Accelerated Sampling of Mutants: An MSM-based Hierarchical Bayesian Strategy

Robert McGibbon and Vijay Pande

Department of Chemistry, Stanford University, Stanford, CA.

STANFORD UNIVERSITY

Introduction

- Mutational analysis is the bread and butter of experimental protein biophysics.
- In simulations, mutational analysis is a major challenge, because the costs scale proportional to the number of mutants.
- Mutational analysis is *useful* because of existence of extensive mutual information between mutants.
- This suggests that extensive simulations of a wild-type protein can be an informative prior on new simulations of a mutant.
- Our Ansatz: mutation perturbs the rates of interconversion between a unknown subset of the conformational states of a protein.

Model

- Use a common discretization of the state space between the mutant and wild type.
- Let $\vec{p_i}^M$ be the outbound transition probabilities from state i in mutant, $\vec{c_i}^{WT}$ be the observed outbound transition counts from state i in the wild type.
- Let the transfer coefficients, $q_i \in (0,1)$, be the degree of information transfer between wildtype and mutant state i.
- Informative prior on $\vec{p_i}^{MT}$:

$$\vec{p_i}^{MT} \sim \text{Dirichlet}(q_i \cdot \vec{c_i}^{WT} + 1/2)$$

- When $q_i = 0$, we have Jeffreys prior, when $q_i > 0$, counts are inherited from the wildtype into the mutant.
- q_i must also be learned from the data. For convenience, the prior on q_i is Beta, with shared hyperparameters, which constrains $q_i \in (0,1)$:

$$q_i \sim Beta(\alpha, \beta)$$

- Posterior distribution on p_i:

$$P(\vec{p_i}^{MT}|\vec{c_i}^{MT}) \propto \int_0^1 dq_i \ \text{Dir}(q_i \cdot \vec{c_i}^{WT} + \vec{c_i}^{MT} + 1/2) \cdot P_{\alpha,\beta}(q_i)$$

Methods

- Our sampling principle: choose actions that maximize the model's expected information gain (EIG).
- The actions considered are "from which state shall I start further sampling?"
- Conditional on q_i, the EIG of observing a "count", e, is given by the expected Kullbeck–Leibler divergence from the current posterior ,$P(\vec{p_i}^{MT}|q_i)$, to the updated posterior, $P(\vec{p_i}^{MT}|q_i, e)$.

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} dx \ln\left(\frac{p(x)}{q(x)}\right) p(x)$$

- For Dirichlet distributions,

$$D_{KL}(\lambda^q||\lambda^p) = \log\frac{\Gamma(\lambda^{qt})}{\Gamma(\lambda^{pt})} + \sum_{s=1}^{m} \log\frac{\Gamma(\lambda_s^p)}{\Gamma(\lambda_s^q)} + \sum_{s=1}^{m} [\lambda_s^q - \lambda_s^p]\left[\Psi(\lambda_s^q) - \Psi(\lambda^{qt})\right]$$

- If e is a single count, distributed according to the current Dirichlet-Multinomial posterior, than this simplifies to

$$E[D_{KL}|q_i] = \Psi(\lambda^t) - \log(\lambda^t) + \frac{1}{\lambda^t}\sum_l \left[\lambda_l \left(\log(\lambda_l) - \Psi(\lambda_l)\right)\right]$$

where $\lambda = q_i \cdot \vec{c_i}^{WT} + \vec{c_i}^{MT} + 1/2$ and Ψ is the digamma function.
- Still requires Markov chain Monte Carlo over q_i.

