Stochastic Models Parameterized by the Infinite Hidden Markov Model Can Predict Flux and Selectivity of Solutes in Nanostructured Membranes.

Benjamin J. Coscia, Chris Calderon Michael R. Shirts

February 7, 2020

1 Introduction

We need highly selective membranes in order to perform efficient separations.

Amphiphilic molecules are capable of self-assembling into ordered nanostructures.

Lyotropic liquid crystals are a class of amphiphilic molecules that can be cross-linked into mechanically strong membranes.

We can only learn so much from experiment. MD can give us mechanistic insights with atomistic resolution so that we can intelligently design new membranes for solute-specific separations.

Unfortunately, the timescales that we can simulate with MD are insufficient to be able to make well-converged predictions of macroscopic transport properties traditionally used to characterize membranes in the lab.

- However, if we use descriptive stochastic models that can capture solute dynamics, then we could project long timescale behavior in addition to gaining a deeper understanding of solute behavior on short timescales.

In our previous work, we designed two different approaches which used solute time series in order to parameterize stochastic models that could be used to project transport on much longer timescales.

- Brief description of anomalous diffusion
- Brief description of MSDDM

Although both models had reasonable success at predicting solute MSDs on simulation timescales, they had shortcomings.

- Why MSDDM failed
- Why anomalous diffusion model could be better.

In this work, we apply the infinite hidden markov model (iHMM), a modeling approach that is agnostic to the source of time series data, in order to automatically detect and infer the parameters of an unknown number of latent autoregressive modes.

We study the same four solutes from our previous work.

We use the iHMM to generate stochastic trajectory realizations that share the same dynamical characteristics as solute trajectories observed in our MD simulations.

- The trajectories are qualitatively similar, showing expected hopping and trapping behavior
- They are quantitatively similar in that they reproduce the MSDs measured in MD

Finally, we use the stochastic trajectory realizations in order to compute the MFPT of solutes in macroscopic length pores.

- We relate the MFPT to flux and compare the fluxes between solutes to get selectivity.

2 Methods

We ran all MD simulations and energy minimizations using GROMACS 2018. We performed all post-simulation trajectory using python scripts which are available online at `https://github.com/shirtsgroup/LLC_Membranes`.

2.1 Molecular Dynamics Simulations

We studied transport of solutes in the H_{II} phase using an atomistic molecular model of four pores in a monoclinic unit cell with 10 % water by weight.

- Approximately one third of the water molecules occupy the tail region with the rest near the pore center.
- We chose to study the 10 wt % water system because solutes move significantly faster than in the 5 wt % system studied previously.
- Appropriate stochastic modeling requires that solutes sample the accessible mechanisms with representative probability.

We chose to study a subset of 4 of the fastest moving solutes from our previous work: methanol, acetic acid, urea and ethylene glycol.

- In addition to exploring membrane structural space the most, these solutes have a relatively diverse set of chemical functionality.
- For each solute we created a separate system and to each system we added 6 solutes per pore for a total of 24 solutes.
- This number of solutes per pore provides a balance of a low degree of interaction between solutes and sufficient amount of data from which to generate statistics on the time scales which we simulate.
- Further details on the setup and equilibration of these systems can be found in our previous work.[1]

We extended the 1 μs simulations of our previous work to 5 μs in order to collect ample data.

- We simulated the system with a time step of 2 fs at a pressure of 1 bar and 300 K controlled by the Parinello-Rahman barostat and the v-rescale thermostat respectively.
- We recorded frames every 0.5 ns

2.2 The Infinite State Hidden Markov Model

Hidden Markov models (HMMs) are a useful and widely used technique for modeling sequences of observations where the probability of the next observation in a sequence depends only on a previous unobserved, latent or hidden, state. [2]

- In the context of our simulations, the observations correspond to the center of mass coordinates of the solutes versus time, and the states correspond to the dynamical behavior which give rise to those types of observations.

- Unfortunately, standard HMMs require the number of hidden states to be known a priori.
- One can partially overcome this by testing a range of numbers of hidden states and determining which is the best representation of the data.

The infinite-state HMM overcomes this drawback by placing a hierarchical Dirichlet process (HDP) prior on the transition probabilities.

- Using some base probability distribution, H, a Dirichlet process (DP) generates distributions over a countably infinite number of probability measures:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \quad \theta_k \sim H, \beta \sim GEM(\gamma) \tag{1}$$

  where the θ_k are values drawn from the base distribution and the weights β_k come from a stick-breaking process parameterized by the concentration parameter γ (equivalently referred to as GEM(γ)).
- The concentration parameter expresses one's confidence in H relative to the posterior and is closely related to the number of data observations.
- Each row, G_j, of the transition matrix is produced by drawing from a DP specified using the β vector as a discrete base distribution and a separate concentration parameter, α.

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \quad \pi_j \sim DP(\alpha, \beta) \tag{2}$$

- This hierarchical specification ensures that the transition probabilities in each row share the same support points $\{\theta_1, ..., \theta_k\}$.
- Once the model has converged only a finite number of states will have significant sampling.

