Addressing the Embeddability Problem in Transition Rate Estimation

Curtis Goolsby and Mahmoud Moradi*
Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701
* Correspondence: moradi@uark.edu

August 2, 2019

Abstract

Markov State Models (MSM) and related techniques have gained significant traction as a tool for analyzing and guiding molecular dynamics simulations due to their ability to extract structural, thermodynamic, and kinetic information on proteins using computationally feasible simulations. Here, we revisit the common practice in extracting the thermodynamic and kinetic information from the empirical transition matrix. We propose to build a rate/generator matrix from the empirical transition matrix to provide an alternative approach for estimating both thermodynamic and kinetic quantities. We also discuss a fundamental issue with this approach, known as the embeddability problem and the ways to address this issue. We use a one-dimensional toy model to show the workings of the proposed methods. We also show that the common method of estimating the rates based on the eigendecomposition and our proposed method based on the rate matrix are complementary in that they can potentially provide an upper and a lower limit for transition rates.

Overview

Biomolecular processes are associated with complex free energy landscapes that are virtually impossible to be fully characterized at the atomic level using current experimental and computational tools. All-atom molecular dynamics (MD) simulations, however, have recently started to explore the possibility of investigating the important regions of free energy landscapes of proteins and other biomolecules. The tools provided by computational methods are limited by their computational costs. With this limitation it is of significant interest to be able to employ statistical techniques which allow the inference of relevant thermodynamic and kinetic properties from shorter simulation times. Markov State Models, MSM [1, 2, 3, 4, 5, 6] provide some of the most powerful tools for analyzing the ensembles of short molecular dynamics trajectories to extract information on both thermodynamics and kinetics of complex biomolecular systems and predict the behavior of such systems at much longer timescales that possible to simulate with current computing capabilities. These methods, however, are based on assumptions and simplifications that introduce limitations to the reliability and interpretability of these methods. Our work is focused on analyzing the underpinnings of MSMs in order to obtain more reliable results.

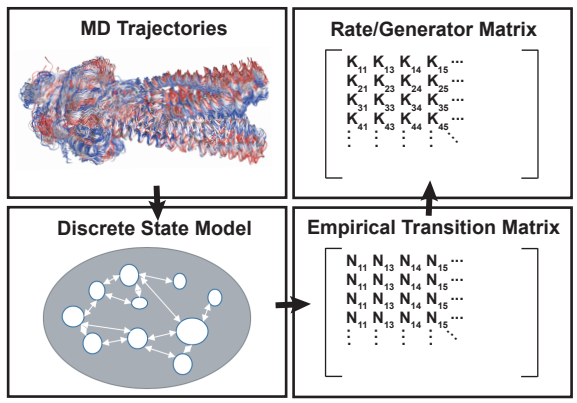


Figure 1: The schematic representation of the method proposed to estimate the rate/generator matrix associated with the discrete model of the states of a protein from the empirical transition matrix obtained from the molecular dynamics trajectoeies.

Building MSMs include (1) discretization of conformational space, (2) extracting transition statistics from simulation trajectories, and (3) analyzing the transition statistics to estimate kinetic and thermodynamic properties. Here, we only focus on the third component, which assumes an empirical transition matrix has been generated. One of the most common approaches for analyzing the empirical transition matrix is its eigendecomposition. A related method is building a rate matrix, also known as a generator matrix, to estimate the kinetic and thermodynamic quantities using theoretically known methods from chemical kinetics literature (Figure 1). Producing a time-independent rate matrix from the empirical transition matrix is known to be associated with the embeddability problem. The embeddability problem is found in taking the matrix logarithm of a time dependent transition probability matrix in order to determine the generator matrix or the rate matrix, as known in chemical kinetics literature. We will hereafter refer to it as the generator matrix. In theory, the time dependent transition probability matrix should only have positive eigenvalues however, in practice insufficient sampling of a MD simulation which has been discretized can result in the presence of negative eigenvalues. Thus, it's matrix logarithm has non-real values and is an invalid generator matrix. Solving this embeddability problem involves accurately predicting the true generator matrix from an invalid generator matrix. We aim to explore the validity of various algorithms in the literature and have developed a few of our own. We show the workings of these methods using a simple one-dimensional toy model, and compare the performance of proposed algorithms empirically.

