University College Dublin

Laboratory Internship Report

Modelling and Inferring Molecular Conformational Changes of Proteins

Author:
Olivier Roux

Internship Supervisor:
Dr. Florian Maire
Academic Supervisor:
Dr. Thibaut Le Gouic

15/02/2016 - 29/07/2016

Acknowledgements

To all the people I had the luck to meet and mix with, over the course of these few months in Ireland, I thank you for making this experience really rewarding.

For his welcome, his guidance, his trust, and many other things, I would like to thank most warmly Florian Maire. Thanks to him, this research internship (which was the first I did in Statistics) has been a great scientifical and personal experience. The best of luck to you Florian!

Dr. Vio Buchete has my deep gratitude for introducing me to the field of protein dynamics and, even with a very demanding schedule, took time to help me and reflect on my work.

I would like to thank Nial Friel and all the group of the Insight Centre For Data Analytics for their warm welcome and good mood.

Finally, I am grateful to my academic supervisor, Dr. Thibaut Le Gouic, for his interest in my work.

Summary

This research internship allowed me to take an interest and learn more about Markov processes, and their application to molecular biophysics. I first familiarized myself with the software Matlab with which I did most of my simulations and algorithms, learning how to simulate simple Monte Carlo Markov Chains and infer them.

As Dr. Florian Maire was curious about the fact that dynamics of small peptides, which are accurately captured by Markov-based models called Coarse Master Equations (CME), can be described by really simple statistics, I had the opportunity to meet Dr. Vio Buchete and take interest in Coarse Graining theory. Slowly at first but surely I was able to understand the framework of master equation models and their interest in understanding and simplifying protein dynamics.

I worked on a method to simulate transition rate matrices, a parameter that govern the dynamics of those processes under Markovian assumption, and studied their spectral properties.

I then tackled the topic of estimating the probability that a simpler representation of the protein dynamics can be considered. Results show that such an alternative representation depends on the number of states and on the degree of connectivity between them. I tried to provide guidelines on how to estimate the error that one is committing when working with the simpler model.

In the light of these analyses I attempted to draw an analogy with the discrete-time general state space model, hoping to find similar results but much remains to be done!

Contents

1 Introduction

1.1 Protein Dynamics

A most important feature of protein dynamics such as protein folding is that proteins chemical properties vary according to their conformation. Modeling the protein dynamics is thus necessary in order to group the states of the protein, which are physically characterized by many parameters such as their energy, position, relative angles, that correspond to stable conformations.

Inference based on experimental data usually yields to a large amount of states, paving the way to complicated models that practitioner may struggle to interpret.

Markovian models such as the Coarse Master Equations, by describing the system at a coarse level, enable analytical treatments while accurately incorporating detailed molecular information.

1.2 The Insight Center for Data Analytics - University College Dublin

Located in Dublin, the Insight Center for Data Analytics is a preponderant data analytics research organisation. Aiming to find solutions for the area of connected health and the discovery economy, it's main areas of priority research are:

- Machine Learning & Statistics
- Semantic Web
- Linked Data
- Media Analytics
- Optimisation & Decision Analytics
- Personal Sensing
- Recommander Systems

My internship took place in the Statistics department - located in the School of Mathematics and Statistics at University College Dublin - led by Prof. Nial Friel and under the Supervision of Dr. Florian Maire. Dr. Florian Maire is a postdoctoral research fellow who takes interest in Markov chain Monte Carlo methods and Expectation Maximization algorithms with application to Big data problems, Image processing and Network models.

2 Coarse Master Equations

2.1 The Statistical Model

Dynamics of a molecular system can be seen as the motion of a polypeptide chain on a complex energy landscape (see [Prinz et al., 2011]). We assume that the molecular system studied lives in a continuous state space Ω consisting of positions, momenta and energy of all the atoms considered. Its time-evolution trajectory $\{X(t),\, t > 0\}$ is defined on the probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ and observes the following properties:

(1) $\{X(t),\, t > 0\}$ is a (memoryless) stochastic process on Ω which satisfies the Markov property that is, for all $(x, y, z) \in \Omega^3$ and $0 < t_0 < t_1 < t_2$:

$$\mathbb{P}(X(t_2)) \in \mathrm{d}x \,|\, X(t_1) = y, X(t_0) = z) = \mathbb{P}(X(t_2) \in \mathrm{d}x \,|\, X(t_1) = y)$$

(2) $\{X(t),\, t > 0\}$ is assumed to be time-homogeneous, that is, for all $(x, y) \in \Omega^2$, $h > 0$ and $t > 0$:

$$\mathbb{P}(X(t + h)) \in \mathrm{d}x \,|\, X(h) = y) = \mathbb{P}(X(t) \in \mathrm{d}x \,|\, X(0) = y)$$

(3) $\{X(t),\, t > 0\}$ is irreducible: any subset of Ω can be reached by an infinitely long trajectory and are visited with a frequency given by the Boltzmann distribution that admits as density

$$\pi(x) \propto \exp\{-\beta H(x)\}\,, \quad x \in \Omega\,,$$

where H is the Hamiltonian operator and β some positive constant.

(4) $\{X(t),\, t > 0\}$ is reversible: for all $(x, y) \in \Omega^2$, the probability density of going from state $X(t) = x$ to state $X(t + \tau) \in \mathrm{d}y$ in time τ, $\mathbb{P}_\tau(x_t, \mathrm{d}y)$, fulfills the detailed balance relation:

$$\pi(\mathrm{d}x)\mathbb{P}(X(t) + \tau) \in \mathrm{d}y \,|\, X(t) = x) = \pi(\mathrm{d}y)\mathbb{P}(X(t) + \tau) \in \mathrm{d}x \,|\, X(t) = y)\,.$$

In most macromolecular systems only the region which contains conformations of relatively low energies is populated. Thus, a common way to characterize this region is to subdivide it into coarse sets, each of which comprising a group of similar molecular structures. Ω is thus subdivided into $N \in \mathbb{N}$ cells, whereby defining discrete microstates. We should note that by subdividing Ω we don't know where exactly $\{X(t),\, t > 0\}$ is and thus lose accuracy in the reproduction of long-time kinetics.

$\{X(t),\, t > 0\}$ (considered to be regular and stable) now takes value in the discrete state $S = \{1, 2, \ldots\}$. We want to express the dynamics in terms of the transition probabilities between these states such that the continuous process is well approximated:

We postulate that there exists a set of functions

$$k_i \quad : \quad S\backslash\{i\} \quad \rightarrow \quad \mathbb{R}^+$$

for $i \in S$ such that for all $j \neq i$ and small $h > 0$

$$\mathbb{P}(X(t+h) = j \,|\, X(t) = i) = k_i(j)h + o(h) \tag{1}$$

and we define the $P(t) \in \mathcal{M}_n(\mathbb{R})$ as the matrix whose elements are:

$$\forall (i,j) \in S^2 \qquad P_{i,j}(t) = \mathbb{P}(X(t) = j \,|\, X(0) = i)$$

We note that:

(1) $\sum_{j \in S} P_{i,j}(t) = 1$ for all $i \in S$ and $t \geq 0$

(2) $P(t)$ satisfies the semi-group property (with the matrix product):

$$\forall h > 0 \qquad P(t+h) = P(t)P(h)$$

as a straightforward application of the Law of Total probability and the Markov property.

(3) $\lim\limits_{t \to 0} P(t) = \mathbb{1}_n$

(4) $P(t)$ is an operator that acts:

(a) on $\mathbb{M}(S)$ the set of probability measure on S:

$$P(t) : \left| \begin{array}{rcl} \mathbb{M}(S) & \longrightarrow & \mathbb{M}(S) \\ \mu & \longmapsto & \mu P(t) \end{array} \right.$$

(b) on $\mathbb{R}_n$:

$$P(t) : \left| \begin{array}{rcl} \mathbb{R}_n & \longrightarrow & \mathbb{R}_n \\ f & \longmapsto & P(t)f \end{array} \right.$$

The derivation operator d/dt on time-dependent matrices of $\mathcal{M}_n(\mathbb{R})$ is $\{dM(t)/dt\}_{i,j} = \{dM_{i,j}(t)/dt\}$

The rate transition matrix K (or infinitesimal generator) with $K \in \mathcal{M}_n(\mathbb{R})$ is thus defined:

$$\lim_{t \to 0} \frac{\mathrm{d}P(t)}{\mathrm{d}t} = K \tag{2}$$

Properties:

(1) for all $j \neq i$, $K_{i,j} = k_i(j)$

(2) for all $i \in S$ $K_{i,i} = -\sum_{j \neq i} K_{i,j}$

(3) Verifies the forward and backward Kolmogorov equations:

$$\forall t \geq 0 \qquad \frac{\mathrm{d}P(t)}{\mathrm{d}t} = P(t)K = KP(t) \tag{3}$$

(4) We can derive P from K as follows:

$$P(t) = \sum_{n=0}^{+\infty} \frac{(tK)^n}{n!} = \exp(Kt) \tag{4}$$

(5) Lastly, it verifies the following master equation:

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = p(t)K \tag{5}$$

with $p(t) = \mathbb{P}(X_t \in \cdot)$ being the probability distribution of the chain at time t

(1) *Proof.* when $t \to 0$, $\mathbb{P}(X_t = j \,|\, X_0 = i) = k_i(j)t$ □

(2) *Proof.* $\forall t \geq 0 \quad \sum_{j=1}^{N} P_{i,j}(t) = 1 \Leftrightarrow \lim_{t \to 0} \sum_{j=1}^{N} \frac{\mathrm{d}P_{i,j}(t)}{\mathrm{d}t} = 0 \Leftrightarrow K_{i,i} = -\sum_{j \neq i} K_{i,j}$ □

(3) *Proof.* from the semi group property of $P(t)$:

$$\forall t \geq 0 \qquad \frac{\mathrm{d}P(t)}{\mathrm{d}t} = \lim_{h \to 0} \frac{P(t+h) - P(t)}{h} = P(t) \lim_{h \to 0} \frac{P(h) - P(0)}{h} = P(t)K$$

We get the backward Kolmogorov equation the same way. □

(4) *Proof.* We have the existence since K is bounded. Starting from the forward or backward Kolmogorov equation and then switching $\frac{\mathrm{d}}{\mathrm{d}t}$ and $\sum_{n=0}^{+\infty}$ both equations are satisfied. □

(5) *Proof.* Let's note that $p(t) = \mathbb{P}(X_t \in \cdot) = \sum_{i \in S} \mathbb{P}(X(0) = i, X(t) \in \cdot) = \sum_{i \in S} \mathbb{P}(X(0) = i) P_{i,\cdot}(t)$.

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = \lim_{h \to 0} \frac{p(t+h) - p(t)}{h} = \lim_{h \to 0} \frac{\sum_{i \in S} \mathbb{P}(X(0) = i) P_{i,\cdot}(t+h) - P_{i,\cdot}(t)}{h}$$

$$= \sum_{i \in S} \mathbb{P}(X_0 = i) \lim_{h \to 0} \frac{P_{i,\cdot}(t+h) - P_{i,\cdot}(t)}{h} = \sum_{i \in S} \mathbb{P}(X(0) = i) \sum_{j \in S} P_{i,j}(t) K_{j,\cdot}$$

$$= \sum_{j \in S} \mathbb{P}(X_t = j) K_{j,\cdot} = p(t)K$$

□

We finally get the time evolution of the chain distribution:

$$p(t) = p(0) \exp(Kt), \quad \forall t > 0 \tag{6}$$

Jensen's method:

A common method to compute CTMC (continuous time markov chain) is to approximate the process by a DTMC (discrete time markov chain, the embedded chain). $\{X(t),\, t>0\}$ is regarded as a DTMC with transition matrix R (instantaneous jump probability matrix) defined as

$$R_{i,j} = k_i(j)/\sum_{\ell\neq i} k_i(\ell)$$

which is the instantaneous probability to jump from i to $j\neq i$ and $R_{i,i}=0$. The probability of the length of staying at state i follows an exponential distribution with parameter $-K_{i,i}$.

