Using Stochastic Modeling to Predict Long Timescale Transport Behavior of Solutes in an H_{II} Phase Lyotropic Liquid Crystal Membrane

Benjamin J. Coscia Michael R. Shirts

January 17, 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.

- H_{II} phase lyotropic liquid crystals have densely packed, uniform sized pores and have the potential to disrupt conventional membrane separation techniques by being selective based not only on size and charge, but on chemical functionality as well.
- Q_I phase LLCs consist of a tortuous network of 3D interconnected pores. They are easier to make.

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.

In our previous work, we studied the transport of 20 small polar molecules in an H_{II} phase LLC membrane.

- In general, we observed subdiffusive transport behavior characterized by intermittent hops separated by periods of entrapment.
- We identified three mechanisms responsible for the solute trapping behavior: entanglement among monomer tails, hydrogen bonding with monomer head groups, and association with sodium counter ions.

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 descritpion 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, a modeling that is agnostic to the source of time series data, in order to detect and parameter an unknown number of autoregressive modes.

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.[**?**]

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, or hidden, state. [**?**]

- 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 their 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 process is characterized by a vector of observations in a time series that are dependent on r previous values of the time series vector, weighted by a coefficient matrix A_i in addition to a white noise term $\mathbf{e}_t$:

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

- We assumed multivariate Gaussian noise and limited our analysis to an autoregressive order of $r = 1$.

- We used a conjugate matrix-normal inverse-Wishart prior on parameters A and Σ in order to analytically draw from the posterior.

Based on the VAR parameters and matrix of transition probabilities, we calculated the most likely sequence of hidden states.

- We repeated this process iteratively until we reached convergence

- Our python implementation of this process is heavily adapted from the MATLAB code of Fox et al. [**?**]

- We refer the interested reader to much more extensive descriptions of this process and its implementation. [**?**, **?**, **?**, **?**, **?**]

3 Results and Discussion

4 Conclusion

We have tested two different mathematical frameworks for describing solute motion in an H_{II} phase LLC membrane.

- Markov state modeling with predefined states gives a nice description of transitions between observed states as well as the type of stochastic behavior shown in each state. However, it doesn't accurately portray correlated time series behavior leading to overpredicted MSDs.
- Subordinated fractional Brownian motion has a nice theoretical foundation in the anomalous diffusion literature. A two mode model that describes dynamics based on whether a solute is in or out of the pore region leads to MSDs fairly consistent with MD simulated trajectories.

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.

TOC Graphic