# 1. Introduction

Transition events in complex systems between long lived states are a key feature of many systems arising in physics, chemistry, biology, etc. It was early recognized that transition processes are characterized by rare but important events, i.e., transition processes are phenomena that take place on a long time scale compared to the time scale characterizing the states of local stability, also called *metastable* states. For example, the timescale for folding of a small protein, i.e. the transition from an unfolded in a folded state is in the range of microseconds to milliseconds, whereas that for small-amplitude motions of amino acid side chains and water solvent is 1 femtosecond.

The first step towards an understanding of rare events was to realize that escape from a metastable state can only happen via noise-assisted hopping events where the amplitude of the noise reflects the finite temperature at which the process takes place. In other words, the dynamics of the process is subject to random perturbations. If we relate the fluctuation induced by the noise to an appropriate energy scale $E_{noise}$, escape from a metastable state will be rare whenever the condition $E_{barrier}/E_{noise} \gg 1$ holds, where $E_{barrier}$ denotes the energy barrier height which separates the metastable state.

Under physical assumptions on the governing dynamics of the process, the time scale of escape from a metastable state depends exponentially on the ratio $E_{barrier}/E_{noise}$. This means that one has to wait exponentially long to observe a single transition. On the other hand, the impact of the motion on the fastest time scale on the global behavior of the process is not negligible. Consequently, any direct numerical simulation of the dynamics in order to get a sufficient statistics on transition events would fail. Hence, alternative and effective strategies are required and had been developed such as Transition State Theory, Transition Path Sampling, and more recently Transition Path Theory.

In the present work we give a unified presentation of Transition Path Theory (TPT) for time-continuous Markov processes and we elucidate its range of applicability on the example of conformational dynamics of bio-molecules.

We consider the most interesting results to include the following:

- Illustration of TPT on several low dimensional examples for Smoluchowski and Langevin dynamics arising from the stochastic modeling of molecular dynamics.
- Derivation of a stable finite discretization scheme of the committor function equation associated with the hypoelliptic Langevin dynamics.
- Adaptation of TPT to the class of time-continuous Markov processes with discrete state space (Markov jump processes).
- Development of efficient graph algorithms for identifying transition pathways for Markov jump processes and in Markov chains.

- Presentation, improvement and comparison of methods to estimate an infinitesimal generator of a Markov jump process if only an incomplete observation of the process is available.
- Derivation of an Metropolis Monte Carlo Markov chain method to investigate the error propagation in the discrete committor function computation for Markov chains.

**Rare Events in Molecular Dynamics** In the classical description of molecular processes the dynamics of the molecule's microscopic configurations (position and momenta) are mathematically modeled in terms of ordinary differential equation, resulting from formulations of Lagrange and Hamilton. Within these models, the physical interactions of atoms are encoded in the interaction *potential* which is composed of sums of contributions of different physical origin as the bond structure of the molecule and electrostatic interactions. But most biomolecular processes can only be understood within a thermodynamical context; instead of a single molecular system as a solution of the classical equations, one is interested in statistical ensembles, since only such ensembles can be object of experimental investigation. Throughout this thesis we will focus on that ensemble view.

Functions of bio-molecules depend on their dynamical properties, and especially on their ability to undergo transitions between long-living states, called *conformations*. A conformation of a molecule is understood as a mean geometric structure of the molecule which is conserved on a large time scale compared to the fastest molecular motions where the system may well rotate, oscillate or fluctuate. From the dynamical point of view, a conformation typically persists for a long time (again compared to the fastest molecular motions) such that the associated subset of microscopic configurations is almost invariant or *metastable* [82] with respect to the dynamics. Hence transitions between different conformations of a molecule are rare events compared to the fluctuations within each conformation.

A very popular model to describe molecular systems including thermal noise is the stochastic Langevin dynamics or Smoluchowski dynamics. A Langevin system can be regarded as a mechanical system with additional noise and friction where the noise can be thought of modeling the influence of a heat bath surrounding the molecule and the friction is chosen such as to counterbalance the energy fluctuations due to the noise [45]. The Smoluchowski dynamics [87] is a Brownian motion which results from the Langevin dynamics in the high friction limit and acts only on the position space.

Mathematically, the Langevin and Smoluchowski dynamics are time-continuous Markov diffusion processes on a continuous state space. Under weak conditions both admit a unique stationary (equilibrium) distribution in configuration space which corresponds to the stationary (canonical) ensemble in experiments under constant volume and temperature, respectively.

