The Bayesian Occam's razor for the hierarchical clustering of MSM's

Ben Harland

October 5, 2019

Occam's razor

"Entities should not be multiplied without necessity."

William of Ockham (c. 1287–1347)

"Everything should be made as simple as possible, but no simpler."

Einstein (New York Times, 1950)

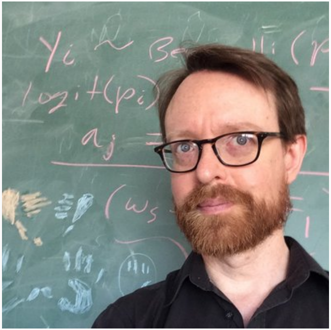
"It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience."

The Ultimate Quotable Einstein

Helpful books (youtube)

David MacKay

Information Theory, Inference, and Learning Algorithms



Richard McElreath

Statistical Rethinking

Hierarchical protein dynamics

REPORT

Direct observation of hierarchical protein dynamics

Józef R. Lewandowski[1,*,†], Meghan E. Halse[1,‡], Martin Blackledge[2,*], Lyndon Emsley[1,3,*]

+ See all authors and affiliations

Science 01 May 2015:
Vol. 348, Issue 6234, pp. 578-581
DOI: 10.1126/science.aaa6111

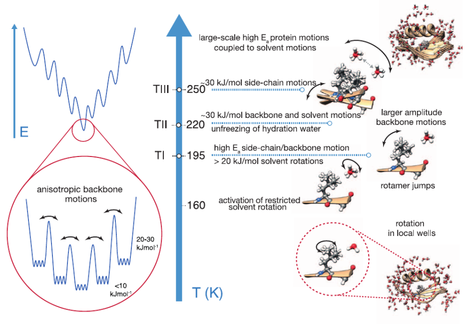
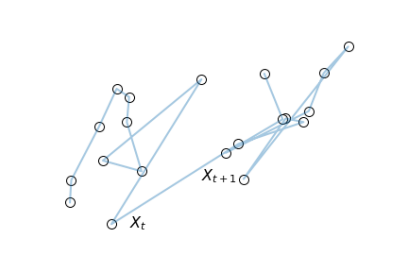


Fig. 4. Summary of hierarchical dynamic behavior of the protein-solvent system as observed by solid-state NMR in a microcrystalline globular protein GB1. The approximate temperature for the

Markov models

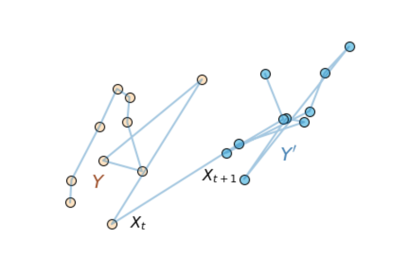
Microstates (or macrostates)



$$X_1 \longrightarrow \cdots \quad X_t \longrightarrow X_{t+1} \quad \cdots \longrightarrow X_N$$

$$P(X_{t+1} = x_j | X_t = x_i) = T_{ij}$$

Macrostate/microstate model



$$\begin{tikzcd}
X_1 & & X_t & X_{t+1} & & X_N \\
Y_1 \arrow[u] \arrow[r] & \cdots & Y_t \arrow[u] \arrow[r] & Y_{t+1} \arrow[u] & \cdots \arrow[r] & Y_N \arrow[u]
\end{tikzcd}$$

$$P(X_{t+1} = x_\nu, Y_{t+1} = y_j | Y_t = y_i) = T_{ij}\theta_{j,\nu}$$

Likelihood

Model

$$x \in \{x_i\}_{i=1}^{n}$$

Data

$$x_N \equiv x_1, x_2, \ldots, x_N$$

Likelihood

$$P(x_N | T, M) \approx \prod_{i=1}^{N-1} T_{x_i, x_{i+1}}$$

Go to transition counts

$$P(C | T, M) \approx \prod_{i=1}^{n} \prod_{j=1}^{n} T_{ij}^{C_{ij}}$$

Maximum likelihood

Maximum likelihood estimate

$$\hat{T} = \arg\max \left\{ P(C|T, M) \middle| \sum_j T_{ij} = 1 \forall i \right\}$$

$$\hat{T}_{ij} = \frac{C_{ij}}{C_i} \qquad C_i = \sum_j C_{ij}$$

Maximum likelihood

$$\log P(C|\hat{T}, M) = \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} \log \frac{C_{ij}}{C_i}$$

Bayesian inference

Type I inference: posterior distribution

$$P(T|C,M) = \frac{P(C|T,M)P(T|M)}{P(C|M)}$$

Type II inference: model comparison

$$P(M|C) = \frac{P(C|M)P(M)}{P(C)}$$

Evidence

$$P(C|M) = \int_T P(C,T|M) = \int_T \underbrace{P(C|T,M)}_{\prod_{i,j} T_{ij}^{C_{ij}}} \quad \underbrace{P(T|M)}_{\prod_i \frac{\Gamma(A_i)}{\prod_j \Gamma(\alpha_{ij})} \prod_j T_{ij}^{\alpha_{ij}-1}}$$

Evidence

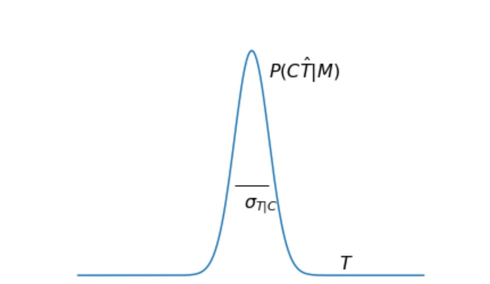
Analytical result (symmetric prior, $\alpha_{ij} = 1/n$)

$$P(C|M) \approx \sum_{j=1}^{n} C_{ij} \log \frac{C_{ij}}{C_i} - n^2 \log n$$

Laplace estimate

$$P(C|M) = \int_T P(C, T|M) \sim \underbrace{P(C, \hat{T}|M)}_{\propto P(T|C,M)} \times \sigma_{C,T}$$

"For many problems, the posterior has a strong peak at the most probable parameters." – Mackay, *Information Theory*, p. 348



Occam's razor

$$
\begin{aligned}
P(C|M) &\sim P(C, \hat{T}|M) \times \sigma_{T|C} \\
&= \underbrace{P(C|\hat{T}, M)}_{\sim \text{max likelihood}} \underbrace{P(\hat{T}|M)}_{\sim \sigma_T^{-1}} \times \sigma_{T|C}
\end{aligned}
$$

Occam factor:

$$
\log \frac{\sigma_{T|C}}{\sigma_T} \sim -n^2 \log n
$$

- $\sigma_{T|C}/\sigma_T < 1$ is the penalty to the evidence for choosing a prior that is broad relative to the volume of parameter space that captures the peak of the joint distribution (at data $= C$).
- "The factor by which M's hypothesis space collapses when the data arrive"