6. Detecting Reaction Pathways via Shortest Paths in Graphs

We have demonstrated that discrete TPT is a powerful tool to analyze transition events in Markov jump processes. The central object is the infinitesimal generator of the process which finally allows to compute reaction rates and to determine a hierarchy of dominant reaction pathways. In the case where the process is only discretely observed in time the generator of the underlying Markov jump process has to be estimated. Alternatively, if the observation time lag is constant, the underlying process can be modeled as a Markov chain, i.e., as a Markov process discrete in space and time.

In this chapter, we will consider the use of shortest-path algorithms in the context of reaction pathway computation in Markov chains. The crucial question for this undertaking is the choice of a weight function that defines the *length* of an edge. We will present two such functions [22] which both have a natural motivation. We will apply both resulting methods on the examples which have already been investigated with discrete TPT in Section 4.3 and we will compare the results.

6.1. Shortest Path in Graphs

6.1.1. Dijkstra Algorithm

The standard algorithm used for computing shortest paths in a graph $G = G(V, E)$ is the Dijkstra algorithm. It solves the so called *Single Source Shortest Path Problem* where the shortest paths from one source vertex $v_s \in V$ to all other vertices $v \in V$ have to be determined. The *Single Source, Single Destination Shortest Path Problem* is a special case in which only one path from v_s to a designated destination vertex v_d has to be determined. In both cases the runtime of the Dijkstra algorithm is $O(|V| \log(|V|) + |E|)$. For a profound discussion of this standard algorithm we refer to, e.g. [54, 18]. In the following we will only roughly sketch its basic principle.

Given a vertex v_s as starting vertex, the algorithm maintains a list of distances to v_s assigned to every other vertex that is initialized with the value ∞ and in the end contains the lengths of the shortest paths from v_s to any vertex. In the first step, the distances of all neighbors of v_s are set to the weight of the edge connecting them to v_s. These vertices form the initial *halo set*, i.e. they are the vertices for which one path from v_s is known but it is not known whether this path is a shortest path. In the main loop of the algorithm, it removes a vertex v_{min} with the minimum known distance from the halo set, and considers all neighbors of v_{min}. If a neighbor is also in the halo set, the algorithm checks whether a path through v_{min} would result in a distance from v_s less than the current known distance. If a neighbor is not yet in the halo set, it is added to it, with its distance value being the sum of the distance of v_{min} and the length of the edge connecting the neighbor to v_{min}. The algorithm

terminates when a prescribed target vertex is reached or when the halo set becomes empty.

By a slight modification, the Dijkstra algorithm can be generalized to find a shortest path from any vertex of a source set $V_s \subset V$ to all remaining vertices $w \in V \setminus V_s$. Unlike in the original algorithm, here the source set V_s forms the initial halo set and in the initialization step all vertices of V_s are assigned the distance value zero.

Algorithm 7 Generalized Dijkstra's Algorithm

Input: A directed graph $G = (V, E)$, weights $w : E \to \mathbb{R}_+$. Source set $V_s \subset V$
Output: Shortest paths from all $v \in V_s$ to all $u \in V$ and their lengths.

(1) Set $\mathrm{dist}(v_s) = 0 \ \ \forall v_s \in V_s$, $\mathrm{dist}(v) := \infty \ \ \forall v \in V \setminus V_s$.
(2) Initialize halo set $H := V_s$.
(3) $v := \mathrm{argmin}_{u \in H} \mathrm{dist}(u)$, set $H := H \setminus \{v\}$.
(4) **FOR ALL** $(v, u) \in E$ **DO**:
(5) **IF** $\mathrm{dist}(u) > \mathrm{dist}(v) + w(v, u)$
(6) **THEN** Set $\mathrm{dist}(u) := \mathrm{dist}(v) + w(v, u)$, $\mathrm{pred}(u) := v$.
(7) **IF** $u \notin H$ **THEN** $H := H \cup \{u\}$.
(8) **END FOR**
(9) **IF** $H \neq \emptyset$ **THEN** go to step (3).