We describe the dynamics of each state using a vector autoregressive (VAR) model.

- A VAR(r) process is characterized by a vector of observations in a time series that are linearly dependent on r previous values of the time series vector:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^{r} A_i \mathbf{y}_{t-i} + \mathbf{e}_t \quad \mathbf{e}_t \sim N(0, \Sigma) \tag{3}$$

  Previous observations are weighted by coefficient matrices A_i. The VAR(r) process is further characterized by a shift in the mean of each dimension by the vector $\mathbf{c}$ and a white noise term $\mathbf{e}_t$. [3]
- We assumed multivariate Gaussian noise, with mean zero and covariance, Σ.
- We limited our analysis to an autoregressive order of $r = 1$.
- We used a conjugate matrix-normal inverse-Wishart prior on parameters A and Σ and a conjugate Gaussian prior on $\mathbf{c}$ in order to analytically draw from the posterior. [4]

Using the iHMM framework, we estimated the most likely number and sequence of hidden states while simultaneously estimating VAR(1) parameters for each state and the overall state transition probability matrix, T.

- We created a python implementation of this process which we heavily adapted from the MATLAB code of Fox et al. [5]
- Parameter estimation is iterative. Therefore, we looked for convergence as shown in SI.
- We refer the interested reader to much more extensive descriptions of this process and its implementation. [2, 6, 7, 4, 8]

We generated stochastic realizations by drawing state sequences based on the rows of T.

2.3 Estimating Flux and Selectivity

We calculate first passage times by propagating stochastic trajectories until they reach distance L.

We determine the mean first passage time (MFPT) using the following equation: [9]

$$P(t) = -\frac{1}{\sqrt{\pi}} e^{-(L-vt)^2/(4Dt)} \left(-\frac{D(L-vt)}{4(Dt)^{3/2}} - \frac{v}{2\sqrt{Dt}} \right) \tag{4}$$

Flux, J, is simply 1 / MFPT by the Hill relation. [10]

In our previous work, we showed that, in the absence of convective solute flux, selectivity towards solute i versus solute j can be calculated by:

$$S_{ij} = \frac{J_i/\Delta C_i}{J_j/\Delta C_j} \tag{5}$$

where ΔC_j is the trans-membrane concentration difference.

3 Results and Discussion

Trajectory realizations qualitatively match MD simulation trajectories.

- Look for hopping and trapping behavior

MSDs generated from stochastic trajectories match those from MD.

- Look at curvature and 1-σ confidence intervals

We can relate the identified states back to transport mechanisms.

- More detailed discussion of identified states
- How size of fluctuations, autoregressive parameters are influenced by trapping mechanisms
- Any new states?

We can predict macroscopic flux and selectivity.

- Flux as function of pore length
- Selectivity as function of pore length (if flux scaling is length-dependent)

4 Conclusion

We have shown that the iHMM can be used to parameterize solute time series with an unknown number of latent dynamical modes.

We can use the iHMM to help identify mechanisms by relating the latent states to observed solute behavior.

We can use the iHMM to predict macroscopic transport properties.

The iHMM is not limited to the H_{II} phase.

Supporting Information

Detailed explanations and expansions upon the results and procedures mentioned in the main text are described in the Supporting Information. This information is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgements

Molecular simulations were performed using the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). This work also utilized the RMACC Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.

References

[1] B. J. Coscia and M. R. Shirts, "Chemically Selective Transport in a Cross-Linked HII Phase Lyotropic Liquid Crystal Membrane," *J. Phys. Chem. B*, June 2019.

[2] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The Infinite Hidden Markov Model," in *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference*, vol. 1, (Cambridge, MA, US), pp. 577–585, MIT Press, Sept. 2002.

[3] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, 1994.

[4] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Nonparametric Bayesian Learning of Switching Linear Dynamical Systems," in *Advances in Neural Information Processing Systems 21* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 457–464, Curran Associates, Inc., 2009.

[5] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "The Sticky HDP-HMM: Bayesian Nonparametric Hidden Markov Models with Persistent States," *MIT Laboratory for Information and Decision Systems*, p. 60, Nov. 2007.

[6] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, Dec. 2006.

[7] J. Van Gael, Y. Saatci, Y. W. Teh, and Z. Ghahramani, "Beam Sampling for the Infinite Hidden Markov Model," in *Proceedings of the 25th international conference on Machine learning - ICML '08*, (Helsinki, Finland), pp. 1088–1095, ACM Press, 2008.

[8] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Bayesian Nonparametric Methods for Learning Markov Switching Processes," *IEEE Signal Processing Magazine*, vol. 27, pp. 43–54, Nov. 2010.

[9] E. L. Cussler, *Diffusion: Mass Transfer in Fluid Systems*. Cambridge University Press, 3 ed., 2009.

[10] T. L. Hill, *Free Energy Transduction and Biochemical Cycle Kinetics*. Springer-Verlag, 1989.

TOC Graphic