We scale the infinitesimal generator K by any positive real number $\bar{k} > 0 > \max_{i\in\mathcal{S}}\{-K_{i,i}\}$ so that transition occur at the same rate in every state (kind of like we supposed that for all $i\in S$, $\sum_{\ell\neq i} k_i$ are all equal). We consider the matrix $\bar{R}$ defined with $\bar{R}_{i,j} = k_i(j)/\bar{k}$ and $\bar{R}_{i,i} = 1-\sum_{j\neq i}\bar{R}_{i,j} = 1-\sum_{j\neq i} k_i(j)/\bar{k}$, that is

$$\bar{R} = \mathrm{I}_n + \bar{k}^{-1}K. \tag{7}$$

Let's consider the following algorithm:

(1) simulate a Poisson process $\{N(t),\ t>0\}$ with parameter $\bar{k}$

(2) simulate a discrete time Markov chain $\{Y_n, n\in\mathbb{N}\}$ with transition matrix $\bar{R}$

Set for all $t\geq 0, X_t = Y_{N(t)}$. Then, $\{X(t),\, t>0\}$ is a CTMC with generator K.

Proof. First, for $i\neq j$:

$$\begin{aligned}\mathbb{P}(X(t)=j|X(0=i) &= \sum_{n=0}^{\infty}\mathbb{P}(X(t)=j, N(t)=n|X(0)=i)\\ &= \sum_{n=0}^{\infty}\mathbb{P}(X(t)=j|N(t)=n, X(0)=i)\mathbb{P}(N(t)=n|X(0)=i)\\ &= \sum_{n=0}^{\infty}\mathbb{P}(Y_{N(t)}=j|N(t)=n, Y(0)=i)\mathbb{P}(N(t)=n|X(0)=i).\end{aligned}$$

The first term of the sum is zero since $\{Y(0)=j\}, \{Y(0)=i\}$ with $i\neq j$ are incompatible. Moreover, for all $n\geq 2$ we have:

$$\mathbb{P}(Y_{N(t)}=j|N(t)=n, Y(0)=i)\mathbb{P}(N(t)=n|X(0)=i) = \underbrace{\bar{R}^n_{i,j}(\bar{k}t)^n\exp(-\bar{k}t)/k!}_{o(t)}$$

It comes

$$\begin{aligned}\mathbb{P}(X(t)=j|X(0)=i) &= \bar{R}_{i,j}(\bar{k}t)\exp(-\bar{k}t) + o(t) = k_i(j)t(1-\bar{k}t) + o(t)\\ &= k_i(j)t + o(t)\end{aligned}$$

which is coherent with 1

Then, for $i = j$:

$$\mathbb{P}(X(t) = j|X(0) = i) = \sum_{n=0}^{\infty} \mathbb{P}(X(t) = i, N(t) = n|X(0) = i)$$

$$= \mathbb{P}(X(t) = i, N(t) = 0|X(t) = i) + \mathbb{P}(X(t) = i, N(t) = 1|X(0) = i)$$

$$+ \sum_{n\geq 2} \mathbb{P}(X(t) = i, N(t) = n|X(0) = i)$$

$$= \mathbb{P}(Y(0) = i|Y(0) = i)\mathbb{P}(N(t) = 0) + \mathbb{P}(Y(1) = i|Y(0) = i)\mathbb{P}(N(t) = 1)$$

$$+ \sum_{n\geq 2} \mathbb{P}(X(t) = i, N(t) = n|X(0) = i)$$

$$= (1 - \bar{k}t + o(t)) + \bar{R}_{i,i}(\bar{k}t + o(t)) + \sum_{n\geq 2} \underbrace{\mathbb{P}(X(t) = i, N(t) = n|X(0) = i)}_{o(t)}$$

$$= 1 - \bar{k}t + \bar{R}_{i,i}\bar{k}t + o(t) = 1 - \bar{k}t + (1 - \sum_{j\neq i} k_i(j)/\bar{k})\bar{k}t + o(t)$$

$$= 1 - \sum_{j\neq i} k_i(j)t + o(t)$$

which is coherent with 1 □

Reducibility:

Remark 1. *An irreducible finite-state markov chain is always positive recurrent.*

Proposition 1. *For an irreducible DTMC on discrete state space with transition matrix R,$R_{i,i} > 0$ implies that i is aperiodic.*

Proof. $R^2_{i,i} = \sum_{j\neq i} R_{i,j}R_{j,i} + R^2_{i,i} > 0$ and by recurrence for all $n \geq 2$, we see that $R^n_{i,i} > 0$ thus the state i is aperiodic. □

Remark 2. *For the uniformization embedded chain, $\bar{k} > 0 > \max_{i\in\mathcal{S}} -K_{i,i} \geq \sum_{l\neq i} k_i(l)$ thus $\forall i \in \mathcal{S}$, $\bar{R}_{i,i} = 1 - \sum_{j\neq i} k_i(j)/\bar{k} > 0$. So the chain is aperiodic.*

Proposition 2. *The CTMC $\{X(t), t > 0\}$ is irreducible if and only if its embedded chain with transition matrix $\bar{R}$ is irreducible.*

Proof. $\{Y_n, n \in \mathbb{N}\}$ be the embedded chain with transition matrix $\bar{R}$. We suppose $\{Y_n, n \in \mathbb{N}\}$ irreducible. Then for all $(i, j) \in S^2$ there exists a $n_{i,j} \in \mathbb{N}$ such that $\bar{R}^{n_{i,j}}_{i,j} > 0$. By construction there will be a time $\tau_{i,j} > 0$ such that

$$\mathbb{P}(X(\tau_{i,j}) = j|X(0) = i) > 0 \text{ and } \tau_{i,j} = \inf_t\{N(t) = n_{i,j}\}.$$

$\{X(t), t > 0\}$ is irreducible. Let's suppose that on the other hand $\{Y\}_n$ is reducible. Then there exists $(i, j) \in S^2$ such that for all $n \in \mathbb{N}$ $\bar{R}^n_{i,j} = 0$, which implies that for all $t > 0, P_{i,j}(t) = 0$. This is contradicting that $\{X(t), t > 0\}$ is irreducible. □

Stationary distribution:

Definition 1. *A distribution $\bar{\pi} = \{\bar{\pi}_i, i \in S\}$ is said to be a stationary distribution for the markov chain $\{Y_n, n \in \mathbb{N}\}$ if $\bar{\pi}\bar{R} = \bar{\pi}$, i.e. $\bar{\pi}_j = \sum_{i \in S} \bar{\pi}_i \bar{R}_{i,j}$, $\forall j \in S$.*

We recall firstly the ergodic theorem for aperiodic and irreducible DTMC:

Theorem 1. *If the DTMC is irreducible and aperiodic, there exists an unique probability distribution $\{\bar{\pi}_1, \cdots, \bar{\pi}_n\}$ such that*

(1) it is stationary for $\bar{R}$, $\bar{\pi}\bar{R} = \bar{\pi}$

(2) it is a limiting distribution, for all $i \in S$, $\lim_{n\to\infty} \mathbb{P}(Y(n) = j | Y(0) = i) = \bar{\pi}_j$

Proof. **Existence**

$\{Y_n, n \in \mathbb{N}\}$ be an irreducible and aperiodic chain with transition matrix $\bar{R}$ and the measure μ_i^j the mean number of times $\{Y_n\}$ is at state j before the first return at state i (noted T_i). We have that:

$$\mu_i^j = \sum_{n=1}^{\infty} \mathbb{P}(Y(n) = j, n \leq T_i | Y(0) = i)$$

$$= \sum_{\ell \in S} \sum_{n=1}^{\infty} \mathbb{P}(Y(n) = j, Y(n-1) = \ell, n-1 \leq T_i | Y(0) = i)$$

$$= \sum_{\ell \in S} \sum_{n=2}^{\infty} \mathbb{P}(Y(n-1) = \ell, n-1 \leq T_i | Y(0) = i) \bar{R}_{j,\ell}$$

$$= (\mu^j \bar{R})_i$$

Since $\{Y_n\}$ is irreducible positive recurrent, $\sum_{\ell \in S} \mu_i^\ell < \infty$, and we get $\{\mu_i^j / \sum_{j \in S} \mu_i^j, j \in S\}$ stationary distribution.

Unicity

Let μ be the measure considered above with $\mu_i = 1$ and ν any stationary measure with $\nu_i = 1$. For any $j \neq i$ we have that

$$\nu_j = \nu_i \bar{R}_{i,j} + \sum_{\ell_1 \neq i} \nu_{\ell_1} \bar{R}_{\ell_1,j}$$

$$= \bar{R}_{i,j} + \sum_{\ell_1 \neq i} \bar{R}_{i,\ell_1} \bar{R}_{\ell_1,j} + \sum_{\ell_1,\ell_2 \neq i} \ell_2 \bar{R}_{\ell_2,\ell_1} \bar{R}_{\ell_1,j}$$

$$\geq \bar{R}_{i,j} + \sum_{n=1}^{\infty} \sum_{\ell_1,\cdots,\ell_n \neq i} \bar{R}_{i,\ell_n} \bar{R}_{\ell_n,\ell_{n-1}} \cdots \bar{R}_{\ell_1,j}$$

$$= \sum_{n=0}^{\infty} \mathbb{P}(Y(n+1) = j, n+1 \leq T_i | Y(0) = i) = \mu_j$$

Now suppose that $\exists j \in S$ such that $\mu_j > \nu_j$. By irreducibility there exists $n \in \mathbb{N}$ such that $\bar{R}^n_{i,j} > 0$. The stationarity of μ and ν implies that

$$\sum_{\ell \in S} \mu_\ell \bar{R}^n_{\ell,i} = \mu_i = \nu_i = \sum_{\ell \in S} \nu_\ell \bar{R}^n_{\ell,i}$$

Thus

$$0 = \sum_{\ell \in S} (\mu_\ell - \nu_\ell) \bar{R}^n_{\ell,i} \geq (\mu_j - \nu_j) \bar{R}^n_{j,i} > 0$$

which is absurd. So we have the unicity of the stationary measure up to a constant multiple and thus we have the unicity of the stationary probability.

Limiting distribution

Let $\{W_n, n \in \mathbb{N}\}$ be a independent copy of $\{Y_n, n \in \mathbb{N}\}$, except that W starts with initial distribution π. We have that $\mathbb{P}(W_n = i) = \pi_i$ for all $n \geq 0$ and $i \in S$. The idea of the proof is to consider the bivariate process $\{Z_n = (Y_n, W_n), n \in \mathbb{N}\}$, Z being an irreducible, recurrent and aperiodic markov chain as Y and W. Let $h = \inf\{n \geq 0, Y_n = W_n\}$.