1 Background

Even with tremendous strides in computing power and efficiency in the past 40 years [7, 8]Enhanced sampling[9] has found success in the field of computational chemistry as an effective remedy for the costs involved with simulating large molecular systems[10, 11]. Methods such as umbrella sampling[12, 13, 14, 15], metadynamics[16, 17, 18, 19, 20], simulated annealing4[21, 22, 23, 24, 25, 26], and replica exchange[27, 28, 29, 30, 31, 3 have grown increasingly popular by aiming to enhance the sampled configurational space through various techniques. In this work we explore the underpinnings of MSMs through the possibility of different algorithmic techniques for solving what is known as the 'embeddability problem'[33, 34] in financial literature, by bypassing the embeddability problem through the use of recursive neural networks, and by utilizing Riemannian manifolds to smooth non-metastable discretizations[35]. Here, we present a comparison between six algorithms, used for predicting bond rating transitions, implemented in Inamura[36] and an extension by Marada[37] of Inamura's work with known free energy estimation methods as well as a rate estimating method developed by Hummer et al[38]. In order to explore the efficacy of each of these methods we have used a bi-stable potential toy model.

Determination of kinetic properties is fundamental to the understanding of molecular systems. In order to use MD as more than a visualization tool, detailed calculations are necessary in order to search the conformational space for properties and conformations of interest. Various methods, such as those mentioned above, have been developed to improve the sampling space of MD simulations in order to determine kinetic properties of ever increasing sized systems. These models are largely based upon defining a collective variable space[39, 40] in which to reduce the dimensionality of the sampling problem. The methods then address how to sample more efficiently along a predefined collective variable space. Interesting work in these areas has also included the use of artificial intelligence methodologies[41, 42] to approximate the high dimensional statistics necessary and other stochastic processes including Markov models. Equilibrium simulations [43] have consistently improved with better forcefields, differing levels of resolution, algorithmic and hardware improvements allowing for deeper searches of the conformational space.

The generator matrix is of fundamental importance in several fields of chemistry and biology. This is due to the information about a system which can be gleaned from it. Knowing the rate at which one species converts into another gives access to equilibrium constants, an understanding of the underlying timescales, and information about how to make adjustments in order to receive desired results. In computational chemistry, having the generator matrix along a transition pathway is a holy grail of information. Calculations about relative free energies, times of transitions, and likelihoods of conformation become simplified and easy to obtain. As such methodologies which can supply the rate matrix are highly desired. Our present work focuses on creating accurate and reliable rate matrices from ever decreasing amounts of data. This is of extraordinary importance for computational work as it will allow for larger and more complex simulations

to be accessible.

Markov State Models[44, 45, 46, 47, 48] allow for the reduction of the complexities of a molecular dynamics simulation into a lower dimensional model. The conformational space can be reduced and then brought into an N×N empirical transition matrix allowing for the determination of various parameters of interest such as the relative free energies and diffusion constants among others. MSMs as with other computational methodologies still require a sufficient amount of sampling in order to obtain these parameters of interest. The present work adapts several algorithms from the field of finance in order to compare their efficacy with that of the state of the art in the chemical literature for addressing problems arising from insufficient sampling. In finance, Markov models are built in order to determine the probabilities of transition between rating grades for bonds[36]. Often there is insufficient data in the bond marketplace to observe transitions from every rating to every other rating in the same fashion that the conformational space of a simulation is too complex in order to sample every transition[36]. The discrete nature of the problem and the assumption of memoryless-ness cause the same models to work in both finance and chemistry.

MSMs inherently introduce some error into the analysis of MD simulations[49]. One focus of this work is to reduce this error from as many angles as possible. The sources of this error are three-fold. The first source of error is statistical error. The concept of enhanced sampling is directly related to this type of error. The algorithms we present are attempts to overcome this error. Statistical error results from insufficient sampling and can never be fully eliminated. The second and third sources of error are from discretization[49]. In order to use a MSM the continuous simulation data must be discretized into separate substates. This allows for the usage of Markovian formalisms. This inherently makes the model slightly non-Markovian as some memory is introduced into the system.

The second form of error is this discretization error which can be reduced by simply using a finer grained discretization. Other methods can be employed such as the use of a non-meta stable discretization[49] in order to focus in on areas of interest. An approach not explored here is to utilize a meta-stable MSM initially, and then refine the discretization towards a non-metastable MSM which focuses on areas identified as of interest.

Finally, the third source of error is known as spectral error. This results from the manner in which MSM formalism tend to ignore fast transitions in favor of perceiving larger scale events. For example, a MSM would observe a conformation change, but would ignore a hydrogen bond vibration. This spectral error tends to zero as the lag time increases. As such, it is possible to run MSM analysis on data and calculate the reduction in spectral error with increasing lag time allowing for intelligent decision making as to which lag time to finally utilize[49].

