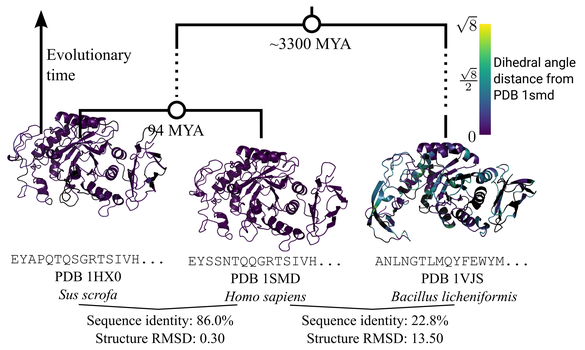
Probabilistic modelling of protein structure evolution

Michael Golden[1,*], Thomas Hamelryck[2], Jotun Hein[3], Oliver Pybus[1]

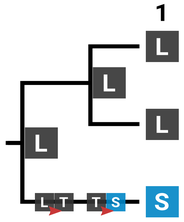
1. Department of Zoology, University of Oxford 2. Department of Statistics, University of Oxford 3. Bioinformatics Centre, University of Copenhagen

Introduction: Evolutionary modelling

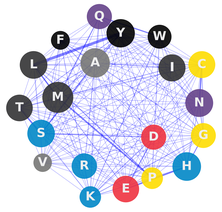
Homologous proteins are correlated due to their shared ancestry (below). Notice how the two amylase proteins on the left are similiar in sequence and structure due to their fairly recent ancestry, whilst the two amylase proteins on the right are more diverged due to them sharing a common ancestor much further back in time. Many substitution models for modelling protein sequence evolution exist, whilst analogous models for protein structure are far more limited.



Substitution models for sequence evolution

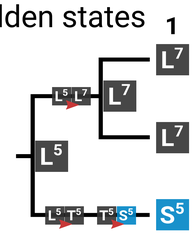


Phylogenetic tree showing the relationship between three sequences for a single site. Also shown on the tree is the substitution history.

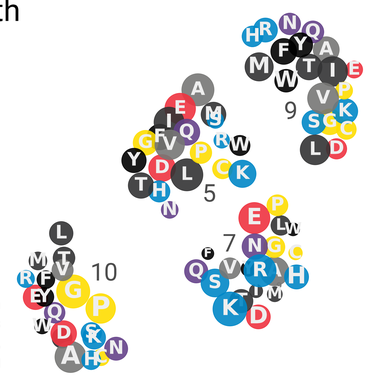


The LG2008 subsitution model is a rate matrix on amino acids that gives a probabilistic description of amino acid substitution histories.

Substitution model with hidden states

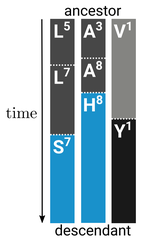


Introducing hidden states provides a richer parameterisation that captures short term site-specific evolutionary regimes (such as hydrophobic sites, basic amino acids, etc.), whilst permitting longer term switching between evolutionary regimes.

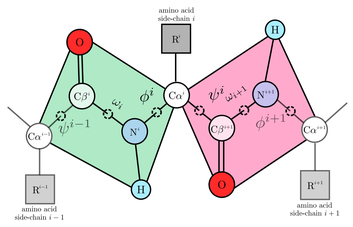


Hidden states with neighbouring dependencies

Standard substitution models do not consider neighbouring dependencies, but instead treat each site as evolving independently. Site-independence ignores obvious features of sequence, such as hydrophobic stretches. Such neighbouring dependencies are critical when considering constraints imposed by protein structure (see Ramachandran plots).



Angle representation



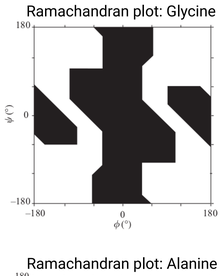
The complete 3D backbone structure of a protein can be exactly specified using three dihedral angles, three additional bond angles and three bond lengths.

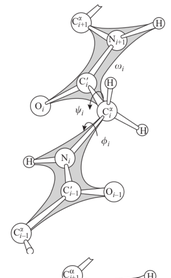
The phi (ϕ) and psi (ψ) dihedral angles are the most important degrees of freedom, because they vary the most. Each amino acid in the peptide backbone is associated with a phi and psi angle (except the N- and C-terminus amino acids).