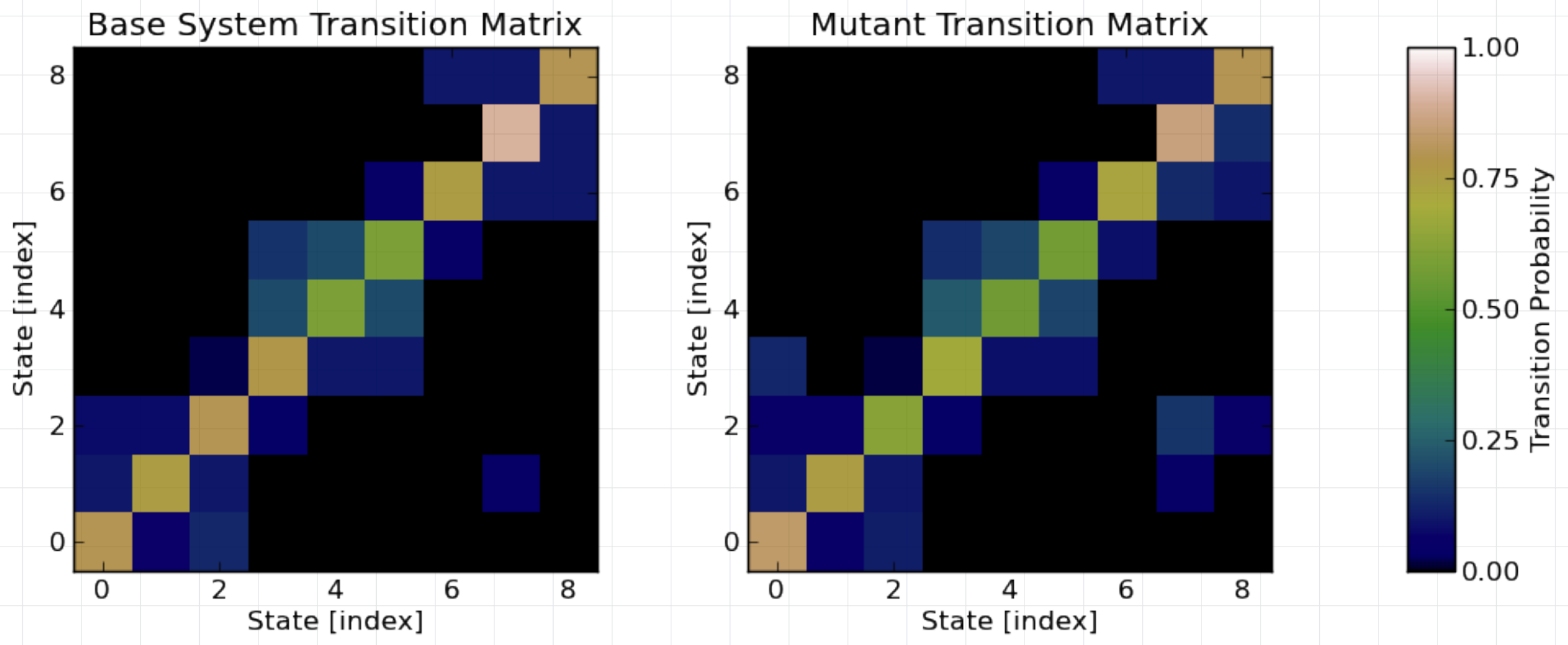

Figure: Transition probability matrix for an example system with nine states. Here, the mutant has new connections $3 \to 0, 2 \to 7, 2 \to 8$

