Bayesian protein superposition using Hamiltonian Monte Carlo

Lys Sanz Moreta¹, Ahmad Salim Al-Sibahi¹, Thomas Hamelryck¹,²

¹Programming languages and theory of computing section, Department of computer science.
²Section 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.

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 (X₂) is rotated and translated. The images are generated with PyMOL. Graph d) shows the pairwise distances (in Å) between the Ca coordinates of the structure pairs for MAP and iterative NUTS (orange) versus conventional RMSD (blue).