The embeddability Problem

This work features implementations of six separate algorithms in the determination of kinetic parameters of interest. These algorithms are applied to a biharmonic toy model. We test the efficacy of the various algorithms under different conditions, first in sufficient sampling to show that the algorithms do not distort the correct answer, and then in the case of increasing lag times as well as the case of insufficient data points. By doing this, we are able to show the various strengths and weaknesses of each algorithm and hope to provide guidance for future researchers in their decision making about this matter.

The empirical transition matrix, N, is defined as a square matrix which contains the probability of transition from state i to state j based upon the observed transitions. As such it has the requisite information for enabling the calculation of relative free energies. This can be accomplished through two classical methods. The first is heretofore referred to as the Naive estimate, which is simply

$$F_i = -k_B T \log(N_{i,i}) \tag{1}$$

The other is a detailed balance (DB) based method, which allows for the calculation of relative free energies by comparing the observed transitions from one state to another. The DB based free energy is found by calculating:

$$\Delta F_{i,j} = -k_b T \ \log(\frac{p_{i,j}}{p_{j,i}}) \tag{2}$$

where $p_{i,j}$ is defined as the time-dependent transition matrix, P, which is simply the normalization of the

empirical transition matrix.

$$p_{i,j} = \frac{N_{i,j}}{N_{i,i}} \tag{3}$$

Detailed balance[50] is a well understood propery for chemical systems in equilibrium. Using this property, we can calculate the relative transitions between two states, and use detailed balance to find their relative free energies. From the normalized transition matrix we can find the generator matrix Q. This matrix is the basis of this papers work. By implementing various algorithms designed to work with generator matrices, we have shown the limitations of these algorithms, and the need of adequate sampling. The generator/rate matrix Q is related to the time-dependent transition matrix P via:

$$Q = \ln(p)/t, \tag{4}$$

where t is the time between observations resulting in the particular empirical transition matrix used to build P. Q has the following properties: • $\sum_{j=1}^{K} q_{i,j} = 0 \ for \ 1 \leq i \leq K$
• $0 \leq q_i \leq \infty for 1 \leq i \leq K$
• $q_{i,j} \geq 0 \ for \ 1 \leq i \leq K \ with \ i \neq j$

where $q_i = -q_{i,i}$. The restrictions, due to the assumption of detailed balance, are namely, that the diagonal elements are equal to the negative sum of the rest of the row, and the off diagonal elements are nonnegative.

The Embeddability Problem

Given an empirical transition matrix, there is a sufficient condition [51] to show that an exact generator matrix does exist. Namely, if there are states i and j such that transitions between the two are possible given infinite time but there is no recorded transition from i to j in the empirical transition matrix, then one can conclude that an exact generator does not exist. Unfortunately, this is more than not the case for empirical transition matrices. Due to undersampling, it is possible to have an Empirical Transition Matrix which has negative eigenvalues. According to equation 4, the matrix logrithm of the Empirical Transition Matrix contains non-real elements, thus creating an invalid generator matrix.

We show several methods in the following section for addressing this problem, namely either adjusting directly the generator matrix produced by the matrix logarithm of our insufficiently sampled transition matrix, performing a maximum likelihood estimation, or using a Markov Chain Monte Carlo technique.

Algorithms

These algorithms range from Bayesian methods such as Gibbs Sampling to very simple methods such as the Diagonal Adjustment Algorithm presented first.

Diagonal Adjustment

The Diagonal Adjustment algorithm[36], DA, is a simple and sometimes effective solution to the embeddability problem. For an $M \times M$ generator matrix, the algorithm consists of two steps.

$$1. \ \tilde{q}_{i,j} = \begin{cases} 0 \ if \ (i \neq j) \ and \ q_{i,j} < 0 \\ q_{i,j} \ otherwise \end{cases} \tag{5}$$

$$2. \ \tilde{q}_{i,j}^{DA} = -\sum_{j=1, j\neq i}^{M} \tilde{q}_{i,j} \tag{6}$$

Weighted Adjustment

The Weighted Adjustment[36], WA, is another simple algorithm very similar to the DA. For an $M \times M$ generator matrix, the algorithm also consists of two steps.

