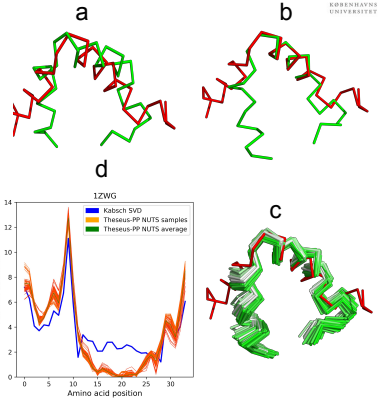
Bayesian protein superposition using Hamiltonian Monte Carlo

Lys Sanz Moreta[1], Ahmad Salim Al-Sibahi[1], Thomas Hamelryck[1,2]

[1] Programming languages and theory of computing section, Department of computer science.
[2] Section for computational and RNA biology, Department of biology.

- Optimal superposition of protein structures is crucial for their comparison, which is necessary for understanding their **structure, function, dynamics and evolution**.
- We adapt THESEUS [1,2], which provides a ML estimate, to **MAP estimation** using **SVI** [3] and Bayesian posterior inference using **Hamiltonian Monte Carlo (NUTS).**
- The model superimposes two structures by **rotating (R)**, **translating (T)** and **perturbing (U)** an underlying **latent mean structure (M)**, using suitable priors.
- We used the deep probabilistic programming languages **Pyro** and **Numpyro**.
- Unlike conventional methods that minimize the sum of the squared distances, THESEUS takes into account **correlations** and **heteroscedasticity** of the atomic positions.
- The model can serve as a **likelihood** in **Bayesian protein structure prediction**.



Theseus-PP model

▷ *Prior over the elements of* $\mathbf{M}$
$\mathbf{m_i} \sim t_1(\mathbf{0}, \sigma_M \mathbf{I}_3)$, where i indicates the atom position
▷ *Prior over the translation*
$\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$
▷ *Prior over rotation*
$\mathbf{u}_j \sim U[0,1]$, for j from 0 to 2
$\mathbf{q} \leftarrow \text{Quaternion}(\mathbf{u})$
$\mathbf{R} \leftarrow \text{RotationMatrix}(\mathbf{q})$
▷ *Prior over diagonal covariance matrix* $\mathbf{U}$
$\sigma_i \sim \mathcal{N}_+(1)$
▷ *Likelihood over the* N *atom coordinates*
$\mathbf{x}_{1,i} \sim t_1(\mathbf{m}_i, \sigma_i \mathbf{I}_3))$
$\mathbf{x}_{2,i} \sim t_1([\mathbf{RM} - \mathbf{1}_N \mathbf{T}]_i, \sigma_i \mathbf{I}_3)$

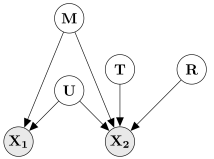


FIGURE 1 : Protein superposition for two conformations of protein **1ZWG** obtained from a) conventional RMSD superimposition, b) THESEUS MAP and c) THESEUS NUTS. The protein in green ($\mathbf{X_2}$) is rotated and translated. The images are generated with PyMOL. Graph d) shows the pairwise distances (in Å) between the Cα coordinates of the structure pairs for MAP and iterative NUTS (orange) versus conventional RMSD (blue).

[1] Theobald DL, Wuttke DS, THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* (2006)
[2] Theobald DL, Wuttke DS, Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proceedings of the National Academy of Sciences USA* (2006)
[3] Moreta LS *et al.*, A probabilistic programming approach to protein structure superposition. *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Siena, Italy (2019)