# Classification of discrete pathways based on the fundamental sequence

Ernesto Suárez and Daniel Zuckerman

June 15, 2017

How can ensembles of transition trajectories be compared on an 'apples-to-apples' basis? One simple approach is trajectory classification - mapping each trajectory of an ensemble into a bin, and generating a histogram with error bars that can be directly compared to other histograms made in the same way from other ensembles. However, we are not aware of established techniques for binning trajectories, which motivates the methods described here.

The concept of the 'fundamental sequence' (FS) will help us to do a meaningful binning/classification. Roughly speaking, the FS of a discrete path is what remains after eliminating the "loops" in the path. A loop is any sub-segment of the path where the initial and final states are the same. Even self transitions –when the systems stays in the same state after $\tau$– can be considered loops of size zero. If two paths share the same FS, then they will belong to the same class. There is not, however, a unique way to remove the loops and we need an unambiguous definition for FS.

Imaging drawing a graph $G$ from each observed (discrete) path $\mathbf{s} = (s_1, s_2, ..., s_m)$ from $A$ to $B$ ($s_1 \in A$, $s_m \in B$), where the nodes of the graph would be all the $N$ states. The graph is built following one single rule; if there is one or more *direct* transitions from $m$ to $l$ in $\mathbf{s}$ (i.e., $\exists i : s_i = m$ and $s_{i+1} = l$ ), a *directed* edge $m \to l$ is added to $G$. At the end, some edges will have a single direction while others will be bidirectional and there will also be in general isolated nodes.

The FS of each path $\mathbf{s}$ is the sequence that maximizes the product of the transition probabilities through all possible paths $\mathbf{q}$ consistent with $G$ ($\Gamma(G)$), that is

$$\mathrm{FS} = \underset{\mathbf{q}\in\Gamma(G)}{\arg\max} \prod_{i=1}^{|\mathbf{q}|-1} k_{q_i,q_{i+1}}. \tag{1}$$

where $|\mathbf{q}|$ is the number of elements in the path $\mathbf{q} \in \Gamma(G)$, being $|\mathbf{q}| - 1$ the number of transitions.

An equivalent and more convenient formulation would be

$$\mathrm{FS} = \underset{\mathbf{q}\in\Gamma(G)}{\arg\min} \left\{ \sum_{i=1}^{|\mathbf{q}|-1} -\log k_{q_i,q_{i+1}} \right\}, \tag{2}$$

where have transformed the problem of maximizing the product of the transition probabilities to the classical problem of finding the shortest distance in a graph, in this case, using the pseudo distance $\delta(m,l) = -\log k_{ml}$. There are well known strategies to solve this kind of problem, here we are going to use the Dijkstra's algorithm.

As an example suppose we have a very simple model with only four states (Fig. 1); a initial state $A$, a final state $B$, and two intermediate states $I1$ and $I2$ that connect $A$ and $B$ as shown in Fig. 1a. In general, any possible path from $A$ to $B$ in this model will correspond to one and only one of the four *fundamental sequences* drawn in Fig. 1b. Fig. 1c are examples of paths that would correspond, respectively, with the fundamental sequences in Fig. 1b.

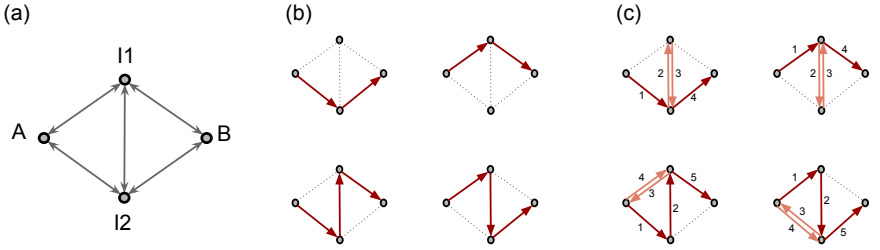


Figure 1: Simple kinetic model. (a) Kinetic model with four states, the arrows are drawn between states kinetically connected. (b) Possible mechanisms from $A$ to $B$. (c) Examples of additional paths with "unproductive" steps

.

For the description of the mechanism, and in particular for equilibrium ensembles, it is more convenient a partition of the path space symmetric by construction. That is, the distribution of the folding paths over the classes should be equal to the distribution of the unfolding paths on the same space. Instead of maximize the likelihood of the sequence given the connectivity of the path, we will maximize the likelihood of the "round trip". The symmetrized fundamental sequence FS$^*$ is defined as

$$\mathrm{FS}^* = \underset{\mathbf{q}\in\Gamma(G)}{\arg\min} \left\{ \sum_{i=1}^{|\mathbf{q}|-1} -\log(k_{q_i,q_{i+1}} k_{q_{i+1},q_i}) \right\}. \tag{3}$$

By default in equilibrium ensembles we are are going to use the symmetrized version. Notice that in this case all the edges in $G$ are considered bidirectional with distance $\delta(m,l) = -\log(k_{ml}k_{lm})$, even when there are directional edges in $G$.