Ramachandran plots

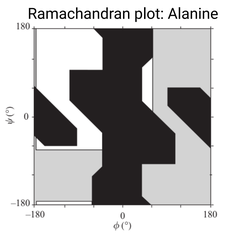
The possible phi-psi conformations depend on the associated amino acid side-chain, as well as neighbouring amino acids, as depicted in the Ramachrandran plots below:

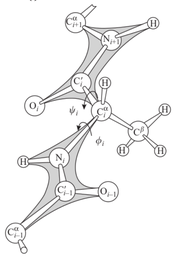
Glycine is the smallest amino acid and therefore the least constrained.



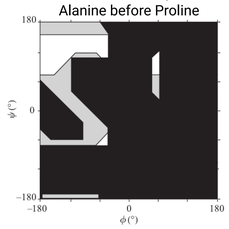


Alanine has a slightly larger side-chain than Glycine, and is therefore slightly more constrained.





Neighbouring amino acids further constrain dihedral angles. Compare to Alanine above.

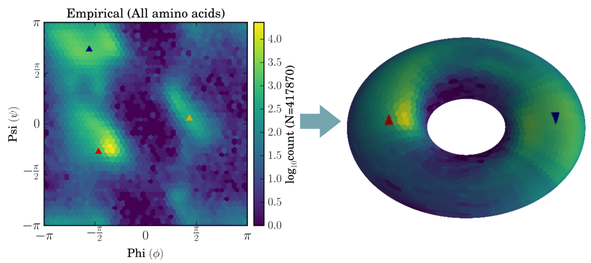


Legend

- Allowed
- Disallowed
- Strained

Density and periodicity

Empirically, dihedral angle constraints are soft like those pictured below, rather than the hard constraints illustrated above, and therefore they require probabilistic treatment. However, when doing so, one needs to account for the periodic nature of angles.

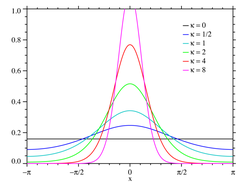


Finkelstein, Alexei V., and Oleg Ptitsyn. Protein physics: a course of lectures. Elsevier, 2016.

Angular distributions

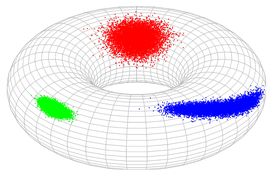
The periodic nature of angles is dealt with by using analogues of the univariate and bivariate normal distributions for modelling their densities. The univariate and bivariate von Mises distributions, respectively:

Univariate von Mises distribution



$$f(x \mid \mu, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$$

Bivariate von Mises distribution

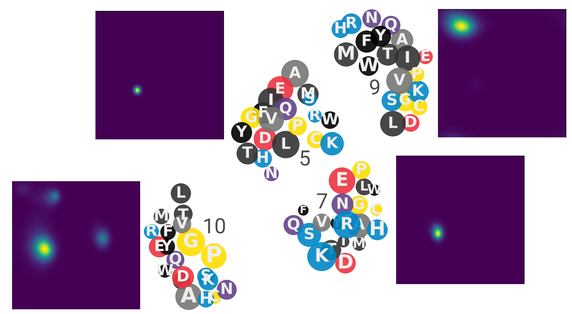


Samples from three bivariate von Mises distributions

Mardia, Kanti. Statistics of directional data. *J. Roy. Statist. Soc. B.*, 1975.

Angular homology modelling

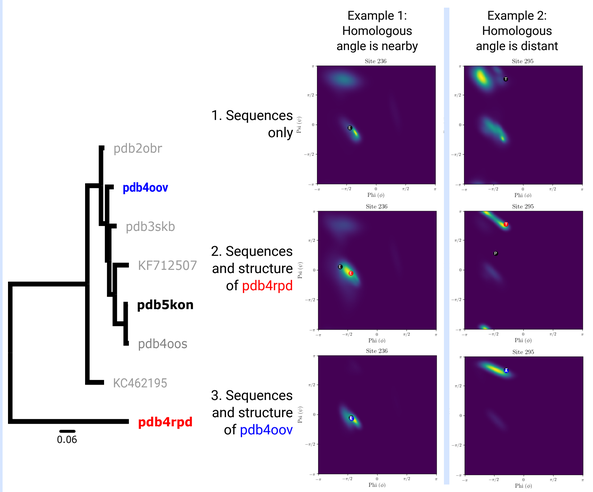
Angle evolution is modelled by associating a bivariate von Mises distribution with each hidden state and amino acid (H x 20 bivariate von Mises distributions). Evolutionary trajectories over hidden states and amino acids therefore give trajectories over dihedral angles and protein structures.