If a vertex $v \in V \setminus V_s$ is not reachable from the source set V_s then $\mathrm{dist}(v) = \infty$ and $\mathrm{pred}(v)$ is not defined. Otherwise, the shortest path from V_s to a vertex v can be reconstructed by following recursively the predecessors until a vertex in $v_s \in V_s$ is reached. If we subsequently reverse the order of the vertices in that path, we end up with a shortest path $z(v)$ from V_s to v,

$$z(v) = (v_s, \ldots, \mathrm{pred}(\mathrm{pred}(v)), \mathrm{pred}(v), v).$$

6.1.2. Bidirectional Dijkstra Algorithm

The purely graph theoretic consideration of shortest paths in the previous section has to be extended by some ideas related to the specialized setting of graphs describing spatial discretizations of Markov diffusion processes. In particular we have in mind the fact that the numerical realizations of these graphs necessarily come with a discretization error which makes it doubtful whether the notion of *the* shortest path between two vertices v_s and v_d is really a meaningful quantity in our applications – even leaving out the possible existence of several shortest paths. Furthermore, we are interested in the revealing of *all* dominant transition channels between metastable sets. Therefore, we are not only interested in one (or all) precisely shortest paths, but we are also interested in a *family of shortest paths* which consists of short paths being only slightly longer than a path with the shortest length.

Edge based family of shortest paths Let $\mathrm{dist}(v_s, v_d)$ denote the length of the shortest path between the vertex $v_s \in V_s$ and the vertex $v_d \in V_d$. We want to calculate all paths from v_s to v_d which have a slightly longer length than the shortest path. To be more precise, we determine all paths in the graph which have a length

of at most

$$(1+\epsilon)\text{dist}(v_s, v_d),$$

where $0 \leq \epsilon$. In order to do so, we need to apply the generalized Dijkstra algorithm, as stated in Algorithm 7, only twice: Firstly, we calculate all distances from v_s to all other vertices and denote these distances by $\text{dist}_1(v)$ for all vertices $v \in V$. Among all distances this also includes the distance between v_s and v_d. Secondly, we consider a new graph $= (V, E^R)$ where E^R consists of the edges in E with direction reversed. Then, we calculate all distances from v_d to all other vertices in G^R, and denote these distances by $\text{dist}_2(v)$ for all vertices $v \in V$. Note that $\text{dist}_2(v)$ is also the distance from v to v_d in G for any vertex $v \in V$.

It is now simple to decide whether or not an edge $(v_i, v_j) \in E$ lies on a path between v_s and v_d of length at most $(1+\epsilon)\text{dist}(v_s, v_d)$. Such a path has to consist of three parts: a path from v_s to v_i, the edge (v_i, v_j) itself and a path from v_j to v_d. The shortest length for the first part is $\text{dist}_1(v_i)$ and the shortest length of the last part is $\text{dist}_2(v_j)$. Thus, an edge (v_i, v_j) lies on a path between v_s and v_d of length at most $(1+\epsilon)\text{dist}(v_s, v_d)$ if and only if

$$\text{dist}_1(v_i) + w(v_i, v_j) + \text{dist}_2(v_j) \leq (1+\epsilon)\text{dist}(v_s, v_d).$$

The result is a subset $E_{sp}(\epsilon) \subset E$ of edges belonging to the family of short paths. The algorithm can easily be extended to the case of more than one source and destination vertex and is stated in Algorithm 8. The computational cost of Algorithm 8 is the

Algorithm 8 Bidirectional Dijkstra algorithm

Input: A directed graph $G = (V, E)$, weights $w : E \to \mathbb{R}_+$, source set $V_s \subset V$ and destination set $V_d \subset V$, threshold ϵ.

Output: Set of edges $E_{sp}(\epsilon) \subset E$ belonging to the family of short paths.

(1) Compute all distances $\text{dist}_1(v)$ in G from V_s to all vertices $v \in V$.
(2) Construct new graph $G^R = (V, E^R)$ by reversing all edges in E.
(3) Compute all distances $\text{dist}_2(v)$ in G^R from V_d to all vertices in $v \in V(G^R)$.
(4) $\text{dist}_{min} := \min_{v_d \in V_d}\{\text{dist}_1(v_d)\}$.
(5) **FOR ALL** edges in $(v_i, v_j) \in E$ **DO**
(6) **IF** $\text{dist}_1(v_i) + w(v_i, v_j) + \text{dist}_2(v_j) \leq (1+\epsilon)\text{dist}_{min}$
(7) **THEN** $E_{sp}(\epsilon) := E_{sp}(\epsilon) \cup \{(v_i, v_j)\}$.
(8) **END FOR**

same as of Dijkstra's algorithm (cf. Algorithm 7).