We have the following property, for all $i, j \in S$:

$$\mathbb{P}(Y_n = j, h \leq n | Y(0) = i) = \mathbb{P}(W_n = j, h \leq n)$$

Besides, $\{Z_n\}$ is recurrent so $\mathbb{P}(h \leq \infty) = 1$ whatever the initial distribution of Y. Thus, we obtain for $i, j \in S$

$$\begin{aligned}
|\mathbb{P}(Y(n) = j | Y(0) = i) - \bar{\pi}_j| &= |\mathbb{P}(Y(n) = j | Y(0) = i) - \mathbb{P}(W_n = j)| \\
&\leq \underbrace{|\mathbb{P}(Y(n) = j, h \leq n | Y(0) = i) - \mathbb{P}(W_n = j, h \leq n)|}_{=0} \\
&+ \underbrace{|\mathbb{P}(Y(n) = j, h > n | Y(0) = i) - \mathbb{P}(W_n = j, h > n)|}_{=\mathbb{P}(h > n)} \\
&= \lim_{n \to \infty} \mathbb{P}(h > n) = 0
\end{aligned}$$

□

We arrive to our main theorem:

Theorem 2. *If the CTMC is irreducible, then it admits an unique limiting distribution which is also its unique stationary distribution.*

Proof. If the CTMC is irreductible then its embedded chain is also irreductible and thus aperiodic. We apply the ergodic theorem as described in Theorem 1. If $\bar{\pi}$ is stationary for $\bar{R}$, then it is stationary as well for the CTMC. Indeed, if we assume $X_0 \sim \bar{\pi}$ we have for all $j \in S$:

$$\mathbb{P}(X(t) = j) = \sum_{i=1}^{N} \sum_{n=1}^{\infty} \mathbb{P}(X(t) = j | X(0) = i, N(t) = n) \mathbb{P}(N(t) = n) \bar{\pi}_i = \cdots = \bar{\pi}_j.$$

Now, from (2) we can write:

$$\forall \epsilon > 0, \exists n_0 \in \mathbb{N}, n \geq n_0 \Rightarrow |\bar{R}^n_{i,j} - \bar{\pi}_j| < \epsilon/2.$$

Independently, since $\lim_{t\to\infty}\mathbb{P}(N(t)<n_0)=0$, we can choose t_0 such that for all $t\geq t_0, \mathbb{P}(N(t)<n_0)<\epsilon/4$. Going back to the CTMC, we want to show that there exists a probability distribution $\{\alpha_1,\cdots,\ \alpha_n\}$, such that for all $i\in S$ $\lim_{t\to\infty}\mathbb{P}(X(t)=j|X(0)=i)=\alpha_j$. Let $\epsilon>0$ and $t\geq t_0$, we have for all $(i,j)\in S^2$:

$$|\mathbb{P}(X(t)=j|X(0)=i)-\bar{\pi}_j|=$$
$$|[\mathbb{P}(X(t)=j|X(0)=i,\ N_t\geq n_0)-\bar{\pi}_j]\mathbb{P}(N(t)\geq n_0)+$$
$$[\mathbb{P}(X(t)=j|X(0)=i,\ N(t)<n_0)-\bar{\pi}_j]\mathbb{P}(N(t)<n_0)|$$
$$\leq \underbrace{|\bar{R}_{i,j}^{n_0+u(t)}-\bar{\pi}_j|}_{<\epsilon/2}\mathbb{P}(N(t)\geq n_0)+|\mathbb{P}(X(t)=j|X(0)=i,\ N(t)<n_0)-\bar{\pi}_j|\underbrace{\mathbb{P}(N(t)<n_0)}_{<\epsilon/4}$$
$$\leq\epsilon$$

since $u(t)\geq 0$ and for two real numbers $(a,\ b)\in[0,\ 1\ ], |a-b|\leq 2$. We conclude that the limiting distribution exists and coincides with the stationary distribution. The uniqueness of the invariant distribution for the CTMC follows from the uniqueness for the DTMC. □

Remark 3. *We can see that $\bar{\pi}$ does not depend on the actual choice of $\bar{k}$ in the definition of $\bar{R}$. Indeed, for all $\bar{k}>\max_{i\in\mathcal{S}}\{-K_{i,i}\}$ we have that*

$$\bar{\pi}\bar{R}=\bar{\pi}\Leftrightarrow\bar{\pi}(\bar{R}-\mathrm{I}_n)=0\Leftrightarrow(\bar{k})^{-1}\bar{\pi}K=0\Leftrightarrow\bar{\pi}K=0$$

Reversibility

The dynamics of the detailed molecular system is considered to be reversible, so that the direction of the time flow does not modify the distribution of the chain. For example,if we consider a DTMC $\{X_n,n\in\mathbb{N}\}$ with transition matrix R then for any $(i,j)\in S^2$, and $n_1<n_2=n_1+u$, we would have

$$\mathbb{P}(X(n_1)=i,X(n_2)=j)=\mathbb{P}(X(n_2)=i,X(n_1)=j)$$
$$\Leftrightarrow\mathbb{P}(X(n_1)=i)\mathbb{P}(X(n_2)=j|X(n_1)=i)=\mathbb{P}(X(n_1)=j)\mathbb{P}(X(n_2)=i|X(n_1)=j)$$
$$\Leftrightarrow\mathbb{P}(X(n_1)=i)R_{i,j}^u=\mathbb{P}(X(n_1)=j)R_{j,i}^u.$$

which is an impossible condition. Thus, when considering time reversibility we assume another condition which is that the chain is supposed to be at stationary regime.

The resulting theorem is:

Theorem 3. *A DTMC is said to be reversible if and only if for all $u\in\mathbb{N}$ and $(i,j)\in S^2$*

$$R_{i,j}^u=R_{j,i}^u\frac{\pi_j}{\pi_i}$$

Remark 4. *(1) For $u=1$ this equation is known as the* ***detailed balance equation****.*

(2) having the detailed balance equation is sufficient for the equation to hold for any u.

Proof. For $u = 2$:

$$\pi_i R_{i,j}^2 = \sum_{n \in S} \underbrace{\pi_i R_{i,n} R_{n,j}}_{\pi_n R_{n,i}} = \sum_{n \in S} R_{n,i} R_{j,n} \pi_j = R_{j,i}^2 \pi_j$$

Then by induction this holds for all $u > 2$ □

The situation is the same for CTMC as the former theorem translates to the following equation:

$$\pi_i P_{i,j}(t) = \pi_j P_{j,i}(t)$$

Theorem 4. *$\{X_t, t > 0\}$ is time reversible at equilibrium if and only if K is π reversible.*

Proof. $\Rightarrow$ We suppose that $\{X_t, t > 0\}$ is time reversible at equilibrium, with $\pi_i = \mathbb{P}(X(t) = i)$. Then for $h > 0$,

$$\mathbb{P}(X(t) = i, X(t+h) = j) = \mathbb{P}(X(t) = j, X(t+h) = i)$$

and so,

$$\pi_i \frac{\mathbb{P}(X(t+h) = j | X(t) = i)}{h} = \pi_j \frac{\mathbb{P}(X(t+h) = i | X(t) = j)}{h}$$

Letting $h \lim 0$ we get that K is π reversible.

$\Leftarrow$ We suppose that K admits detailed balance for π. We recall that

$$P(t) = \exp(Kt)$$

so that

$$P_{i,j}(t) \sum_{n=0}^{+\infty} \frac{t^n}{n!} \{K^n\}_{i,j}$$

it is clear that if K is π reversible, then $\{X_t, t > 0\}$ is time reversible at equilibrium. □

Thus, we will call **detailed balance equation**:

$$\pi_i K_{i,j} = \pi_j K_{j,i}, \ \ \forall i, j \in S \tag{8}$$

2.2 Spectral Properties

2.2.1 Spectral decomposition

To know that $\{Y_n, n \in \mathbb{N}\}$ or $\{X_t, t > 0\}$ are reversible do not provide information on whether K and R are symmetric matrices. In fact, they are not unless the stationary distribution π is the uniform distribution. So we have to assess if K (or R) are diagonalizable. We consider the Hilbert space $\mathcal{L}_2(\pi) \subset \mathbb{R}^N$ with N the size of S:

(1) $(x, y) \in \mathbb{R}^N$, $\langle x, y\rangle = \sum_{i=1}^N \pi_i x_i y_i$

(2) $||x||^2 = \sum_{i=1}^N \pi_i (x_i)^2 < \infty$

(3) And the operator norm: $||A|| = \sup_{x \in \mathbb{R}^N, ||x|| > 0} \frac{||Ax||}{||x||}$

Be π the stationary distribution of $\{X_t, t > 0\}$ (resp. $\{Y_n, n \in \mathbb{N}\}$), then K (resp. R) is a self-adjoint operator, thanks to detailed balance equation, on $\mathcal{L}_2(\pi) \subset \mathbb{R}^N$:

$$\langle Kx, y\rangle = \sum_{i=1}^N \pi_i (\sum_{j=1}^N K_{i,j} x_j) y_i = \sum_{j=1}^N \pi_j x_j \underbrace{\sum_{i=1}^N K_{j,i} y_i}_{\{Ky\}_j} = \langle x, Ky\rangle$$

Thus, the spectral theorem tells us that K (resp. R) is diagonalizable and there is an orthonormal basis of eigenvectors such that:

$$K = \Psi \Delta \Psi^{-1} \tag{9}$$

with:

(1) Ψ's columns are K right eigenvectors.

(2) Δ is the diagonal matrix formed of the eigenvalues of K: $\Delta = diag(\lambda_1, \ldots, \lambda_N)$.

(3) Ψ^{-1}'s rows are K left eigenvectors.

Spectral features of R

Jensen's inequality tells us that for any $x \in \mathbb{R}^n$

$$\{Rx\}_i^2 = (\sum_{j=1}^N x_j R_{i,j})^2 \leq \sum_{j=1}^N x_j^2 R_{i,j}$$

since $R_{i,j}$ are probabilities. It comes that

$$||Rx||^2 = \sum_{i=1}^N \pi_i \{Rx\}_i^2 \leq \sum_{j=1}^N x_j^2 \sum_{i=1}^N \pi_i R_{i,j} = \sum_{j=1}^N \pi_j x_j^2 = ||x||^2$$

implying $||R|| \leq 1$. $\xi_1 = 1$ is the eigenvalue of R with right eigenvector $x_1 = 1_N$ and left eigenvector $y_1 = \pi$.

$$Rx_1 = x_1 \Rightarrow ||Rx_1|| = ||x_1||$$

implying $||R|| \geq 1$. Thus, $||R|| = 1$ Let's consider another eigenvalue ξ_2 with right eigenvector x_2 and left eigenvector y_2, we have that

$$\frac{||Rx_2||}{||x_2||} \Rightarrow |\xi_2| \leq 1$$

The Perron-Frobenius theorem states that a square matrix with positive entries (which is the case for our stochastic matrices R and $P(t)$) has a unique largest eigenvalue, which in this case is $\xi_1 = 1$. Thus, $|\xi_2| < 1$ and likewise for the other eigenvalues of R.

Spectral features of K

We note $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ the N eigenvalues of K and $\{\psi_i^{(R)}, i \in \{1, \cdots, N\}\}$ their respective right eigenvectors. For $t > 0$ and $i \in \{1, \ldots, N\}$ we have that:

$$P(t)\psi_i^{(R)} = \sum_{n=0}^{N} \frac{t^n}{n!} K^n \psi_i^{(R)} = \sum_{n=0}^{N} \frac{(\lambda_i t)^n}{n!} \psi_i^{(R)} = \exp(\lambda_i t)\psi_i^{(R)}$$

For any $i \in \{1, \ldots, N\}$, $\psi_i^{(R)}$ is also an eigenvector of $P(t)$ with eigenvalue $\xi_i = \exp(\lambda_i t)$. For $t > 0$

$$P(t) = \Psi \exp(\Delta t)\Psi^{-1}$$

Since $P(t)$ is a stochastic matrix we can apply the previous analysis and see that for all $i \in \{2, \ldots, N\}$ and $t > 0$:

$$\xi_i < \xi_1 = 1 \Rightarrow \exp(\lambda_i t) < \exp(\lambda_1 t) = 1$$

Thus,

$$\lambda_1 = 0 > \lambda_2 \geq \cdots \geq \lambda_N \tag{10}$$

We can now rewrite 6:

$$p(t) = p(0)\exp(Kt) = p(0)\Psi \exp(\Delta t)\Psi^{-1}$$

$$= p(0)\{\sum_{n=1}^{N} \Psi_{i,n} \exp(\lambda_n t)\Psi_{n,j}^{-1}\}_{1 \leq i,j \leq N}$$

$$= \{\sum_{i=1}^{N} p_i(0) \sum_{n=1}^{N} \Psi_{i,n} \exp(\lambda_n t)\Psi_{n,j}^{-1}\}_{1 \leq ,j \leq N}$$

Thus:

$$p(t) = \sum_{i=1}^{N} \sum_{j=1}^{N} (p_j(0)\psi_i^{(R)}(j)) \exp(\lambda_i t)\psi_i^{(L)} \tag{11}$$

with $\Psi_{i,:}^{-1} = \psi_i^{(L)}$ and $\Psi_{:,i} = \psi_i^{(R)}$ for all $i \in \{1, \ldots, N\}$.