Example

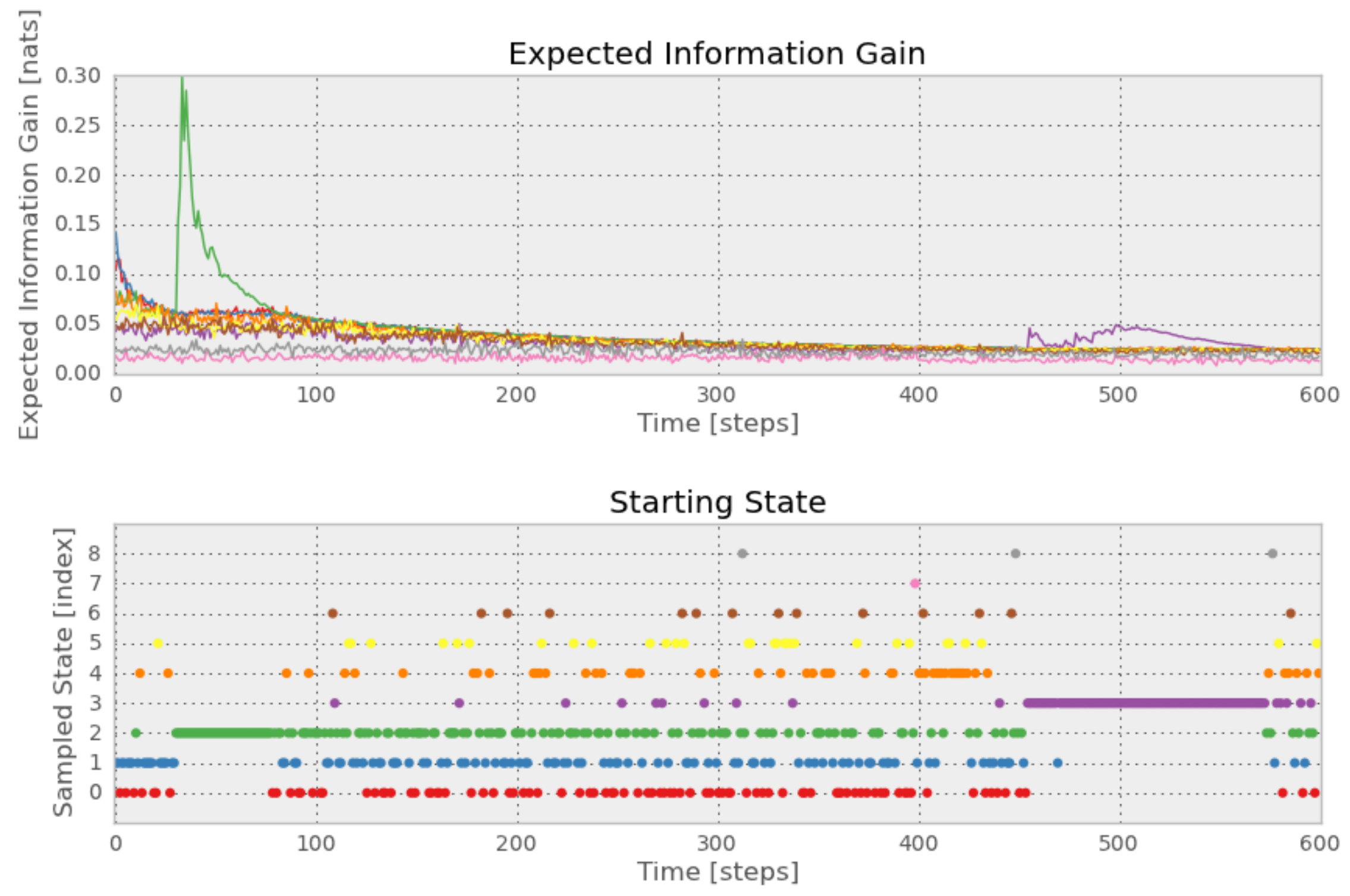

Figure: When the model discovers unexpected behavior anomalies, like the transitions from 2 and 3, it focuses its sampling there.