As mentioned above, the problem of identifying conformations amounts to the identification of metastable sets in configuration space. The characterization of metastability within the canonical ensemble hence requires the mathematical description of the propagation of sub-ensembles. This is accomplished by the *transfer operator approach* [80]; if we define a transition probability from a sub-ensemble $C$ into another sub-ensemble $B$ in time $\tau$, denoted by $p(\tau, C, B)$ then $C$ will be called

metastable on a time slice $\tau$ if the fraction of the systems in that sub-ensemble which stays in $C$ after time $\tau$ is almost one, i.e. $p(\tau, C, C) \approx 1$ [51]. Finally, the algorithmic strategy to decompose the state space into metastable states is based on spectral properties of the *transfer operator* [24].

**Transition State Theory** Since the 1930s transition state theory (TST) and evolutions thereof based on the reactive flux formalism have provided the main theoretical framework for the description of rare events [37, 95, 97, 7, 15]. Originally, TST was derived in the context of analyzing the rate of chemical reactions $R \to P$, where $R$ denotes the reactant and $P$ the product. The idea behind TST is to approximate the reaction rate $k$ by the mean crossing frequency $k^{TST}$ of transitions from $R$ to $P$ through a *transition state*, the dynamical bottleneck for the reaction. Generally, the transition state can be any dividing surface separating the reactant state $R$ from the product state $P$. Then the TST rate, $k^{TST}$, is proportional to the total flux of *reactive trajectories*, i.e., trajectories from the reactant to the product side of the dividing surface, and can be expressed in terms of thermodynamical quantities.

The TST rate is always an upper bound of the true reaction rate because reactive trajectories can recross the transition state many times during one reaction. Therefore, the true rate is given by

$$k = \kappa k^{TST},$$

where $\kappa$, the *transition coefficient*, is a correcting factor accounting for these recrossings. Due to this overestimation, several strategies have been proposed to improve the TST rate. For example, the earliest one is called *variational* TST [50] and amounts to choose the dividing surface which minimizes the TST rate constant (see also [91, 94]).

Performing the computation in practice, however, may prove very challenging, and this difficulty is related to a deficiency of the theory. TST is based on partitioning the system into two, leaving the reactant state on one side of a dividing surface and the product state on the other, and the theory only tells how this surface is crossed during the reaction. As a result, TST provides very little information about the mechanism of the transition, which has bad consequences e.g. if this mechanism is totally unknown *a priori*. In this case, it is difficult to choose a suitable dividing surface and a bad choice will lead to a very poor estimate of the rate by TST (too many spurious crossings of the surface that do not correspond to actual reactive events). The TST estimate is then extremely difficult to correct. The situation is even worse when the reaction is of diffusive type, since in this case all surfaces are crossed many times during a single reactive event and there is simply no good TST dividing surface that exists.

**Transition Path Sampling** How to go beyond TST and describe rare events whose mechanism is unknown *a priori* is an active area of research and several new techniques have been developed to tackle these situations. Most notable among these techniques are the transition path sampling (TPS) technique of Bolhuis, Chandler, Dellago, and Geissler [72, 21] and the action method of Elber [35, 36] which allow to sample directly the ensemble of reactive trajectories, i.e. the trajectories by which the reaction occurs.

The basic idea behind TPS is a generalization of standard Monte Carlo Markov Chain (MCMC) [39, 56] procedures on the trajectory space of the considered dynamics. Generally, an MCMC procedure performs a biased random walk on the configuration space such that the number of visits of a configuration $x$ is proportional to its probability $p(x)$. In TPS a configuration $X(\mathcal{T}) = (x_0, x_{\Delta t} \ldots, x_{\mathcal{T}})$ is a sequence of states representing a time-discretization of a true dynamical trajectory of fixed length $\mathcal{T}$ rather than individual states of the dynamics itself. The statistical weight $p(X(\mathcal{T}))$ depends on the initial conditions and on the underlying dynamics. Since one is only interested in reactive trajectories connecting $A$ and $B$, TPS finally performs a random walk on the *transition path ensemble* with respect to the *reactive path probability*

$$p_{AB}(X(\mathcal{T})) = Z_{AB}^{-1}(\mathcal{T}) \mathbb{1}_A(x_0) p(X(\mathcal{T})) \mathbb{1}_B(x_{\mathcal{T}}),$$

where $Z_{AB}$ normalizes the distribution of the transition path ensemble and the characteristic $\mathbb{1}_A(x)$ is equal one if $x \in A$ and 0 otherwise ($\mathbb{1}_B(x)$ is defined analogously).