Vertex based family of shortest paths In order to motivate the vertex based approach, consider for an edge $(v_i, v_j) \in E_{sp}(\epsilon)$ the associated shortest path

$$z_S(v_i, v_j) = \left(z(v_i), z^R(v_j)\right),$$

where $z^R(v_j)$ denotes the shortest path from v_j to V_d in G. The crucial observation is now that in general the path $z(v_i, v_j)$ is different from the shortest path composed of the shortest path from V_s to v_i and the shortest path from v_i to V_d. In other words, the edge (v_i, v_j) does not have to lie on the shortest path connecting V_s and V_d via v_i. The same argument holds for the vertex v_j too.

For this reason, we consider in the vertex based approach the vertex set $V_{sp}(\epsilon) \subset V$ which comprises all vertices being involved on a short path with length at most $(1+\epsilon)\text{dist}_{min}$. Formally, a vertex $v \in V$ lies on a short path between V_s and V_d of length of at most $(1+\epsilon)\text{dist}_{min}$ if and only if

$$\text{dist}_1(v) + \text{dist}_2(v) < (1+\epsilon)\text{dist}_{min}.$$

Reaction pathways and short paths It immediately follows from the definition of a shortest path that it does not have any self-intersections (loops). Unfortunately, both presented approaches - the edge and vertex based approach - lead to short paths which could have self-intersections. To see this, notice that in both approaches a short path is composed of two shortest paths which are computed independently of each other. Since we are eventually interested in reaction pathways (cf. Def. 4.1.15), we finally have to sort out from the family of short paths all short paths with loops.

6.2. Choice of Edge Weights

The choice of the edge weights depends on the way how to compare paths. The first natural choice of edge weights is based on the *likelihood* of sample paths. Unfortunately, this approach leads in our context to reaction pathways which are in contradiction to the underlying dynamics and its physical interpretation. As a remedy, we follow an alternative approach which takes into account the *free energy* barriers along a reaction pathway. This approach is more adapted to the underlying dynamics and, moreover, is less sensitive with respect to the underlying box discretization of the state space. In what follows we interpret a Markov chain, described by its transition matrix $P = (p_{ij})_{i,j\in S}$, as a directed graph $G = (V, E)$ (cf. notations introduced in Section 4.1.1).

6.2.1. Likelihood Approach

Suppose we are given two sample paths p_1 and p_2 of a Markov chain and, moreover, suppose that both starting in a state i_A and ending in the state i_B. One option to compare these two paths is to ask which of both is the more preferred one by the dynamics, i.e. which one is more *likely*? The respective likelihoods of the sample paths can be expressed in terms of the transition probabilities of the Markov chain. Let $z = (i_1, i_2, i_3, \ldots, i_n)$ be a finite sample path of the Markov chain. Then the likelihood of z is given by

$$\mathcal{L}_d(z) = \prod_{k=1}^{n-1} p_{i_k, i_{k+1}},$$

being the probability that the Markov chain produces that sample path conditional on starting in the state i_1. The edge weights are now chosen such that the more likely a sample paths is the less is its length. To this end, we define the weight of an edge $(i, j) \in E$ by

$$w_L(i,j) \stackrel{def}{=} -\log(p_{ij}) \tag{6.1}$$

and the resulting length of a path p,

$$l(z) = \sum_{k=1}^{n-1} w_L(i_k, i_{k+1}) = -\log(\prod_{k=1}^{n-1} p_{i_k,i_{k+1}}),$$

is the negative *log-likelihood* of that sample path (cf. Sect. 5.2). It should be clear that the shortest path with respect to the weights in (6.1) between two different states is the most probable one and vice versa.