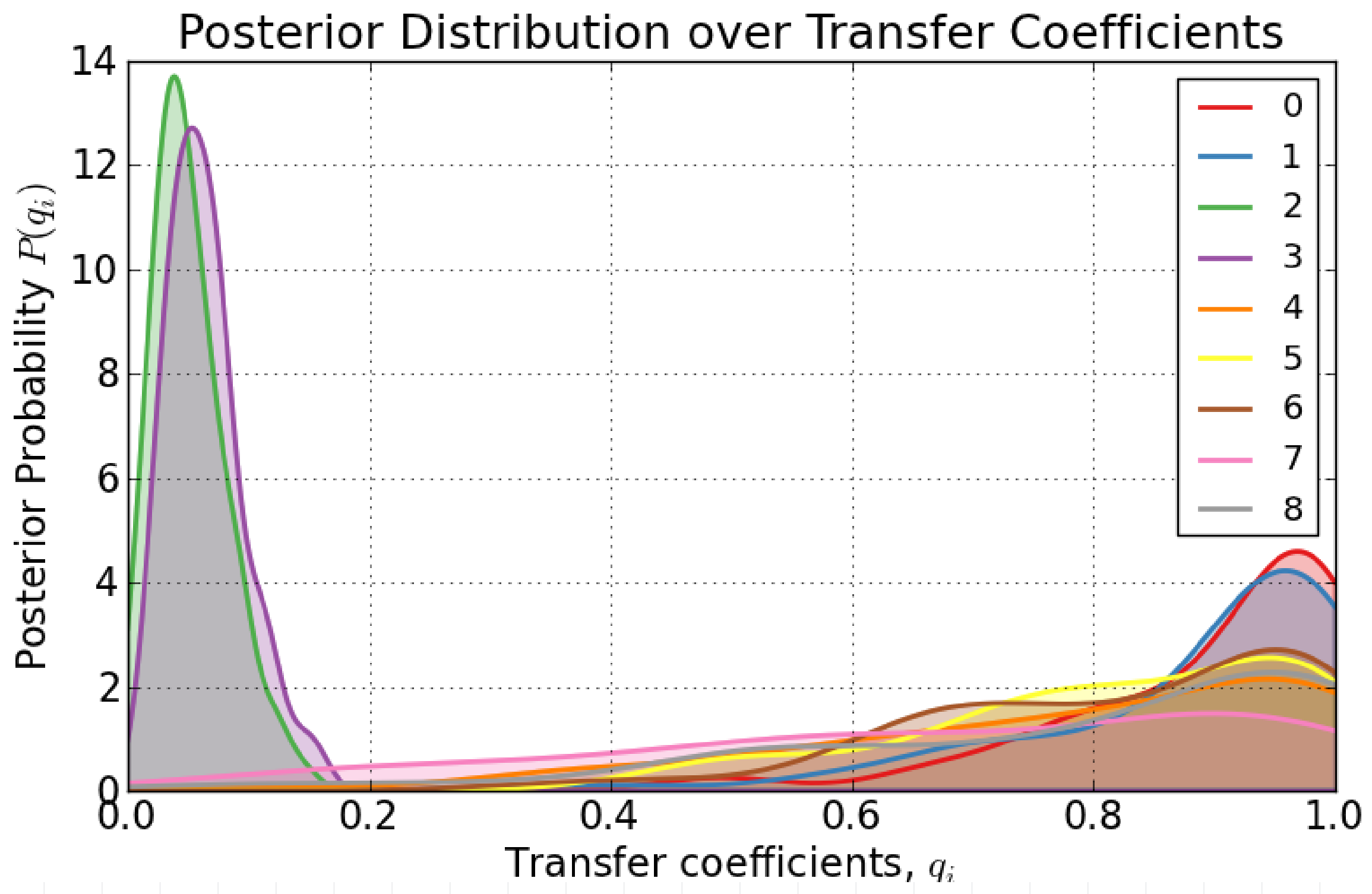

Figure: The posterior distributions over q show that the model has correctly "discovered" that states 2, 3 are dissimilar in the mutant. Some states (e.g. 8) that are undersampled still feature very wide posteriors.

Limitations

- Conformational states are assumed to be the same for wildtype and mutant.
- No model for reversibility or detailed balance constraints.
- Uncertainty due to low wildtype counts in low-population unfolded microstates can swamp model, focus sampling there.

References

- Patil, A.; Huard, D.; Fonnesbeck, C. J. "PyMC: Bayesian stochastic modelling in Python." *J. Stat. Softw.* 35 1 (2010)
- Cohn, D; Atlas, L.; Ladner, R. "Improving generalization with active learning." *Machine Learning* 15 201 (1994)