We note that a similar expression could be found for the DTMC R since it is π-reversible and is thus diagonalisable.

We know that $\psi_1^{(R)} = 1$ and $\psi_1^{(L)} = \pi$. Since K is self-ajoint we have the following normalization condition for our left and right eigenvectors, $\forall i, j \in \{1, \ldots, N\}$:

(1)

$$\psi_i^{(L)}(j) = \pi_j \psi_i^{(R)}(j) \tag{12}$$

(2)

$$\begin{aligned} &\sum_{n=1}^{N} \psi_i^{(R)}(n)\psi_j^{(R)}(n)\pi_n \\ &= \sum_{n=1}^{N} \psi_i^{(L)}(n)\psi_j^{(L)}(n)/\pi_n \\ &= \sum_{n=1}^{N} \psi_i^{(L)}(n)\psi_j^{(R)}(n) \\ &= \delta_{ij} \end{aligned}$$

The probability distribution of $\{X(t),\, t > 0\}$ corresponds to the superposition of N processes:

(1) a stationary process, π

(2) $N-1$ transient processes $\rho_i(t), t > 0, i \in \{2, \cdots, N\}$ such that:

$$\rho_i(t) = \sum_{j=1}^{N} p_j(0)\psi_i^{(R)}(j) \exp(-t/\tau_i)\psi_i^{(L)} \tag{13}$$

with $\tau_i = 1/\lambda_i$ being the relaxation time of process i

We draw attention to the following remarks:

Remark 5. *The left eigenvectors of K apart from $\psi_1^{(L)} = \pi$ will all add up to 0. Indeed, for $i \in \{2, \ldots, N\}$:*

$$\sum_{n=1}^{N} \psi_i^{(L)}(n) = (1/\lambda_i) \sum_{n=1}^{N} \sum_{m=1}^{N} \psi_i^{(L)}(m) K_{m,n} = (1/\lambda_i) \sum_{m=1}^{N} \psi_i^{(L)} \sum_{n=1}^{N} K_{m,n} = 0$$

Thus, the transient processes $\rho_i(t), t > 0, i \in \{2, \ldots, N\}$ are viewed as flows which move the probability mass across different states. Those flows are proportional to the left eigenvectors which therefore act as probability mass transfer directions.

Remark 6. *If $p(0) = \pi$ then $\forall t > 0$, $p(t) = \pi$*

Indeed, $\forall i, j \in \{1, \ldots, N\}$ $\psi_i^{(L)}\psi_j^{(R)} = \delta_{ij}$ and $\psi_1^{(L)} = \pi$.

Finally,

Remark 7. *If the system is at equilibrium but perturbed by a vector proportional to one of the left eigenvectors $\psi_i^{(L)}$ then the typical time to reach back equilibrium is $\tau_i = 1/\lambda_i$. Indeed, $p(0) = \pi + \delta\psi_i^{(L)}$ and the probability density vector is*

$$p(t) = \pi + \delta \sum_{j=1}^{N} \psi_i^{(L)}(j)\psi_i^{(R)}(j) \exp(\lambda_i t)\psi_i^{(L)} \tag{14}$$

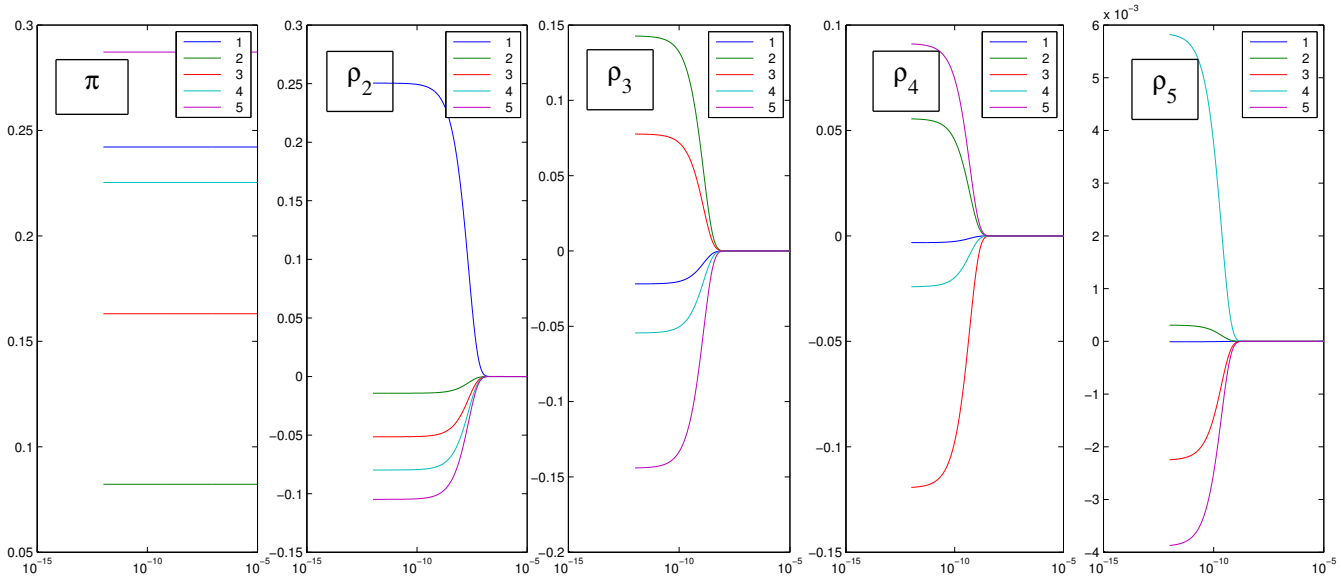

Figure 1: Representation of the stationary process and the $N-1$ transient processes throughout time (in log scale) for five states.

We can see that all the transient processes add up to zero for $t > 0$.

2.2.2 Conformational Clustering: Two-States Representation

2.2.2.1 Folded/Unfolded Despite the complexity of protein folding, experiences tend to show that, in some situations, it can be approximated by a simpler process consisting of only two macroscopic states (which are aggregations of the microstates) namely folded and unfolded.

Two-state kinetic processes are physically justified if protein molecules rapidly equilibrate between different unfolded conformations prior to complete folding (or conversely). Majority of the protein population is localized in two basins separated by a narrow bottleneck, so that the intrabasin relaxation (the equilibration within the folded and unfolded states) occurs much faster than the interbasin exchange (which corresponds to the folding relaxation time) and thus kinetics of equilibration is essentially single exponential.

2.2.2.2 The Spectral Gap A distinctive feature of two-state proteins is thus the presence for the CTMC $\{X(t),\, t > 0\}$ of a gap in the eigenvalue spectrum after the first non-zero eigenvalue λ_2, so such that:

$$\lambda_1(=0) - \lambda_2 \ll \lambda_2 - \lambda_3.$$

We will call spectral gap the following expression:

$$\gamma = \frac{\lambda_3 - \lambda_2}{\lambda_2} = \frac{\lambda_3}{\lambda_2} - 1 = \frac{\tau_2}{\tau_3} - 1 \tag{15}$$

$\{X(t),\, t > 0\}$ is said to have a spectral gap if: $\gamma \geq 10$. Indeed, if we set:

1. $\omega = t/\tau_2$
2. $\alpha_i = \sum_{j=1}^{N} p_j(0)\psi_i^{(R)}(j)$ for $i \in \{2, \ldots, N\}$

We have:

$$p(\omega) = \pi + \alpha_2 \exp(-\omega)\psi_2^{(L)} + \alpha_3 \exp(-\omega(\gamma+1))\psi_3^{(L)} + \cdots$$

If $\gamma \geq 10$ then all the flows $\rho_i,\ \ i \geq 3$ are negligible compared to ρ_2:

The amplitude between ρ_3 and ρ_2 at $t = \tau_2$ is

$$\approx \frac{\alpha_3}{\alpha_2} \exp(-10)$$

Thus we can ignore the flows above ρ_2 as an approximation.

Usually, in a two-states like system, folded and unfolded macrostates respectively correspond to the following U and F here described arbitrarily:

1. $U = \{1 \leq i \leq N, \alpha_2\psi_2^{(L)}(i) < 0\}$
2. $F = \{1 \leq i \leq N, \alpha_2\psi_2^{(L)}(i) > 0\}$

The flow ρ_2 transfers probability mass from the states in U to those in F. The speed of the probability mass transfer between states through ρ_2 is (in the case $\gamma \gg 1$) much lower than that of the other flows. Since each flow converges to zero, the error committed by dropping the $n-2$ fastest flows will have no impact on the dynamic of $p(t)$ once the characteristic time τ_3 is passed.

2.2.2.3 Splitting probability In a variant of the above approach, one can identify microstates that can be faithfully clustered according to their splitting probability σ_i.

We define it as follow:

Starting out in state i, the protein folds before it unfolds with probability σ_i.

In such a system all microstates can be separated into three groups: Unfolded U ,folded F, and intermediate I. Microstates forming the intermediate group correspond to those of the bottleneck region (they are far fewer than the rest), they are called the transition states. Hence, $\sigma_i = 1$ and $\sigma_i = 0$ for respectively $i \in F$ and $i \in U$. We can write:

$$\sigma_i = \mathbb{P}(\exists t_1, t_2 \in \mathbb{R}, t_1 < t_2, X(t_2) \in U, X(t_1) \in F | X(0) = i) \tag{16}$$

We define the transition state ensemble as the one formed by the states for which σ_i is close to 0.5. The choice of the bottleneck boundaries is arbitrary. In the case of a two-state system, Berezhkovskii and Szabo showed (see [Berezhkovskii and Szabo, 2004]) that σ_i can be approximated to a good approximation as:

$$\sigma_i = \frac{\psi_2^{(R)}(i) - min_j(\psi_2^{(R)}(j))}{max_j(\psi_2^{(R)}(j)) - min_j(\psi_2^{(R)}(j))} \tag{17}$$

with $\psi_2^{(R)}$ corresponding to the second right-hand eigenvector of K, shifted and scaled to the interval $[0, 1]$. The states are thus grouped in such a way that those with $\sigma_i < 1/2$ belong to one class, and those with $\sigma_i \geq 1/2$, to the other.

2.3 Assigning the Conformations

Buchete, Hummer, Buchner, Murphy and Kubelka (see [Buchete and Hummer, 2008] and [Buchner et al., 2011]) showed how to construct coarse master equation models from simulation data. We will give here a glimpse of the proceeding to have a proper assignment of conformations to coarse states.

One has to wonder how to initially partition the state space in a way that doesn't lose too much of the information of the time-evolution trajectory $\{X(t),\, t > 0\}$ of the system.