6.2.2. Free Energy Approach

Motivational Example To motivate the alternative free energy approach we present an example for which the likelihood approach yields misleading results. To this end we consider a pure diffusion process, i.e. a diffusion process in a flat potential landscape, on a squared domain with reflecting boundary conditions. Additionally, we cut out of the domain a small square and apply reflecting boundary conditions on its boundary too. In Figure 6.1 we give a schematic picture of the situation. The decomposition of the domain is chosen such that the probability to encounter the equilibrated process in the boxes $A, 1, 2, 3$ and B is $1/8$ and for the box 4 is $3/8$. The dynamics between the boxes is given by a reversible Markov chain where the transition probabilities between adjacent boxes are given by

$$\begin{aligned} p(A,1) &= p(1,A) = p(1,2) = p(2,1) \\ &= P(2,3) = p(3,2) = p(3,B) = p(B,3) \\ &= p(A,4) = p(B,4) = a,\ 0 < a < \frac{1}{2}, \\ p(4,A) &= p(4,B) = \frac{1}{3}a, \end{aligned}$$

where the probability a depends on the size of the squares. The condition $0 < a < \frac{1}{2}$ guarantees that the probability to make a self-transition is positive.

Suppose, we start the diffusion process in box A. From the symmetry of the domain and the nature of diffusion, it should be clear that the probability to reach the box B via the upper way is the same, namely $\frac{1}{2}$, as for the lower way. The discrete likelihoods of the upper and lower way in the Markov chain are given by

$$\begin{aligned} \mathcal{L}_d\big((A,1,2,3,B)\big) &= a^4, \\ \mathcal{L}_d\big((A,4,B)\big) &= \frac{1}{3}a^2. \end{aligned}$$

The likelihoods of both discrete paths are equal if and only if $a = \sqrt{1/3} \approx 0.577 > 0.5$ which contradicts the condition $0 < a < 0.5$. But this means that the likelihood approach would indicate that one of the paths is preferred which is not consistent with the continuous picture.

New Choice of Edge Weight

The new choice of edge weights is based on the discrete analog of the free energy (cf. Sect. A.5).

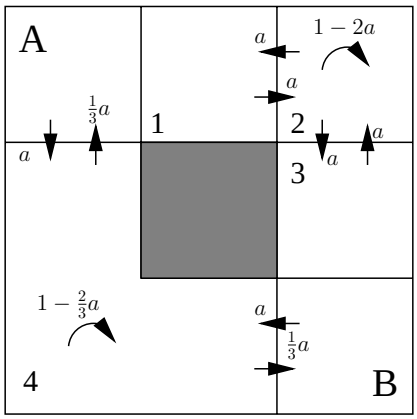

Figure 6.1.: Schematic representation of the motivational example.

Recall that under the assumption that the Markov chain admits a unique stationary distribution $\pi = (\pi_i)_{i \in S}$ the discrete free energy $F = (F_i)_{i \in S}$ is defined by

$$F_i \overset{def}{=} -\log \pi_i > 0, \quad i \in S.$$

The new edge weights are constructed such that the shortest path between two states is the one which overcomes the lowest discrete free energy barriers. To this end, we introduce the weights

$$w_F(i,j) = |F_j - F_i|. \tag{6.2}$$

In order to explain that the free energy weights yield the desired result, firstly consider a path $z = (i_1, \ldots, i_s)$ with monotonously increasing discrete free energies along it, i.e.,

$$F_{i_j} \leq F_{i_{j+1}} \Leftrightarrow \pi_{i_j} \geq \pi_{i_{j+1}}, \quad j = 1, \ldots, s-1. \tag{6.3}$$

Then the length of such a path z,

$$l(p) = \sum_{j=1}^{s-1} w_F(i_j, i_{j+1}) = F_{i_s} - F_{i_1},$$

is simply given by the free energy difference between the last and the first state of the path. Moreover, if we fix the states i_1 and i_s, then all paths connecting these two states and satisfying (6.3), have the same length. Next, consider a path $z = (i_1, \ldots, i_n)$ which can be decomposed into two parts $z_1 = (i_1, \ldots, i_s)$ and $z_2 = (i_s, \ldots, i_n)$ such that

$$\begin{cases} F_{i_j} \leq F_{i_{j+1}}, & j = 1, \ldots, s-1, \\ F_{i_j} \geq F_{i_{j+1}}, & j = s, \ldots, n-1. \end{cases}$$

One immediately verifies that again the length of the path z,

$$l(z) = 2F_{i_s} - (F_{i_1} + F_{i_n}) \geq 0,$$

only depends on free energy differences, namely the barriers $F_{i_s} - F_{i_1}$ and $F_{i_s} - F_{i_n}$. Consequently, if we fix the states i_1 and i_n then the shortest path between i_1 and i_n with respect to the weights in (6.2) is the one which crosses the lowest free energy barriers.