Example: angle predictions

Our model gives a complete probabilistic description of of protein sequence and protein backbone structure evolution along a phylogenetic tree. Amongst many other tasks, this enables structure prediction (or more accurately: homology modelling) using flexible combinations of sequence or structure observations.

Prediction of angles in **pdb5kon**



Angular distance

The cosine angular distance (Downs and Mardia, 2002) was used to measure distances between pairs of angles:

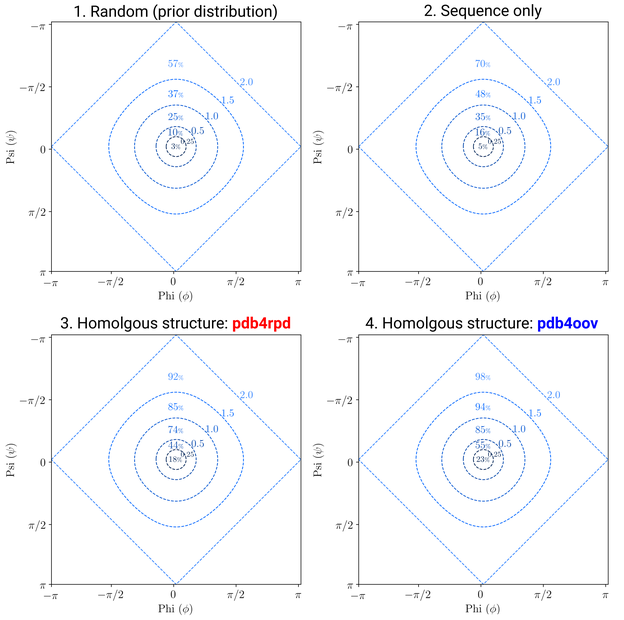
$$d(\langle\phi_a, \psi_a\rangle, \langle\phi_b, \psi_b\rangle) = \sqrt{4 - 2\cos(\phi_a - \phi_b) - 2\cos(\psi_a - \psi_b)}.$$

When $\phi_a - \phi_b \approx 0$ and $\psi_a - \psi_b \approx 0$ are small, the small angle approximation for cosine ($\cos\theta \approx 1 - \frac{\theta^2}{2}$) can be used, giving the Euclidean distance:

$$d(\langle\phi_a, \psi_a\rangle, \langle\phi_b, \psi_b\rangle) \approx \sqrt{4 - 2(1 - (\phi_a - \phi_b)^2/2) - 2(1 - (\psi_a - \psi_b)^2/2)}$$
$$= \sqrt{(\phi_a - \phi_b)^2 + (\psi_a - \psi_b)^2}.$$

Predictive accuracy

The cosine angular distance was used to compare our model's predictions of phi-psi angles in **pdb5kon** under four increasingly informative observation conditions (1-4). The plots below show the percentage of phi-psi angles correctly predicted within five different cosine angular distance radii (0.25, 0.5, 1.0, 1.5, and 2.0; dashed blue lines). If the centre of the plot is taken to be the true angle, the goal is to the maximise the percentage of angles predicted within the innermost radii.



Future perspective: 3D homology modelling and ancestral structure reconstruction

Errors in angle predictions can lead to large errors in 3D structure prediction. For example: a single incorrectly predicted angle can cause an entire structure to self-intersect in 3D.

State-of-the-art machine learning models of protein structure prediction (DeepMind's AlphaFold and AlQuraishi's end-to-end differentiable model) utilise angle models for predicting an intermediate angle representation. Angle models are used due to their ability to capture local structural conformations, although they have the same problem with angle errors. These state-of-the-art models correct errors in 3D structure due to errors in the angles using subsequent modelling layers, resulting in improved 3D structure predictions.

It should be easy to incorporate our model as a layer into these existing models. The use of homology information should enhance these models' ability to predict 3D structures. Furthermore, using an explicit evolutionary model permits the 3D reconstruction of ancestral protein structures.

*Corresponding author: golden@phylo.dev