Indeed, to get a clustering of the energy landscape of a wide selection of biomolecular systems which justifies the assumption made of Markovian dynamics, no general procedure is available.

However, for small peptide like systems (the kind of protein we are considering in this report), a satisfying way to do so is to partition the state space by geometrical clustering where each cluster should represent a set of structures that the dynamics remains in for a long time before jumping to another cluster.

A good example of such geometrical clustering is the CBA (conformation based assignment) where the Ramachandran Free Energy Surface is considered. The energy of the protein is seen as a function of two torsion angles of the polypeptide chain which also called Ramachandran angles (and denoted Phi and Psi).

Following the value of the angles the conformations are separated into different states.

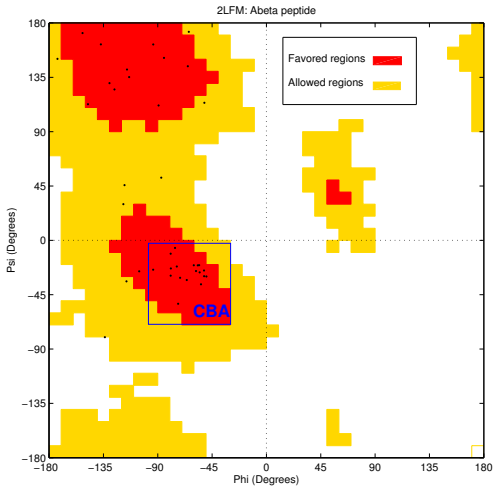

Figure 2: Ramachandran plot for an Abeta peptide, done thanks to the Ramachandran Toolbox from Matlab. The blue square illustrates the choice of a conformational state by the CBA method.

3 Results and Discussion

3.1 Designed Method to simulate Random Rate matrices

3.1.1 Fitting the Model

As we would like to find a convenient method to simulate transition rate matrices for a given number N of microstates, we have to define some conditions to make our results meaningful.

3.1.1.1 Distribution of the Rates of Transition We want positive transition rates with big entropy, in order to keep a broad approach of the problem. I considered for our simulations the exponential distribution since it is the least biased distribution of non-negative numbers.

$K_{i,j} \sim \exp(10^{\beta})$ with β being the tuning parameter.

3.1.1.2 Connectivity We have to define which microstates are connected. For the purpose of simplicity we will usually consider that transitions only occur between neighboring microstates, which otherwise would greatly enhance the computational complexity. A more complete model could be considered in which transition rates are decreasing the further their related microstates are far from each other. We define the bandwidth m which corresponds to the number of non-zero lower diagonals.

3.1.1.3 Relaxation Times Finally, we try to get realistic relaxation times. We want relaxation times which are on par with the ones of small peptides, so between 10^{-6} and 10^{-12} seconds. To do so we make the tuning parameter and the number of diagonals set to zero $(N - m - 1)$ fluctuate.

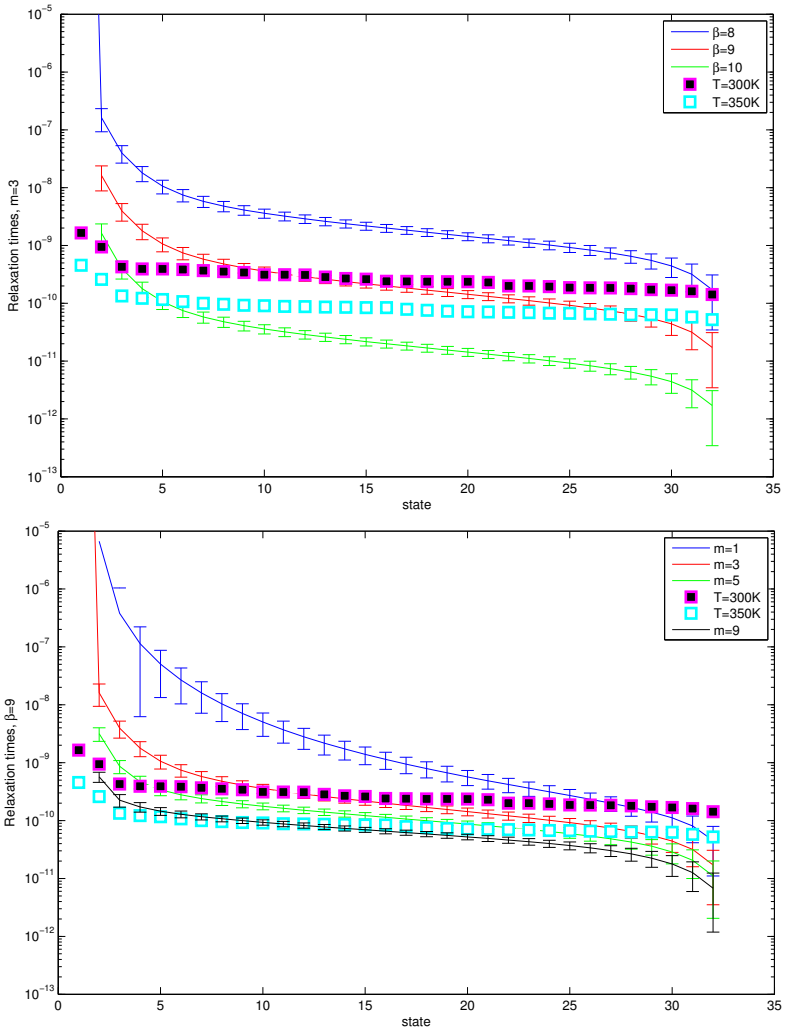

Figure 3: Distribution of the relaxation times for random exponential rate matrix of $N = 32$ microstates, compared to the ones of a small peptide, with several parameters and connectivity considered.

We then chose

$$\beta = 9$$

and keep m adjustable.

3.1.2 Computational Simulation

3.1.2.1 Algorithm We get for our Random Matrix Simulation the following algorithm:

(1) simulate the vector of equilbrium probability

$$\pi = (\alpha_1, \ldots, \alpha_N) / \sum_{i=1}^{N} \alpha_i$$

where $\alpha_i \sim \text{unif}(0, 1)$.

(2) raw rate matrix $\mathbf{K}$ (squarred matrix of size N)

- lower triangular:
  - iid entries with exponential distribution $K_{ij} \sim \exp(10^9)$ on the first m diagonals ($i > j$, m positive integer).
  - other entries are set to zero.
- upper triangular: deterministic so as to satisfy detailed balance

$$K_{i,j} = K_{j,i} \frac{\pi_j}{\pi_i}$$

- diagonal: deterministic entries to make the sum of the rows equals zero.

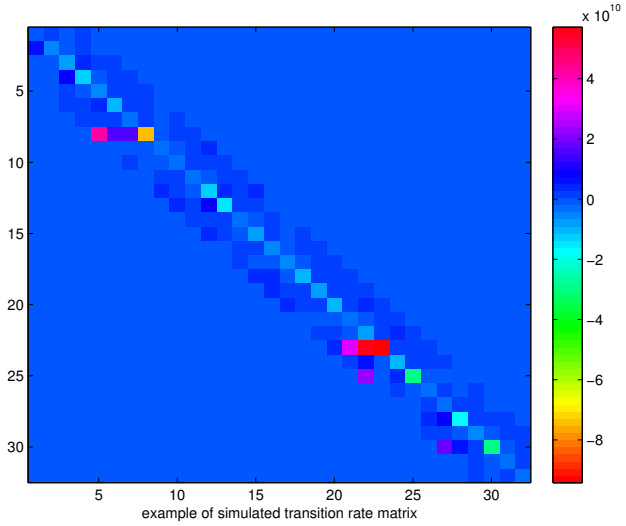

Figure 4: Example of simulated exponential rate matrix K (for $N = 32$ microstates) scaled in s^{-1}.

3.1.2.2 Monte Carlo Method We simulate a large number of iid rate transition matrices and store each time the interesting output value. To do so allow us to get the distribution of the output of interest. This method is really interesting in that by generating a large number of data it allows us to get graphical results and to get quite easily the "sensitivity" of inputs on results.

We recall that the monte carlo method is based on the **weak law of large numbers**.

Theorem: Let $X_1, X_2, \ldots, X_N$ be iid Lebesgue integrable random variables with expected value $E(X_i) = x$ for all i. We make the assumption of finite variance. Then for any positive number ϵ,

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - x| \geq \epsilon) = 0$$

Proof: We apply Chebyshev inequality to the random variable $\bar{X}_n$ with

$$\bar{X}_n = \frac{1}{n}(\sum_{i=1}^{n} X_i).$$

3.1.2.3 Bootstrapping To control and check the stability of our results, we will use the Bootstrap method.

Principe:

We simulate several times a random sample of transition rate matrices (iid) of same size N. So, although each resample will have the same number of elements, each one will randomly depart from the others. And because the elements in these resamples vary slightly, the statistic of interest written $\hat{\theta}_i$, calculated from one of these resample will take on slightly different values. The bootstrap method asserts that the relative frequency distribution of these $\hat{\theta}_i$ is an estimate of the sampling distribution of the statistic of interest $\hat{\theta}$.

We will base outselves on random samples of around $10,000$ iid rate transition matrices. The following graph gives an instance of gap probability density function in the case $N = 32$ and $m = 3$. For each sample, we calculate for each iteration of rate transition matrix its eigenvalues and then store its spectral gap, getting through normal kernel density estimation (done thanks to the ksdensity matlab function which is non-parametric) the gap distribution, we bootstrap to get the final gap distribution.

3.2 Two-States Representation

We recall that the spectral gap of a rate transition matrix is indicative about knowing wether a simpler representation of the protein dynamics is possible (for $gap \approx 10$ the process is considered to be two-states like).

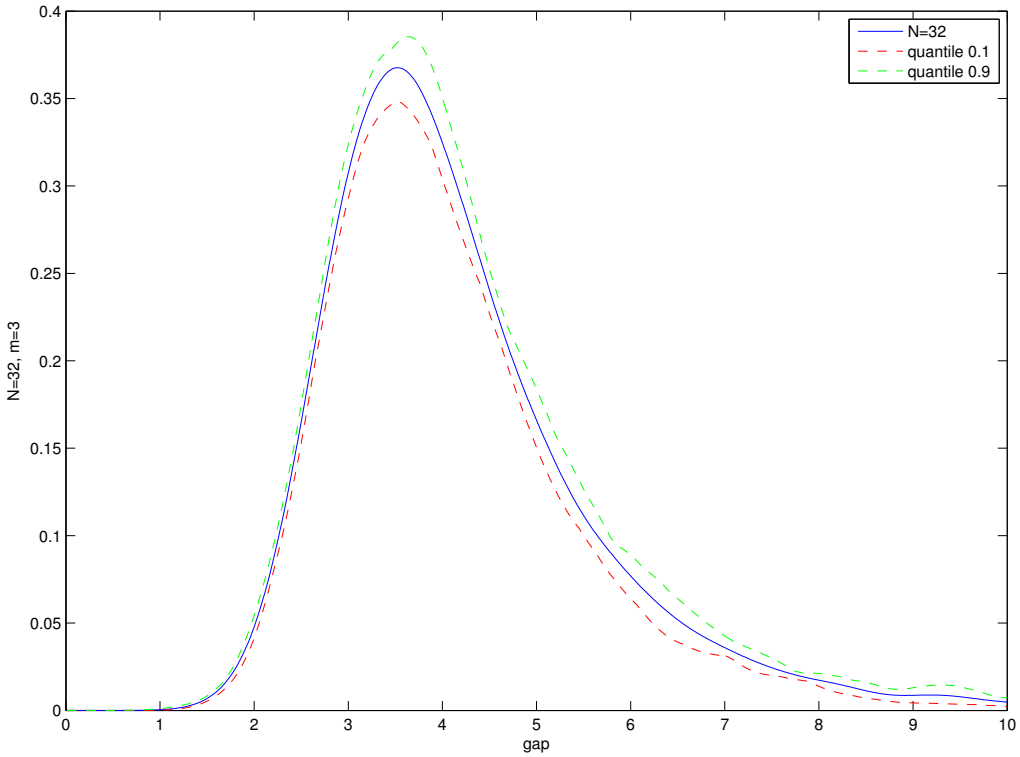

Figure 5: Gap probability density function for random rate matrix (band diagonal) of size $N = 32$ and bandwidth $m = 3$, with quantiles.