Remark 6.2.1. *Notice that in the free energy approach every barrier along a certain path contributes to the length twofold: first from V_s towards V_d and second from V_d towards V_s. However, in a post-processing step one can determine all barriers along a reaction pathway and use these information for a further analysis or a re-computation of the shortest paths. Notice further that in our motivational example the free energy approach would also tell that the lower discrete path is the preferred path of the underlying diffusion. Nevertheless, our numerical experiments will show that the free energy approach is more insensitive with respect to a decomposition of a diffusion process.*

6.3. Numerical Experiments

In this section we will illustrate the method of detecting reaction pathways via shortest paths in graphs on some examples which have already been investigated via discrete TPT. In the following, we only use the vertex based approach.

In the first example we consider the Smoluchowski dynamics in the three-hole example introduced in Section 3.7.1. In order to start with the shortest paths method, we have to provide a transition matrix which captures the dynamics on a coarse grained level. Instead of generating such a transition matrix via a time series resulting from a direct numerical simulation of the Smoluchowski dynamics in (3.22), we utilized the generator L, given in (4.44), of the approximating Birth-Death process (cf. Sect. 4.3.1). For the time lag $\tau = 1.2$ we generated a transition matrix $P(\tau)$ via the relation

$$P(\tau) = \exp(\tau L).$$

As the source set $A(= V_s)$, we chose the states (mesh points) in S which cover the left deep minima. The destination set $B(= V_d)$ was chosen with respect to the right minima. The stationary distribution $\pi = (\pi_i)_{i \in S}$ was numerically computed via $P(\tau)^T \pi = \pi$. In the following results, we always used the vertex based approach.

For the comparison of the likelihood approach and the free energy approach, we asked two questions:

1. Do both approaches detect the two transition channels?
2. Do both approaches reproduce the entropic switching behavior?

The Figure 6.2 gives an answer to the first question. In the left column we plot the families of reaction pathways resulting from the likelihood approach. From top to bottom we used the thresholds $\epsilon = 0.1, \epsilon = 0.3$ and $\epsilon = 0.6$. The right column shows the families of reaction pathways resulting from the free energy approach for the thresholds $\epsilon = 0.05$ (top),$\epsilon = 0.06$ and $\epsilon = 0.08$ (bottom). The darker the color of a drawn edge is the shorter is the path in which the edge is involved, i.e, the more likely is the path in the likelihood approach and the lower is the overcome barrier in the free energy approach, respectively. One can clearly see, that with increasing threshold the likelihood approach results rather in reaction pathways which cross the local maximum than in pathways proceeding from A to B in the upper channel. This behavior is in contradiction to the underlying diffusion and can be ascribed to the sensitivity of the likelihood approach with respect to the underlying