Following [72]:

> *Metaphorically, TPS is akin to "throwing ropes over rough mountains passes, in the dark" where "throwing ropes" stands for shooting trajectories, attempting to reach one metastable state from another and "in the dark" because high-dimensional systems are so complex that it is generally impossible to make any prediction on the relevant energy surfaces.*

We want to emphasize that reactive trajectories in the transition path ensemble are true dynamical trajectories, free of any bias by non-physical forces, constraints or assumptions on the reaction mechanism. The mechanism of the reaction and possibly its rate can then be obtained *a posteriori* by analyzing the ensemble of reactive trajectories. However, these operations are far from trivial. TPS or the action method *per se* do not tell how this analysis must be done and simple inspection of the reactive trajectories may not be sufficient to understand the mechanism of the reaction. This may sound paradoxical at first, but the problem is that the reactive trajectories may be very complicated objects from which it is difficult to extract the quantities of real interest such as the probability density that a reactive trajectory be at a given location in state-space, the probability current of these reactive trajectories, or their rate of appearance. In a way, this difficulty is the same that one would encounter having generated a long trajectory from the law of classical mechanics but ignoring all about statistical mechanics: how to interpret this trajectory would then be unclear. Similarly, the statistical framework to interpret the reactive trajectories is not given by the trajectories themselves, and further analysis beyond TPS or the action method is necessary (for an attempt in this direction, see [52]).

**Transition Path Theory** Recently, a theoretical framework to describe the statistical properties of the reactive trajectories in the context of Markov diffusion processes has been introduced [34, 92]. This framework, termed transition path theory (TPT), goes beyond standard equilibrium statistical mechanics and accounts for the non-trivial bias that the very definition of the reactive trajectories imply – they must be involved in a reaction.

TPT allows to understand the statistical properties of the ensemble of all reactive trajectories (not only reactive trajectories with respect to a fixed length as in TPS) by giving precise answers to the following questions:

- What is the probability to encounter a reactive trajectory in a given state, i.e. what is the *probability density function of reactive trajectories*?
- What is the net amount of reactive trajectories going through a given state, i.e. what is the *probability current of reactive trajectories*?
- What is the mean frequency of transitions between two sets, say $A$ and $B$, i.e. what is the *rate of reaction*?
- What are the mechanisms of transitions, i.e. what are the *transition tubes* or *transition pathways*?

The key ingredient in the main objects provided by TPT is the *committor function* $q_{AB}(x) \equiv q(x)$ which is the probability to go rather to the set $B$ than to the set $A$ conditional on the process has started in the state $x$. The committor function $q(x)$ can be seen as an abstract reaction coordinate, because under appropriate conditions on the dynamics the levels sets of the committor function foliate the state space in sets of equal probability to rather end up in $B$ than $A$, i.e. it describes the progress of reaction from $A$ to $B$ in terms of probabilities.

For Markov diffusion processes, the committor function satisfies a boundary value problem where the involved partial differential operator is the generator of the diffusion process under consideration. Solving the committor equation numerically in high dimensions is infeasible and, hence, TPT is impractical for the analysis of high dimensional complex processes.

As a remedy to avoid the "curse of dimension" we will follow a two-step procedure. Instead of considering the system in all its degrees of freedom, we will choose appropriate low-dimensional observables which allow to describe the effective dynamics of the system. In the second step the dynamics in these observables is considered on a coarse grained level, e.g. on a discretization of the image space of the observables, and modeled as a Markov jump process. As a result the essential dynamics of the complex system is captured in a discrete transition network (see Figure 1).

For discrete representatives of the sets $A$ and $B$, discrete TPT [66] allows to analyze the statistical properties of the associated reactive trajectories, i.e. these trajectories by which the walkers transit on the discrete state space from $A$ to $B$ driven by the underlying Markov jump process. Discrete TPT provides discrete analogs of the probability density, the transition rate and the probability current of reaction trajectories. Again, these objects depend on a discrete committor function which satisfies a linear system of equations involving the infinitesimal generator of the considered jump process. Within this discrete setting, then it is easy to compute transition rates and, moreover, to identify transition pathways by utilizing Graph algorithms.