We can see that the vast majority of the transition rate matrices will have a spectral gap $\gamma < 10$, implying a priori that most processes won't be well represented by a two-states like system. Thus, it draws into question the relevance of the two-states approximation and the use of the splitting probability. In practice, we considered that a gap $7 \leq \gamma \leq 10$ was enough to represent the system as two-states like.

3.2.1 Size/Connectivity

In order to get a good grasp at how common a two-states representation is, and its dependences on the characteristics of the matrix, we study the distribution of the gap in several cases of number of microstates N and number of non zero lower diagonals m (the bandwidth).

For performance issues we only considered matrices from size 8×8 to 200×200. In terms of connectivity we worked from single diagonal matrices to full matrices. We get the following results:

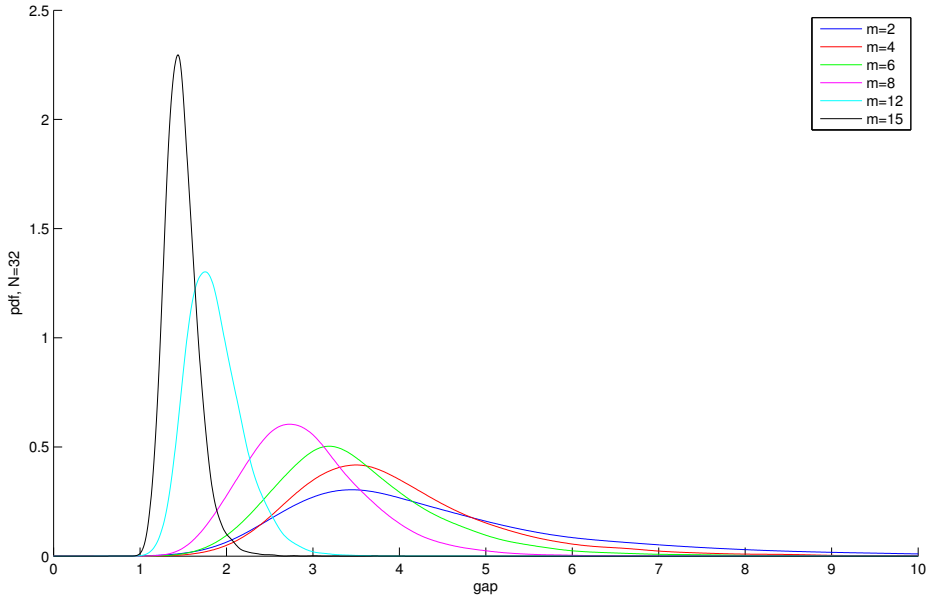

Figure 6: Gap probability density function for band diagonal random rate matrix of different bandwidths.

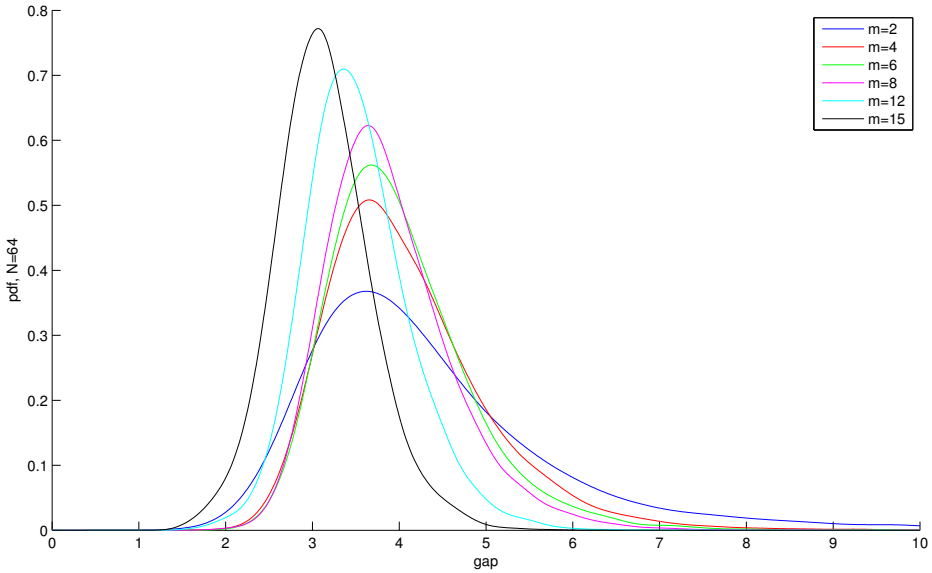

Figure 7: Gap probability density function for band diagonal random rate matrix of different bandwidths.

We can see that as it was expected the higher the bandwidth m is the lower will be the probability to get a good two-states like system.

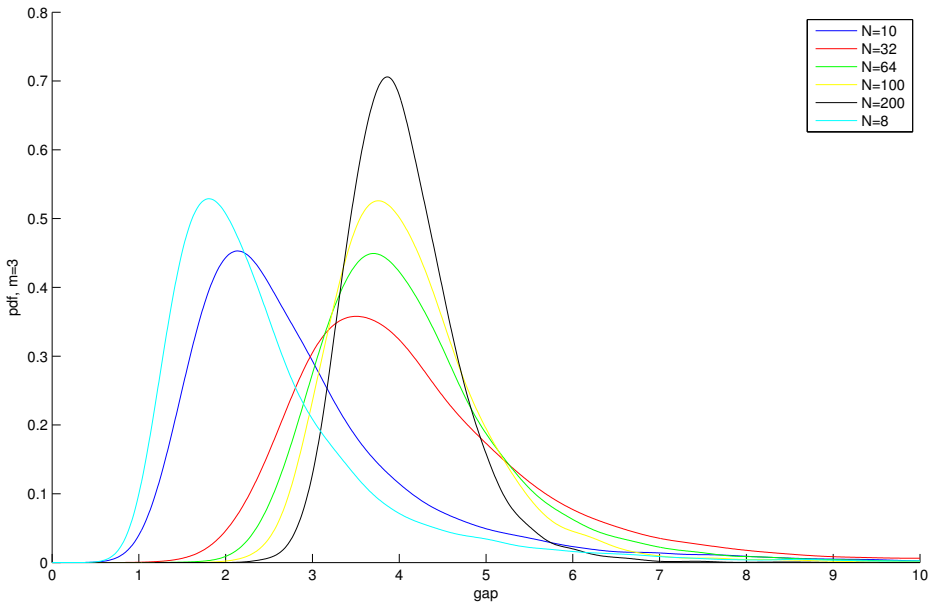

Figure 8: Gap probability density function for band diagonal random rate matrix of different sizes.

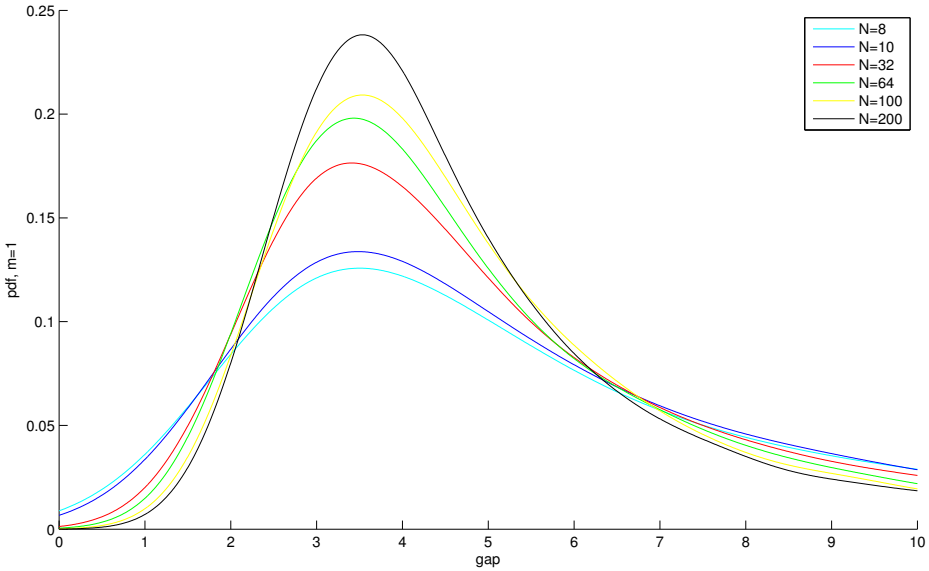

Figure 9: Gap probability density function for band diagonal random rate matrix of different sizes.

For N the interpretation is made more difficult since it raises into question the sparsity of the matrix. Indeed, for small matrices the value of N will matter less than the density of the rate transition matrix concerning the probability to get a good two-states like system. Nonetheless, we observe that the bigger N, the lower the variance of the gap will be.

3.2.2 The Error Committed

We want to quantify the two-states approximation we are doing while neglecting the flows ρ_i for $i \in \{3, \ldots, N\}$.

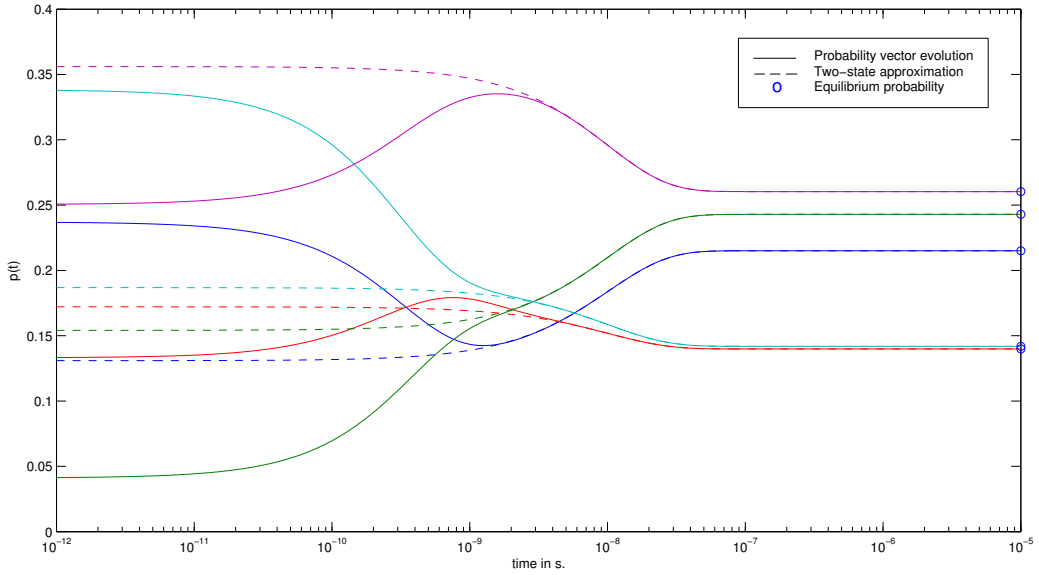

Figure 10: Representation of the probabilities $p(t)$ and its approximation $\bar{p}(t)$ throughout time (in log scale) for five states, with $\gamma = 7.8$.