$$1. \ \tilde{q}_{i,j} = \begin{cases} 0 \ if \ (i \neq j) \ and \ q_{i,j} < 0 \\ q_{i,j} \ otherwise \end{cases} \tag{7}$$

$$2.\ \ \tilde{q}_{i,j}^{WA} = q_{i,j} - |\tilde{q}_{i,j}| \frac{\sum_{j=1}^{M} \tilde{q}_{i,j}}{\sum_{j=1}^{M} \tilde{|q_{i,j}|}} \tag{8}$$

Component-Wise Optimization

Component-Wise Optimization[37], CWO, is a somewhat more complex version of DA or WA. In fact it performs quite well for its relative simplicity when compared with stochastic algorithms. The general idea is to divide the problem into M-1 seperate optimization problems. First, a DA or WA is performed followed by optimizing each row according to:

$$\tilde{q}_{i,j} = \underset{q_{i,j} \in [0,c]}{\arg\min} ||exp(\Delta t \mathbf{Q}(q_{i,j})) - \mathbf{P}|| \tag{9}$$

Where $\mathbf{P}$ is the given or observed normalized empirical transition matrix and c is some constant chosen from expectations. The smaller the c, the faster the convergence. Marada[37] goes on to note that the CWO is not capable of distinguishing between local and global minima, and as such should be used judiciously, possibly as a way to further optimize results from other algorithms.

Expectation Maximization

To see a full derivation of EM[36] see McLachan & Krishnan[52]. The procedure of the EM algorithm is completed by calculating a generator matrix for each element, and then using (4) to construct a new generator matrix. Then iteration until convergence completes the algorithm.

Hamiltonian-Sampling

Our method is a hybrid of Gibbs and Metropolis-Hasting Monte Carlo [53] sampling methods. This method works by drawing a new transition matrix from the previous iteration's generator matrix according to a gamma distribution. The samples are accepted or rejected using the relative probabilities of the states from the initial empirical transition matrix with random acceptance occurring as in the Metropolis-Hastings algorithm. These generated samples are then used to create another generator matrix, which is accepted with probability one, as in Gibbs Sampling, and another iteration is performed. This process allows for the fleshing out of the generator matrix and thus solves the embeddability problem.

Maximum Likelihood Estimator

One can also create a maximum likelihood estimator for the generator matrix[38]. This solution utilizes the relationship between the likelihood of the observations and the generator matrix utilizing the detailed balance condition as shown.

$$L = \prod_{\alpha} p(i_\alpha, t_\alpha | j_\alpha, 0) = \prod_{\alpha} (e^{t_\alpha \mathbf{Q}})_{i_\alpha, t_\alpha} \tag{10}$$

$$1.\ \ \tilde{q}_{i,j} = \begin{cases} q_{ij}\ if\ (i > j), \\ -\sum_{l \neq i}\ q_{ij}\ if\ i = j, \\ q_{ji}\frac{P_i}{P_j}\ if\ i < j \end{cases} \tag{11}$$

Assuming a uniform prior, we can use equation 11 to reduce the number of free parameters. This allows for adjustment of individual values of the generator matrix, until a maximum likelihood is obtained.

Mean First Passage Time Calculations

The Mean First Passage Time, MFPT, i.e. the average time to travel from state A to state B, is theoretically obtainable from the generator matrix. Given a toy model with a known potential function and diffusion constant we can calculate the exact value using theory from Lifson and Jackson[54] where x_n is state A,

x_m is state B, a is a point at which the potential is $\pm$inf but in practice is just arbitrarily large, $D(y)$ is a function of the diffusion coefficients at y, and β is the usual thermodynamic quantity $k_B T$.

$$MFPT = \int_{x_n}^{x_m} dy e^{\beta V(y)} \frac{1}{D(y)} \int_a^y dz e^{-\beta V(z)} \tag{12}$$

We can also empirically determine the MFPT by simply observing the transition in a molecular dynamics trajectory. The interesting application of this work is that given a generator matrix it is possible to predict the MFPT using one of two methods. The first is known as the implied timescale method in the mathematical literature. Given a generator matrix, one can solve for the MFPT, with λ_i as the ith eigenvalue and τ as the lag time.

$$MFPT = \frac{-\tau}{ln\ \lambda_2} \tag{13}$$

Additionally, utilizing Hummer's work [38] we can predict the MFPT from a valid generator by discretizing Lifson and Jackson's formula in order to produce a summation.

$$MFPT = \sum_{x_n}^{x_m} \Delta y e^{\beta V(y)} \frac{1}{D(y)} \sum_a^y \Delta z e^{-\beta V(z)} \tag{14}$$

Toy Model

We have implemented the algorithms discussed above to estimate both the free energy profile and the MFPT for a simple one-dimensional toy model, described by potential $f(x) = \frac{k}{4}(x^2 - 1)^2$. We use an overdamped Langevin equation to generate many long trajectories starting at different positions around the minima and between the minima. We then systematically use shorter and shorter trajectories with different lag times to determine the behavior of different algorithms with the trajectory length, the number of trajectories, and the lag time.