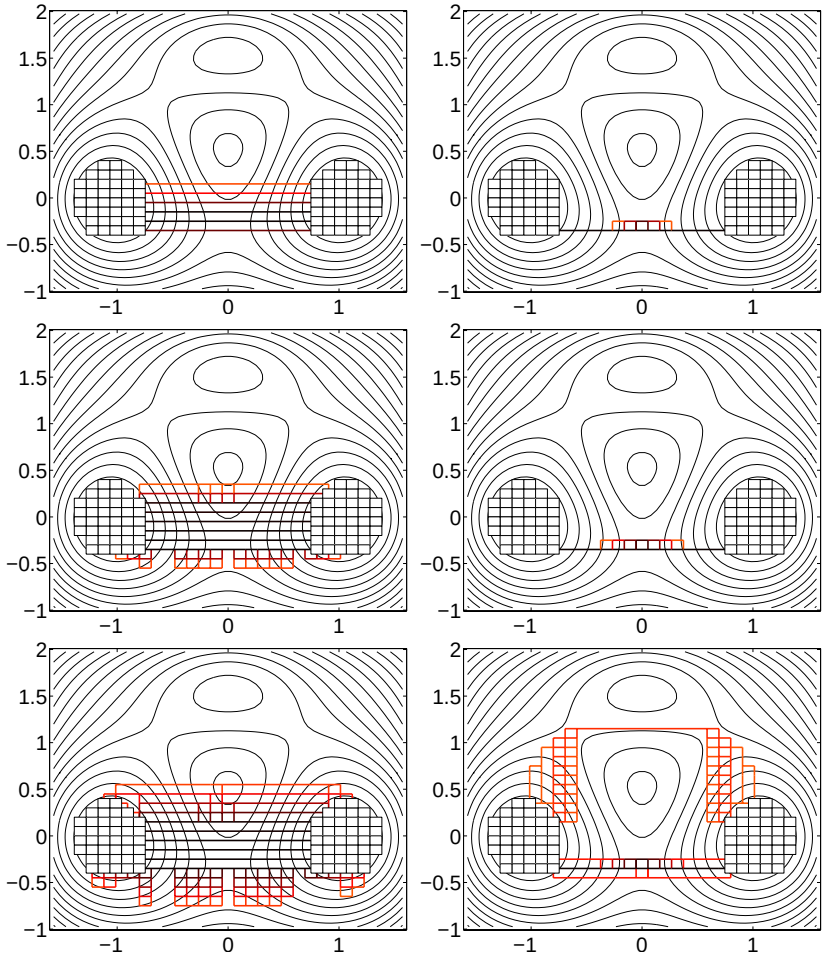

Figure 6.2.: Comparison of vertex based families of reaction pathways resulting from the likelihood approach (left column) and from the free energy approach (right column). The darker the color of an edge the shorter is the pathway in which the edge is involved. For the likelihood approach we chose from top to the bottom the threshold $\epsilon = 0.1, \epsilon = 0.3$ and $\epsilon = 0.6$. For the free energy approach we chose $\epsilon = 0.05, \epsilon = 0.06$ and $\epsilon = 0.08$. The sets A and B (depicted by boxes) were chosen such that they cover the two local minima, respectively. Results for a 30×30 mesh discretization of the rectangular state space and temperature $\beta = 1.67$.

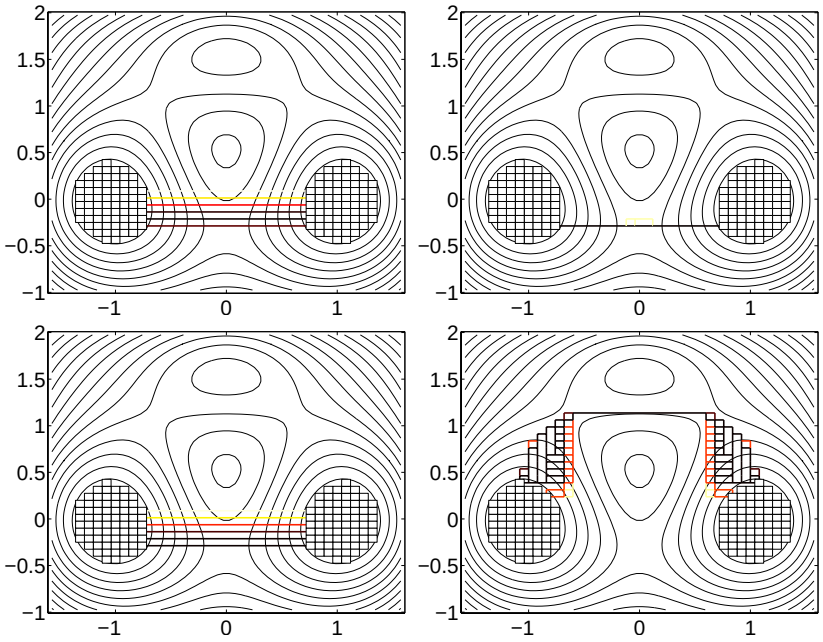

Figure 6.3.: Comparison of reaction pathways resulting from the likelihood approach (left column, $\epsilon = 0.1$) and from the free energy approach (right column, $\epsilon = 0.01$) at two different temperatures: $\beta = 6.67$ (top row) and $\beta = 1.67$ (bottom row). Results for a 40×40 mesh discretization of the rectangular state space.

discretization. As expected, the families of reaction pathways resulting from the free energy approach are less sensitive to the respective discretization because for the low threshold $\epsilon = 0.08$ both channels were detected.

