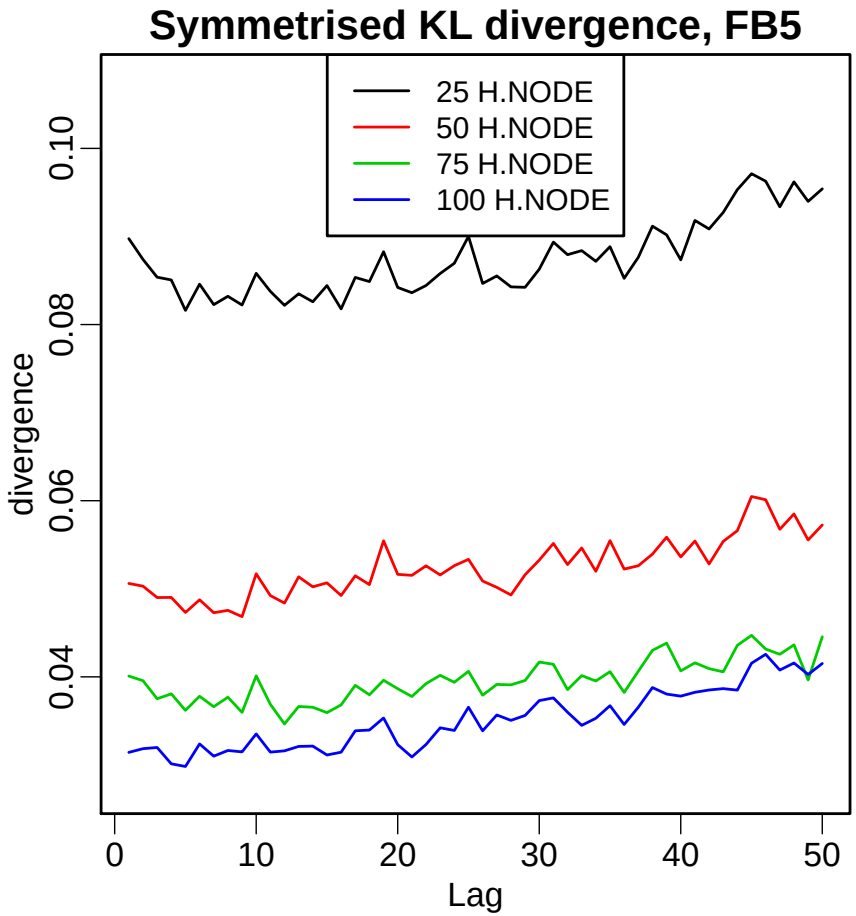
- For learning the proposed HMM models we consider:
- The following observed nodes:
  - amino-acid sequences
  - secondary structures
  - θ, τ angles
- Four different number of hidden nodes
  - $H = \{25, 50, 75, 100\}$
- Three HMMs:
  - (a) FB5-HMM
  - (b) BVM-HMM
  - (c) BVN-HMM
- Therefore we have 12 trained models, and for each trained model we simulate 100 sequence of proteins, each with length 100

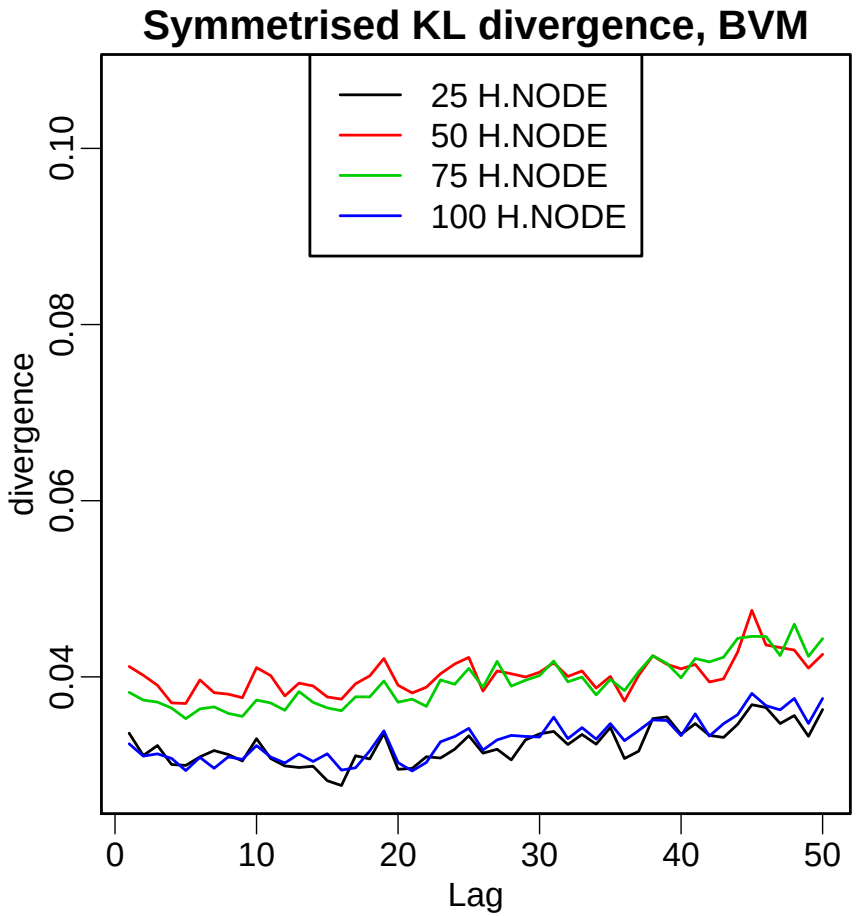
How to do Model Assessment? cont.