This has allowed for an empirical investigation of the errors involved in the discretization of a continuous space, as well as statistical errors which derive from sparsely sampled state spaces. This error shows up in three categories, discretization error, i.e. the error associated with taking a Markovian process to one which has a slight amount of built in memory, spectral error, i.e. the error associated with an inexact decorrelation of data, and statistical error, i.e. the error associated with insufficient sampling.

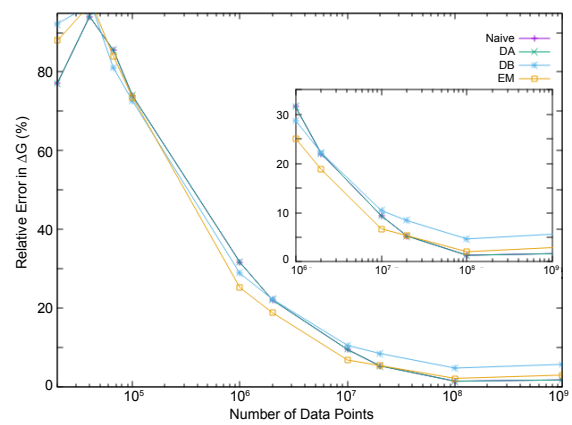
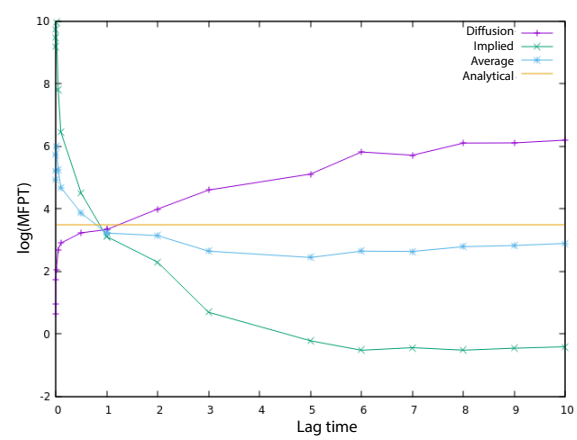


Figure 2: Data showing the performance of a DB calculation versus two other algorithms at predicting the free energy of a model bistable potential functions. $f(x) = \frac{k}{4}(x^2 - 1)^2$. Notice how the EM function is consistently on the bottom, showing the highest accuracy.

The embeddability problem has to do with insufficient sampling. Several algorithms have been implemented to take an invalid rate matrix and produce a valid rate matrix. To this end it can be seen in 2 that the Bayesian approach, titled Expectation Maximization, surpasses the other algorithms in our model bistable potential for predicting the free energy of the transition state.

We have also explored two different methods of calculating the Mean First Passage Time. It can been seen in the Figure 3 that the method built from Hummer's[38] work appears to produce an upper bound for the MFPT, while the Implied timescale method, based upon standard Markov Model eigendecomposition method seems to produce a lower bound. The analytically determined MFPT ends up in the middle of both of these predictions. If the mean of both the Hummer prediction and the implied timescale is taken, a very close approximation is obtained. This shows the two methods are complementary and can provide upper and lower bounds for the estimatation of MFPT.



References

[1] Andrieu, C., A. Doucet, and R. Holenstein. 2010. Particle Markov Chain Monte Carlo Methods. *JOURNAL Of The Royal Statistical Society Series B-Statistical Methodology*. 72:269–342. ISSN 1369-7412. doi:10.1111/j.1467-9868.2009.00736.x.

[2] Ephraim, Y. and N. Merhav. 2002. Hidden Markov Processes. *IEEE Transactions On Information Theory*. 48:1518–1569. ISSN 0018-9448. doi:10.1109/TIT.2002.1003838.

[3] Pande, V. S., K. Beauchamp, and G. R. Bowman. 2010. Everything You Wanted To Know About Markov State Models But Were Afraid To Ask. *METHODS*. 52:99–105. ISSN 1046-2023. doi:10.1016/j.ymeth.2010.06.002.