We calculate the mean squared deviation and try to find its correlation with the spectral gap γ. In order to find meaningful results we only simulate matrices that have their second eigenvalue λ_2 in a given (small) interval. Be $\bar{p}(t),\ t > 0$ the approximation and ϵ the MSD. Then:

$$\epsilon = \int_t ||p(t) - \bar{p}(t)||^2 \mathrm{d}t$$

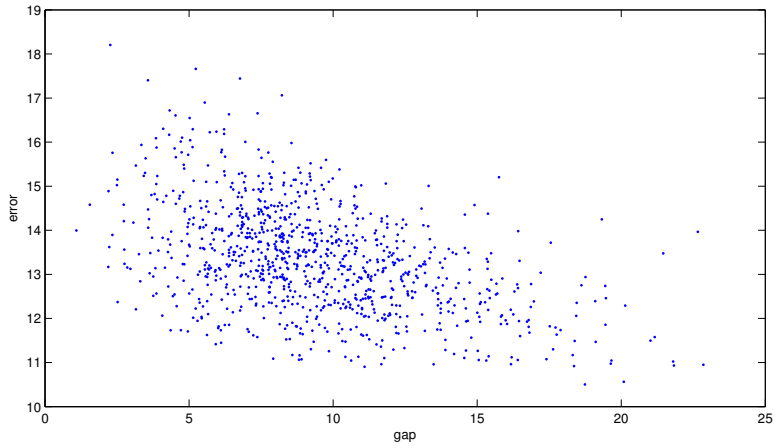

Figure 11: Correlation between γ and the error.

We do not obtain a clear correlation between the value of γ and the effectiveness of the two-states approximation as we would have hoped.

As we shall see now, we lack good control over our two-states approximation. Since

$$p(0) = \pi + \sum_{i=2}^{N} \sum_{j=1}^{N} (p_j(0)\psi_i^{(R)}(j))\psi_i^{(L)}$$

we can write that:

$$p(t) = p(0) + \sum_{i=2}^{N} \underbrace{\sum_{j=1}^{N} (p_j(0)\psi_i^{(R)}(j))(\exp(\lambda_i t) - 1)\psi_i^{(L)}}_{\phi_i(t)} \tag{18}$$

where ϕ_i acts as a flow on the initial probability measure. However, the interpretation of ϕ_i is made difficult for the reason that, considered independently, the sole action of ϕ_i on $p(0)$ may actually not result in a probability measure.

Let $\mathbb{M}_N := \{\mu \in \mathbb{R}^N, \mu_i \in (0,1), \sum_i \mu_i = 1\}$ be the set of probability measures on $\mathbb{R}^N$.

Let's suppose now that for a subset $J \subseteq \{2, \ldots, n\}$

$$\bar{p}_J(t) := p(0) + \sum_{i \in J} \phi_i(t) \in \mathbb{M}_N$$

can serve as an approximation of $p(t)$.

Proposition 3. *Let K be a transition rate matrix, $(p(0), p) \in \mathbb{M}_N^2$. For any subset $J \subseteq \{2, \cdots, N\}$, if $\bar{p}_J(0) \in \mathbb{M}_N$ then for all $t > 0$, $\bar{p}_J(t) \in \mathbb{M}_N$.*

Proof. We know thanks to remark 5 that all the left eigenvectors of K apart from $\psi_1^{(L)} = \pi$ add up to 0. Thus, for any subset $J \subseteq \{2, \ldots, N\}$ and $t > 0$,

$$\sum_{i=1}^{N} \bar{p}_{J,i}(t) = 1$$

Next, we have that for all $i \in \{1, \ldots, N\}$, $J \subseteq \{2, \ldots, N\}$ and $t > 0$,

$$\bar{p}_{J,i}(t) = \pi_i + \sum_{j \in J} \sum_{\ell=1}^{N} (p_\ell(0)\psi_j^{(R)}(\ell)) \exp(\lambda_j t)\psi_j^{(L)}(i) \in (0, 1)$$

$$\Leftrightarrow |\sum_{j \in J} \sum_{\ell=1}^{N} (p_\ell(0)\psi_j^{(R)}(\ell)) \exp(\lambda_j t)\psi_j^{(L)}(i)| \leq \min\{\pi_i, 1 - \pi_i\}$$

and $\bar{p}_J(0) \in \mathbb{M}_N$ is a sufficient condition for this inequality to hold. Thus, for all $t > 0$, $\bar{p}_J(t) \in \mathbb{M}_N$. □

However,

Proposition 4. *Let K be a transition rate matrix, $(p(0), p) \in \mathbb{M}_N^2$. For any subset $J \subseteq \{2, \ldots, N\}$, it is not true that $\bar{p}_J(0) \in \mathbb{M}_N$*

Proof. We give a counterexample: Consider the CTMC where $N = 3$, $J = 2$, $X(0) = 1$, and π is the uniform distribution. Thus $k_i(j) = 1/3$ for $i \neq j$. The three left eigenvectors of K are $\psi_1^{(L)} = [1/3, 1/3, 1/3]$, $\psi_2^{(L)} = [1, -1, 0]$ and $\psi_3^{(L)} = [-1/2, -1/2, 1]$. Since in this case K is symmetric then for all $n \in \{1, \ldots, N\}$ $\psi_n^{(R)} = (\psi_n^{(L)})^T$. Then:

$$\bar{p}_J(0) = \pi + \psi_2^{(L)} = [4/3, -2/3, 1/3]$$

which is not a probability measure. □

Thus, our two-states approximation is doable only for those processes that have $\bar{p}_{J=2}(0)$ as a probability measure.

It seems however that the study of γ is not enough to thoroughly assess whether a two-states representation of the protein dynamics is possible.

3.2.3 Relevance of the Splitting Probability

In the case of a two-states like system an ordering of the states between Folded and Unfolded macrostates is doable, thanks to the splitting probability.

We note that a study of the occurence of transition states is possible, which we recall are microstates for which σ_i is close to 0.5, the choice of boundaries being arbitrary.

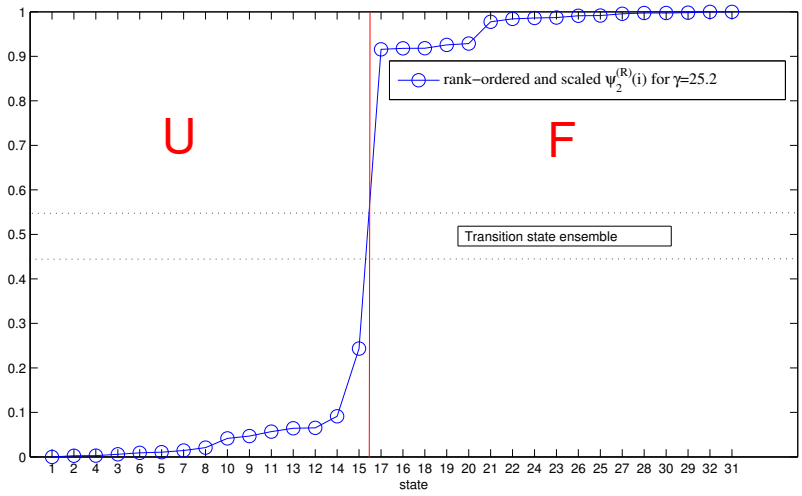

Figure 12: Example of rank-ordered states in the case of a good gap for $N = 32$.

Simulations of MCMC show that the expression of the splitting probability given by Berezhkovskii and Szabo is relevant for $\gamma \geq 7$.

In an attempt to better understand the method Berezhkovskii and Szabo used to find their approximation of the splitting probability (see [Berezhkovskii and Szabo, 2004]), I tried to find it back. Since we use our two-states approximation:

$$\bar{p}_{J=2}(t) := p(0) + \sum_{i=1}^{N} (p_i(0)\psi_2^{(R)}(i))(\exp(\lambda_2 t) - 1)\psi_2^{(L)}$$

We consider only two-states like processes that have $\bar{p}_{J=2}(0)$ as a probability measure.

Notation:

1. $T_F = \inf_{t>0}\{X_t \in F\}$
2. $T_U = \inf_{t>0}\{X_t \in U\}$

We call Splitting Probability the following expression:

$$\sigma_i = \mathbb{P}(T_F < T_U | X(0) = i) = \mathbb{P}_i(T_F < T_U)$$

for $i \in S = \{1, \ldots, N\}$.

We have that:

$$\begin{aligned}
\sigma_i &= \int_0^\infty \mathbb{P}_i(T_F < T_U, T_U \in \mathrm{d}t) \\
&= \int_0^\infty \int_0^t \mathbb{P}_i(T_F < T_U, T_U \in \mathrm{d}t, T_F \in \mathrm{d}s) \\
&\quad + \underbrace{\int_0^\infty \int_0^t \mathbb{P}_i(T_F < T_U, T_U \in \mathrm{d}t, T_F \in \mathrm{d}s)}_{=0} \\
&= \int_0^\infty \int_0^t \underbrace{\mathbb{P}_i(T_F < T_U | T_U = t, T_F = s)}_{=1} \mathbb{P}_i(T_U \in \mathrm{d}t, T_F \in \mathrm{d}s) \\
&= \int_0^\infty \int_0^t \mathbb{P}_i(T_U \in \mathrm{d}t | T_F = s) \mathbb{P}_i(T_F \in \mathrm{d}s)
\end{aligned}$$

Notation:

(1)

$$f_{ij}(\mathrm{d}t) = \mathbb{P}_i(T_j \in \mathrm{d}t) = f_{ij}(t)\mathrm{d}t$$

for $i, j \in \{1, \ldots, N\}$

(2)

$$\begin{aligned}
p_{ij}(t) &= \mathbb{P}_i(X(t) = j) \\
&= \int_0^t \mathbb{P}_i(X(t) = j, T_j \in \mathrm{d}s) \\
&= \int_0^t \mathbb{P}_i(X(t) = j | T_j = s) \mathbb{P}_i(T_j \in \mathrm{d}s) \\
&= p_{ii} * f_{ij}(t)
\end{aligned}$$

(3)

$$\widetilde{p}_{ij}(\lambda) = \int_0^\infty p_{ij}(t) \exp(-\lambda t) \mathrm{d}t$$

Thus,

$$\sigma_i = \sum_{\ell \in U} \sum_{j \in F} \int_0^\infty \int_0^t f_{ij}(s) f_{j\ell}(t-s) \mathrm{d}s \mathrm{d}t$$

We get:

$$\sigma_i = \sum_{\ell \in U} \sum_{j \in F} \int_0^\infty f_{ij} * f_{j\ell}(t) \mathrm{d}t$$

$$\sigma_i = \sum_{\ell \in U} \sum_{j \in F} \int_0^\infty \mathcal{L}^{-1}\left(\frac{\widetilde{p}_{ij}}{\widetilde{p}_{ii}} \times \frac{\widetilde{p}_{j\ell}}{\widetilde{p}_{jj}}\right)(t) \mathrm{d}t$$

We inject $\bar{p}_{J=2}(t)$.