How to do Model Assessment? cont.





How to do Model Assessment? cont.

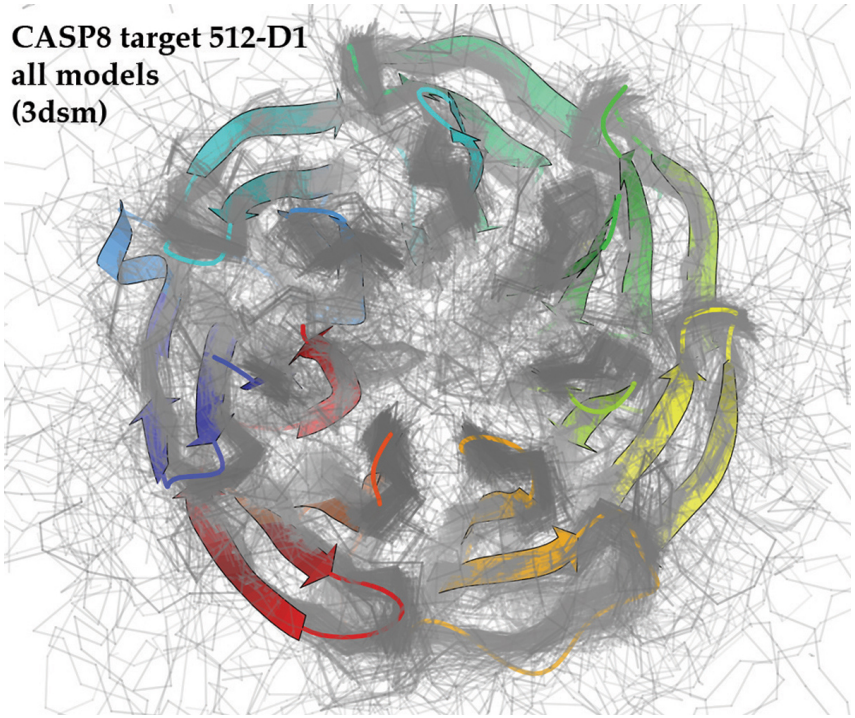
Table : Comparisons of symmetrised KL divergence (SKLD) between the HMM models with FB5, BVN, and BVM angular distributions for different numbers of hidden nodes

Number of hidden nodes	Model Distribution FB5	BVN	BVM
H=25	0.08473 (0.00944)	0.05220 (0.00598)	0.03140 (0.00158)
H=50	0.04884 (0.00340)	0.05106 (0.00749)	0.03890 (0.00566)
H=75	0.03796 (0.00159)	0.04928 (0.00486)	0.03661 (0.00551)
H=100	0.03152 (0.00281)	0.04571 (0.00509)	0.03090 (0.00281)