[4] Bowman, G. R., V. A. Voelz, and V. S. Pande. 2011. Taming The Complexity Of Protein Folding. *CURRENT Opinion In Structural Biology*. 21:4–11. ISSN 0959-440X. doi:10.1016/j.sbi.2010.10.006.

[5] Sisson, S. 2005. Transdimensional Markov Chains: A Decade Of Progress And Future Perspectives. *JOURNAL Of The American Statistical Association*. 100:1077–1089. ISSN 0162-1459. doi:10.1198/016214505000000664.

[6] Shukla, D., C. X. Hernandez, J. K. Weber, and V. S. Pande. 2015. Markov State Models Provide Insights Into Dynamic Modulation Of Protein Function. *ACCOUNTS Of Chemical Research*. 48:414–422. ISSN 0001-4842. doi:10.1021/ar5002999.

[7] Schulz, R., B. Lindner, L. Petridis, and J. C. Smith. 2009. Scaling of multimillion-atom biological molecular dynamics simulation on a petascale supercomputer. *Journal Of Chemical Theory And Computation*. 5:2798–2808. doi:10.1021/ct900292r. PMID: 26631792.

[8] Durrant, J. D. and J. A. Mccammon. 2011. Molecular Dynamics Simulations And Drug Discovery. *BMC Biology*. 9. ISSN 1741-7007. doi:10.1186/1741-7007-9-71.

[9] Bernardi, R. C., M. C. R. Melo, and K. Schulten. 2015. Enhanced Sampling Techniques In Molecular Dynamics Simulations Of Biological Systems. *Biochimica Et Biophysica Acta-General Subjects*. 1850:872–877. ISSN 0304-4165. doi:10.1016/J.Bbagen.2014.10.019.

[10] Karplus, M. and J. Mccammon. 2002. Molecular Dynamics Simulations Of Biomolecules. *NATURE Structural Biology*. 9:646–652. ISSN 1072-8368. doi:10.1038/nsb0902-646.

[11] Orozco, M. 2014. A Theoretical View Of Protein Dynamics. *CHEMICAL Society Reviews*. 43:5051–5066. ISSN 0306-0012. doi:10.1039/c3cs60474h.

[12] Kumar, S., D. Bouzida, R. Swendsen, P. Kollman, and J. Rosenberg. 1992. The Weighted Histogram Analysis Method For Free-Energy Calculations On Biomolecules .1. The Method. *Journal Of Computational Chemistry*. 13:1011–1021. ISSN 0192-8651. doi:10.1002/Jcc.540130812.

[13] Torrie, G. and J. Valleau. 1977. Non-Physical Sampling Distributions In Monte-Carlo Free-Energy Estimation - Umbrella Sampling. *Journal Of Computational Physics*. 23:187–199. ISSN 0021-9991. doi:10.1016/0021-9991(77)90121-8.

[14] Roux, B. 1995. The Calculation Of The Potential Of Mean Force Using Computer-Simulations. *Computer Physics Communications*. 91:275–282. ISSN 0010-4655. doi:10.1016/0010-4655(95)00053-I.

[15] Kumar, S., J. Rosenberg, D. Bouzida, R. Swendsen, and P. Kollman. 1995. Multidimensional Free-Energy Calculations Using The Weighted Histogram Analysis Method. *Journal Of Computational Chemistry*. 16:1339–1350. ISSN 0192-8651. doi:10.1002/Jcc.540161104.

[16] Laio, A. and M. Parrinello. 2002. Escaping free-energy minima. *Proceedings Of The National Academy Of Sciences*. 99:12562–12566. ISSN 0027-8424. doi:10.1073/pnas.202427399.

[17] Barducci, A., G. Bussi, and M. Parrinello. 2008. Well-Tempered Metadynamics: A Smoothly Converging And Tunable Free-Energy Method. *Physical Review Letters*. 100. ISSN 0031-9007. doi: 10.1103/Physrevlett.100.020603.

[18] Laio, A. and F. L. Gervasio. 2008. Metadynamics: A Method To Simulate Rare Events And Reconstruct The Free Energy In Biophysics, Chemistry And Material Science. *Reports On Progress In Physics*. 71. ISSN 0034-4885. doi:10.1088/0034-4885/71/12/126601.