The Figure 6.3 reveals the behavior of both approaches if the temperature in the underlying diffusion is varied. In the first column we show the reaction pathways ($\epsilon = 0.1$) resulting from the likelihood approach for the low temperature $\beta = 6.67$ (top) and the high temperature $\beta = 1.67$ (bottom). The right column shows the results of the free energy approach ($\epsilon = 0.01$) for the same temperatures. Apparently, only the free energy approach reproduces an entropic switching behavior of the underlying dynamics. But if we compare the preferred channels with the channels obtained via the discrete TPT (cf. Fig. 4.7 in Section 4.3.1) we see that the results are in opposite to each other; e.g., the preferred channel resulting from the shortest path methods at $\beta = 6.67$ corresponds to the preferred channel resulting from discrete TPT at $\beta = 1.67$. This observation can be explained by recalling that the length of a path in the free energy approach reflects barriers which the path overcomes. At low temperature ($\beta = 6.67$) the only chance to encounter the underlying process is in one of the minima. Hence, the upper shallow minima is separated from the rest by a extremely high free energy barrier and, hence, the lower direct channel is detected.

We end this section by presenting the results of the free energy approach applied to two others examples which have been investigated with discrete TPT in Section 4.3. The reaction pathways for the genetic toggle switch example (cf. Sect. 4.3.3 extracted from the vertex based family of short paths ($\epsilon = 0.005$) are depicted in the left panel

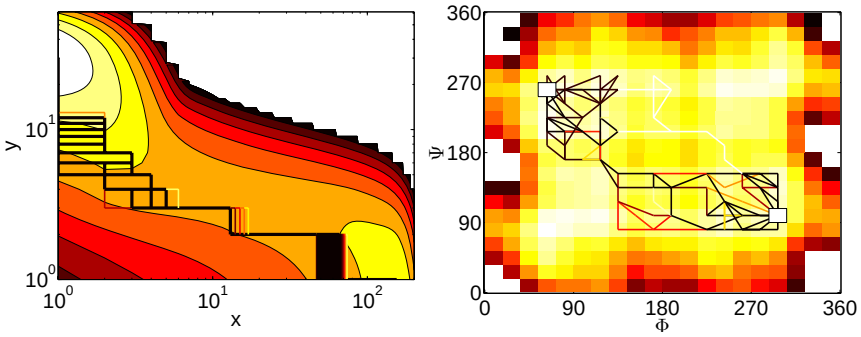

Figure 6.4.: Left: Reaction pathways for the genetic toggle switch example (cf. Fig. 4.20 in Section 4.3.3) extracted from the vertex based family of short paths ($\epsilon = 0.005$) in the free energy approach. Right: Reaction pathways in the torsion angles space of the glycine in solvent example (cf. Sect. 4.3.2) extracted from the vertex based family of short paths ($\epsilon = 0.01$) in the free energy approach. The darker the color of an edge is the shorter is the length of the pathway in which the edge is involved, i.e., the lower is the overcome free energy barrier of the pathways. The reaction pathways are embedded in a log-log contour plot of the discrete free energy, respectively.

of Figure 6.4. Apparently, the reaction pathway crossing the lowest discrete free energy barrier is consistent with dominant reaction pathways found via discrete TPT (cf. Fig. 4.20).

The right panel of Figure 6.4 illustrates the reaction pathways in the torsion angles space of the glycine in solvent example (cf. Sect. 4.3.2) extracted from the family of short paths ($\epsilon = 0.01$). The detected reaction pathways suggest that the lower channel is the preferred one which stands in contradiction to the results found via discrete TPT (cf. Fig. 4.13).

Finally, we draw the conclusion that only the free energy approach is able to detect different transition channels which are consistent with the underlying diffusion. But in the case of multiple reaction channels it does not allow to make any predictions about which channel is preferred one by the underlying diffusion. Nevertheless, the shortest-path approach is useful to get a first impression of possible reaction channels.