Finally, it is worth to point out that TPT is the theoretical background beyond the string method [30, 31, 32, 33, 75, 60], which is a numerical technique to compute the statistical properties of the reactive trajectories directly (that is, without having to identify these trajectories themselves beforehand as in TPS or the action method) in complicated systems with many degrees of freedom.

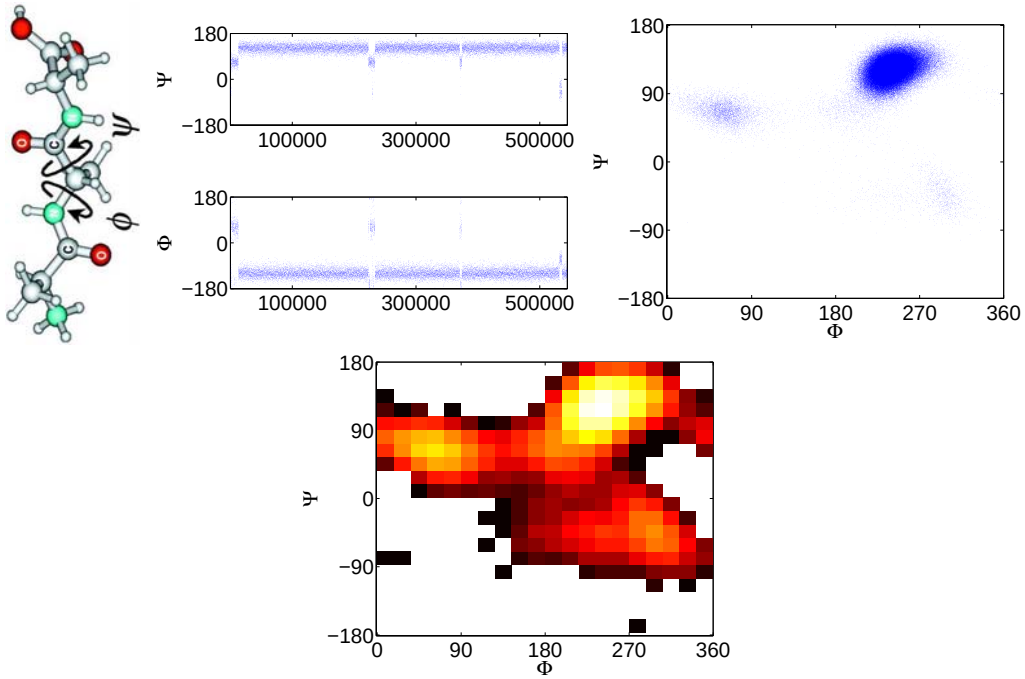


Figure 1.1.: In this figure we exemplify our strategy to capture the essential dynamics of a bio-molecule in a coarse grained model. The top left panel shows the ball-and-stick representation of the trialanine dipeptide analog. Top right: Projection of the time series (all atomic positions) onto the torsion angle space spanned by $\Phi$ and $\Psi$, which reveals the metastable behavior. Bottom left: The Ramachandran plot of the torsion angle time series. At first glance, trialanine attains three different conformations, indicated by the three clusters. Bottom right: The discrete free energy, $-\log \pi$, associated with the stationary distribution $\pi$ of a Markov jump process which models the effective dynamics of a system in terms of the torsion angles $\Phi$ and $\Psi$. The jump process was estimated from the underlying time series with respect to a $20 \times 20$ box discretization of the torsion angle space. The lighter the color of a box the more probable to encounter the process in that box.

**Acknowledgments** I would like to express my gratitude to all those who gave me the possibility to complete this thesis. In particular, I would like to thank my visors Prof. Dr. Christof Schütte and Prof. Dr. Eric Vanden-Eijnden for their continuous support and patience during my studies.

Special thanks to Jessica Walter for convincing me to do my Ph.D. in the Bio Computing group, Alexander Fischer and Illia Horenko for taking me by the hands on my first steps into the field of molecular dynamics and Eike Meerbach for providing me data from molecular dynamics simulations (not to mention his tolerance for thousand of hours of stimulating dark wave music). I'm indebted to thank Heidi for not just simply being there during one year. Finally, without the support of my parents, my sister and my friends this all would not have been possible: thank you.

This work was funded by the DFG Research Center Matheon "Mathematics for Key Technologies" (FZT86) in Berlin.