[19] Barducci, A., M. Bonomi, and M. Parrinello. 2011. Metadynamics. *Wiley Interdisciplinary Reviews-Computational Molecular Science*. 1:826–843. ISSN 1759-0876. doi:10.1002/Wcms.31.

[20] Laio, A., A. Rodriguez-Fortea, F. Gervasio, M. Ceccarelli, and M. Parrinello. 2005. Assessing The Accuracy Of Metadynamics. *Journal Of Physical Chemistry B*. 109:6714–6721. ISSN 1520-6106. doi: 10.1021/Jp045424K.

[21] Zang, H., S. Zhang, and K. Hapeshi. 2010. A review of nature-inspired algorithms. *Journal Of Bionic Engineering*. 7:S232 – S237. ISSN 1672-6529. doi:https://doi.org/10.1016/S1672-6529(09)60240-7.

[22] Tsallis, C. and D. A. Stariolo. 1996. Generalized simulated annealing. *Physica A: Statistical Mechanics And Its Applications*. 233:395 – 406. ISSN 0378-4371. doi:https://doi.org/10.1016/S0378-4371(96)00271-3.

[23] KIRKPATRICK, S., C. Gelatt, and M. Vecchi. 1983. OPTIMIZATION By Simulated Annealing. *SCIENCE*. 220:671–680. ISSN 0036-8075. doi:10.1126/science.220.4598.671.

[24] Szu, H. and R. Hartley. 1987. Fast simulated annealing. *Physics Letters A*. 122:157 – 162. ISSN 0375-9601. doi:https://doi.org/10.1016/0375-9601(87)90796-1.

[25] Marinari, E. and G. Parisi. 1992. Simulated Tempering - A New Monte-Carlo Scheme. *Europhysics Letters*. 19:451–458. ISSN 0295-5075. doi:10.1209/0295-5075/19/6/002.

[26] Goffe, W., G. Ferrier, and J. Rogers. 1994. Global Optimization Of Statistical Functions With Simulated Annealing. *Journal Of Econometrics*. 60:65–99. ISSN 0304-4076. doi:10.1016/0304-4076(94)90038-8.

[27] Sugita, Y. and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*. 314:141 – 151. ISSN 0009-2614. doi:https://doi.org/10.1016/S0009-2614(99)01123-9.

[28] Sugita, Y. and Y. Okamoto. 1999. Replica-Exchange Molecular Dynamics Method For Protein Folding. *Chemical Physics Letters*. 314:141–151. ISSN 0009-2614. doi:10.1016/S0009-2614(99)01123-9.

[29] Hukushima, K. and K. Nemoto. 1996. Exchange Monte Carlo Method And Application To Spin Glass Simulations. *Journal Of The Physical Society Of Japan*. 65:1604–1608. ISSN 0031-9015. doi: 10.1143/Jpsj.65.1604.

[30] Earl, D. and M. Deem. 2005. Parallel Tempering: Theory, Applications, And New Perspectives. *Physical Chemistry Chemical Physics*. 7:3910–3916. ISSN 1463-9076. doi:10.1039/B509983H.

[31] Sugita, Y., A. Kitao, and Y. Okamoto. 2000. Multidimensional Replica-Exchange Method For Free-Energy Calculations. *Journal Of Chemical Physics*. 113:6042–6051. ISSN 0021-9606. doi: 10.1063/1.1308516.

[32] Fukunishi, H., O. Watanabe, and S. Takada. 2002. On The Hamiltonian Replica Exchange Method For Efficient Sampling Of Biomolecular Systems: Application To Protein Structure Prediction. *Journal Of Chemical Physics*. 116:9058–9067. ISSN 0021-9606. doi:10.1063/1.1472510.

[33] Guerry, M.-A. 2014. Some Results On The Embeddable Problem For Discrete-Time Markov Models In Manpower Planning. *Communications In Statistics-Theory And Methods*. 43:1575–1584. ISSN 0361-0926. doi:10.1080/03610926.2012.742543.

[34] Guerry, M.-A. 2013. On The Embedding Problem For Discrete-Time Markov Chains. *Journal Of Applied Probability*. 50:918–930. ISSN 0021-9002.

[35] Koltai, P., G. Ciccotti, and C. Schuette. 2016. On Metastability And Markov State Models For Non-Stationary Molecular Dynamics. *JOURNAL Of Chemical Physics*. 145. ISSN 0021-9606. doi: 10.1063/1.4966157.

[36] Inamura, Y. and Others. 2006. Estimating continuous time transition matrices from discretely observed data. *Bank Of Japan*:06–07.