After resolution, we get that

$$\sigma_i = a - \psi_2^{(R)}(i)b$$

for $i \in S = \{1, \cdots, N\}$

with

$$a = \sum_{j\in S}\sum_{\ell\in S} \frac{\mu(\mu + \psi_2^{(R)}(\ell)\psi_2^{(R)}(j))}{(\mu + (\psi_2^{(R)}(j))^2)(\mu + (\psi_2^{(R)}(\ell))^2)}$$
$$+\frac{1}{\psi_2^{(R)}(j) + \psi_2^{(R)}(\ell)}\left(\frac{(\mu + (\psi_2^{(R)}(j))^2)(\psi_2^{(R)}(j))^3}{(\mu + (\psi_2^{(R)}(j))^2)^2} + \frac{(\psi_2^{(R)}(\ell))^3(\mu + (\psi_2^{(R)}(\ell))^2)}{(\mu + (\psi_2^{(R)}(\ell))^2)^2}\right)$$

and

$$b = \sum_{j\in S}\sum_{\ell\in S} \frac{(\psi_2^{(R)}(j))^2(\mu + (\psi_2^{(R)}(j))^2)}{(\mu + (\psi_2^{(R)}(j))^2)^2)(\psi_2^{(R)}(\ell) + \psi_2^{(R)}(j))}$$
$$- \frac{\psi_2^{(R)}(j)}{\mu + (\psi_2^{(R)}(\ell))^2}\left(\frac{\mu + \psi_2^{(R)}(j)\psi_2^{(R)}(\ell)}{\mu + (\psi_2^{(R)}(j))^2} - \frac{\psi_2^{(R)}(\ell)(\mu + (\psi_2^{(R)}(\ell))^2)}{\psi_2^{(R)}(\ell) + \psi_2^{(R)}(j)}\right)$$

and

$$\mu = \sum_{i=1}^{N} (\psi_2^{(R)}(i))^2 \pi_i$$

I could not obtain a simpler expression...

3.3 The Weibull Distribution

We consider the 2-Weibull Distribution (whose exponential distribution is taken from) whose probability density function is:

$$f(x;a,b) = \begin{cases} b/a(x/a)^{b-1}\exp(-(x/a)^b) & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{19}$$

And we consider the spacings s between adjacent eigenvalues (sorted in descending order):

$$s_i = |\lambda_i - \lambda_{i+1}| \tag{20}$$

for $i \in S = \{1, \ldots, N-1\}$.

Simulations show that our spacings are not iid, as we would have guessed since the elements of our band diagonal matrices are not iid (arbitrarily only the lower triangular ones are).

But, more interestingly, in the case of large number of microstates, the spacings in the bulk of the spectrum behave similarly, following very closely a 2-Weibull Distribution so that we have:

$$\mathbb{P}(s_i) \approx b/a(s_i/a)^{b-1}\exp(-(s_i/a)^b) \;\; \forall i \in \;\; bulk. \tag{21}$$

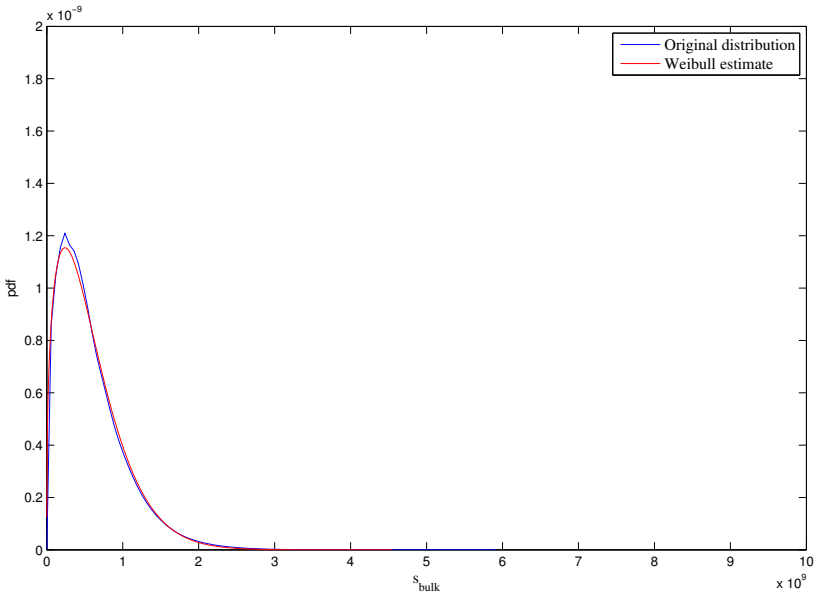

Figure 13: Density function of a spacing in the bulk of the spectrum, and its weibull estimate.

Work can be done to link this result with already existing random matrix theory. Indeed, it would be interesting to compare the spectral behavior of our matrices (non-symmetric band-diagonal real matrices, even if self-adjoint on the hilbert space $\mathcal{L}_2(\pi) \subset \mathbb{R}^N$) with the one of symmetric real matrices. For some interesting info see [Liu, 2000] and [Timm, 2009].

4 Discrete Time, General State Space

At the end of my internship, I tried to draw a parallel bewteen the spectral properties found for CTMC or DTMC on finite state space and the ones of DTMC on the general state space, which are widely used in MCMC algorithms (see for example [Rosenthal, 2003]). Unfortunately I didn't have the time to find promising results. I will nonetheless give a little overview of markov operator on general state space (in discrete time).

Let $\{X_n, n \in \mathbb{N}\}$ be a stochastic process defined on the general state space $(\mathbb{R}, \mathcal{B})$ that evolves as follows:

$$\forall x_n \in \mathbb{N},\ \ A \in \mathcal{B},\ \ \mathbb{P}(X_{n+1} \in A | X_0 = x_0, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} \in A | X_n = x_n) = \mathbb{P}(x_n, A).$$

Under this assumption $\{X_n, n \in \mathbb{N}\}$ is a DTMC on general state space and is characterized by:

(1) an initial distribution p_0 on $(\mathbb{R}, \mathcal{B})$

(2) A conditional probability distribution $P(x, \cdot)$ on $(\mathbb{R}, \mathcal{B})$ determined by a function $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}+$ satisfying:

- for all $x \in \mathbb{R}$
$$P(x, \mathbb{R}) = \int_{\mathbb{R}} \rho(x, \mathrm{d}y) = 1$$
- for all $x \in \mathbb{R},\ A \in \mathcal{B}$
$$P(x, A) = \int_{A} \rho(x, \mathrm{d}y) \in (0, 1)$$

We consider the Hilbert space $\mathcal{L}_2(\pi)$, space of π-measurable functions where π is a probability measure on the measurable space $(\mathbb{R}, \mathcal{B})$:

$$\mathcal{L}_2(\pi) = \{f : \mathbb{R} \to \mathbb{R}, \int \pi(\mathrm{d}x) f(x)^2 < \infty\}$$

$P(x, \cdot)$ is an operator that acts:

1. on the set of probability measure on $(\mathbb{R}, \mathcal{B})$:
$$P : \left| \begin{array}{rcl} \mathbb{M}(\mathbb{R}, \mathcal{B}) & \longrightarrow & \mathbb{M}(\mathbb{R}, \mathcal{B}) \\ \mu & \longmapsto & \mu P = \int_{\mathbb{R}} \mu(\mathrm{d}x) P(x, \cdot) \end{array} \right.$$

2. on $\mathcal{L}_2(\pi)$:
$$P : \left| \begin{array}{rcl} \mathcal{L}_2(\pi) & \longrightarrow & \mathcal{L}_2(\pi) \\ f & \longmapsto & Pf = \int_{\mathbb{R}} P(\cdot, \mathrm{d}x) f(x) \end{array} \right.$$

$\mathcal{L}_2(\pi)$ is equipped with the scalar product:

$$\langle f, g \rangle = \int \pi(\mathrm{d}x) f(x) g(x),\ \ \forall (f, g) \in (\mathcal{L}_2(\pi))^2$$

and the related norm:

$$||f||^2 = \pi(\mathrm{d}x) f(x)^2$$

We define the operator norm as:

$$||P|| = \sup_{f \in \mathcal{L}_2(\pi), ||f|| > 0} \frac{||Pf||}{||f||}$$

Jensen's inequality tells us that, for all $f \in \mathcal{L}_2$,

$$(Pf)^2 = (\int P(\cdot, \mathrm{d}x) f(x))^2 \leq \int P(\cdot, \mathrm{d}x) f(x)^2$$

Thus by π invariance,

$$||Pf||^2 \leq ||f||^2$$

So, we have that $||P|| \leq 1$. By the same reasoning that we used for DTMC on discrete state space we finally find that $||P|| = 1$ with eigenvalue $\lambda = 1$ and eigenfunction $f_1 = 1$.

We assume that $\{X_n, n \in \mathbb{N}\}$ is π-reversible, for all $A, B \in \mathcal{B}$

$$\int_A \pi(\mathrm{d}x) P(x, B) = \int_B \pi(\mathrm{d}x) P(x, A)$$

We have the following theorem:

Theorem 5. *P π-reversible is equivalent to P self-adjoint*

Proof. for all $(f, g) \in (\mathcal{L}_2(\pi))^2$,

$$\langle Pf, g \rangle = \int \int \pi(\mathrm{d}x) P(x, \mathrm{d}y) f(y) g(x) = \int \int \pi(\mathrm{d}x) P(x, \mathrm{d}y) f(x) g(y) = \langle f, Pg \rangle$$

□

The spectrum of P is defined as

$$Sp(P) := \{\lambda \in \mathbb{R}, \exists f, Pf = \lambda f\}$$

We have that $Sp(P) \in [-1, 1]$, indeed let $\lambda_0 \in Sp(P)$. There exists eigenfunction $f_0 \in \mathcal{L}_2(\pi)$ such that

$$\langle Pf_0, Pf_0 \rangle = \lambda_0^2 ||f_0||^2$$

So,

$$|\lambda_0| = \frac{||Pf_0||}{||f_0||} \leq 1$$

I did not study the spectral decomposition of $\{X_n, n \in \mathbb{N}\}$, thus much remains to be done!

5 Conclusion

The spectral analysis of the continuous time markov process $\{X(t), t > 0\}$ allows to derive reliable statistics giving a good insight on the process of protein folding.

The spectral decomposition of $\{X(t), t > 0\}$ allows indeed an interpretation of the chain dynamics in terms of flux of probability mass between high density regions.

Simulations tend to show that the presence of a spectral gap (with $\gamma \geq 7$) between the second and third eigenvalue of the transition rate matrix K support a two-states approximation, as well as the importance of the structure of K on the distribution of γ.

However, a better gap won't necessarily yield a better two-states approximation.

Likewise, the two-state approximation of the time evolution of $\{X(t), t > 0\}$ we propose is not always valid.

Work could be done about knowing whether a higher order representation (3 or 4 macrostates for example) of a protein dynamics is possible, as well as pursuing the study of the parallels between spectral analysis of markov processes on discrete and general state space.

References

[Berezhkovskii and Szabo, 2004] Berezhkovskii, A. and Szabo, A. (2004). Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. *The Journal of chemical physics*, 121(18):9186–9187.

[Buchete and Hummer, 2008] Buchete, N.-V. and Hummer, G. (2008). Coarse master equations for peptide folding dynamics. *The Journal of Physical Chemistry B*, 112(19):6057–6069.

[Buchner et al., 2011] Buchner, G. S., Murphy, R. D., Buchete, N.-V., and Kubelka, J. (2011). Dynamics of protein folding: Probing the kinetic network of folding–unfolding transitions with experiment and theory. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1814(8):1001–1020.

[Liu, 2000] Liu, Y. (2000). Statistical behavior of the eigenvalues of random matrices.

[Prinz et al., 2011] Prinz, J.-H., Keller, B., and Noé, F. (2011). Probing molecular kinetics with markov models: metastable states, transition pathways and spectroscopic observables. *Physical Chemistry Chemical Physics*, 13(38):16912–16927.

[Rosenthal, 2003] Rosenthal, J. S. (2003). Asymptotic variance and convergence rates of nearly-periodic markov chain monte carlo algorithms. *Journal of the American Statistical Association*, 98(461):169–177.

[Timm, 2009] Timm, C. (2009). Random transition-rate matrices for the master equation. *Physical Review E*, 80(2):021140.