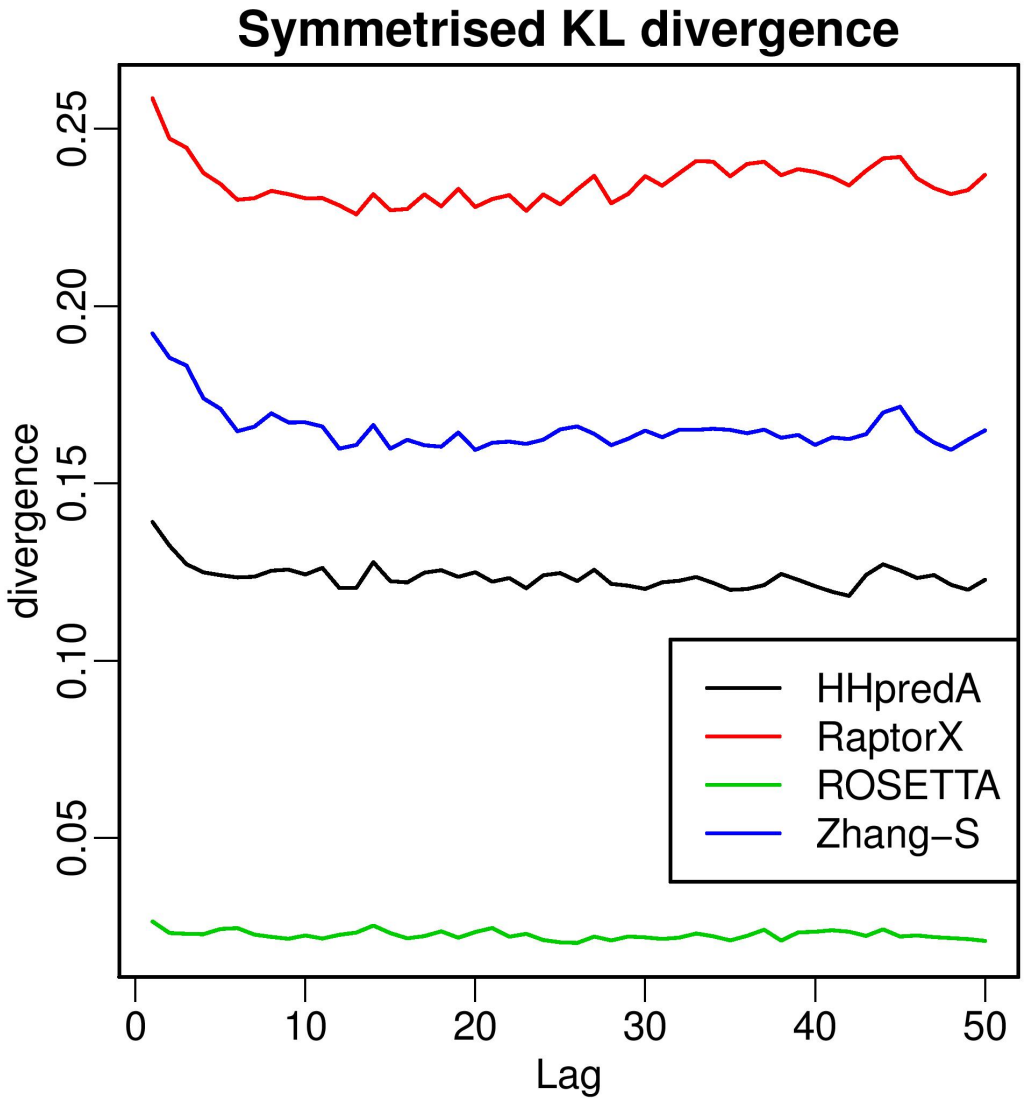
Critical Assessment of Protein Structure Prediction

Amino acid sequence (328 AA):

ASGLFITNEGNFQYSNATLSYYDPATCEVENEVFYRANGFKLGDVAQSMVIRDGIGWIVVNNSHVIFAIDINTFKEVGRITGFTSPRYI
HFLSDEKAYVTQIWDYRIFIINPKTYEITGYIECPDMDMESGSTEQMVQYGKYVYVNCWSYQNRILKIDTETDKVVDELTIGIQPTS
LVMDKYNKMWTITDGGYEGSPYGYEAPSLYRIDAETFTVEKQFKFKLGDWPSEVQLNGTRDTLYWINNDIWRMPVEADRVPVR
PFLEFRDTTKYYGLTVNPNNGEVYVADAIDYQQQGIVYRYSPQGKLIDEFYVGIIPGAFCWKLEHHHHHH



LagSVD for CASP9.



Conclusions and Future Works

- Use of LagSVD to explore the dependence structure of dihedral/planar angles in lower dimensions
- Marginal bivariate lag-distributions can be used for model assessment.
- The dependency structure of the dihedral/planar angles vanishes for the lags beyond $nine$ in the real protein structures.
- The commonly used FB5 model is not necessarily the best option.
- Future Works
  - Nonparametric density estimation of the angular data over spherical domain using bivariate B-splines.
  - Explore the dynamic of variations over the lags by using Functional PCA

Thank You

References

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, **105**(26), 8932–8937.

Bystroff, C., Thorsson, V., and Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins1. *Journal of Molecular Biology*, **301**(1), 173–190.

Hamelryck, T., Kent, J. T. T., and Krogh, A. (2006). Sampling Realistic Protein Conformations Using Local Structural Bias. *PLoS Comput Biol*, **2**(9).

Maadooliat, M., Gao, X., and Huang, J. Z. (2013). Assessing protein conformational sampling methods based on bivariate lag-distributions of backbone angles. *Briefings in Bioinformatics*, **to Appear**.