[37] Marada, T. Quantitative credit risk modeling: Ratings under stochastic time.

[38] Hummer, G. 2005. Position-Dependent Diffusion Coefficients And Free Energies From Bayesian Analysis Of Equilibrium And Replica Molecular Dynamics Simulations. *New Journal Of Physics*. 7. ISSN 1367-2630. doi:10.1088/1367-2630/7/1/034.

[39] Kurylyak, I. and I. Yukhnovskii. 1982. The Method Of Collective Variables In The Equilibrium Statistical-Theory Of Bounded Systems Of Charged-Particles .1. Continuum Model Of An Electrolyte Solution Occupying A Half-Space. *Theoretical And Mathematical Physics*. 52:691–699. ISSN 0040-5779. doi:10.1007/Bf01027790.

[40] Gimondi, I., G. A. Tribello, and M. Salvalaglio. 2018. Building Maps In Collective Variable Space. *Journal Of Chemical Physics*. 149. ISSN 0021-9606. doi:10.1063/1.5027528.

[41] Yedidia, J., W. Freeman, and Y. Weiss. 2005. Constructing Free-Energy Approximations And Generalized Belief Propagation Algorithms. *Ieee Transactions On Information Theory*. 51:2282–2312. ISSN 0018-9448. doi:10.1109/Tit.2005.850085.

[42] Pirdashti, M., S. Curteanu, M. H. Kamangar, M. H. Hassim, and M. A. Khatami. 2013. Artificial Neural Networks: Applications In Chemical Engineering. *Reviews In Chemical Engineering*. 29:205–239. ISSN 0167-8299. doi:10.1515/Revce-2013-0013.

[43] Gillespie, D. T. 2007. Stochastic Simulation Of Chemical Kinetics. *Annual Review Of Physical Chemistry*. 58:35–55. ISSN 0066-426X. doi:10.1146/Annurev.Physchem.58.032806.104637.

[44] Chodera, J. D., N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope. 2007. Automatic Discovery Of Metastable States For The Construction Of Markov Models Of Macromolecular Conformational Dynamics. *Journal Of Chemical Physics*. 126. ISSN 0021-9606. doi:10.1063/1.2714538.

[45] Prinz, J.-H., H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schuette, and F. Noe. 2011. Markov Models Of Molecular Kinetics: Generation And Validation. *Journal Of Chemical Physics*. 134. ISSN 0021-9606. doi:10.1063/1.3565032.

[46] Bowman, G. R., K. A. Beauchamp, G. Boxer, and V. S. Pande. 2009. Progress And Challenges In The Automated Construction Of Markov State Models For Full Protein Systems. *Journal Of Chemical Physics*. 131. ISSN 0021-9606. doi:10.1063/1.3216567.

[47] Beauchamp, K. A., G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande. 2011. Msmbuilder2: Modeling Conformational Dynamics On The Picosecond To Millisecond Scale. *Journal Of Chemical Theory And Computation*. 7:3412–3419. ISSN 1549-9618. doi:10.1021/Ct200463M.

[48] Chodera, J. D. and F. Noe. 2014. Markov State Models Of Biomolecular Conformational Dynamics. *CURRENT Opinion In Structural Biology*. 25:135–144. ISSN 0959-440X. doi:10.1016/j.sbi.2014.04.002.

[49] Prinz, J.-H., H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. SchÜTte, and F. NoÉ. 2011. Markov models of molecular kinetics: Generation and validation. *The Journal Of Chemical Physics*. 134:174105.

[50] Klein, M. J. 1955. Principle of detailed balance. *Physical Review*. 97:1446.

[51] Israel, R., J. Rosenthal, and J. Wei. 2001. Finding Generators For Markov Chains Via Empirical Transition Matrices, With Applications To Credit Ratings. *Mathematical Finance*. 11:245–265. ISSN 0960-1627. doi:10.1111/1467-9965.00114.

[52] Ng, S. K., T. Krishnan, and G. J. Mclachlan. 2012. The em algorithm. In *Handbook Of Computational Statistics*. Springer, pages 139–172.

[53] Metropolis, N. and S. Ulam. 1949. The monte carlo method. *Journal Of The American Statistical Association*. 44:335–341. doi:10.1080/01621459.1949.10483310. PMID: 18139350.

[54] Lifson, S. and J. L. Jackson. 1962. On the self-diffusion of ions in a polyelectrolyte solution. *The Journal Of Chemical Physics*. 36:2410–2414.