Stochastic dynamics and dominant protein folding pathways

P. Faccioli \textsuperscript{a, b}, M. Sega \textsuperscript{c}, F. Pederiva \textsuperscript{d}, H. Orland \textsuperscript{c}

\textsuperscript{a} Dipartimento di Fisica, Università degli Studi di Trento, Trento, Italy
\textsuperscript{b} I.N.F.N., Gruppo Collegato di Trento, Trento, Italy
\textsuperscript{c} Frankfurt Institute for Advanced Studies, Frankfurt, Germany
\textsuperscript{d} Institut de Physique Théorique, Centre d'Etudes de Saclay, France


To link to this Article DOI: 10.1080/14786430802609644

URL: http://dx.doi.org/10.1080/14786430802609644
Stochastic dynamics and dominant protein folding pathways

P. Faccioli\textsuperscript{a,b*}, M. Sega\textsuperscript{c}, F. Pederiva\textsuperscript{d} and H. Orland\textsuperscript{e}

\textsuperscript{a}Dipartimento di Fisica, Universit\`a degli Studi di Trento, Trento, Italy; \textsuperscript{b}I.N.F.N., Gruppo Collegato di Trento, Trento, Italy; \textsuperscript{c}Frankfurt Institute for Advanced Studies, Frankfurt, Germany; \textsuperscript{d}Institut de Physique Théorique, Centre d'Etudes de Saclay, Gif-sur-Yvette, France

(Received 27 June 2008; final version received 7 November 2008)

We present the results of a recently developed theoretical framework denominated Dominant Reaction Pathways (DRP), to study thermally activated reactions in multi-dimensional systems. In particular, we focus on application to the protein folding reaction. By applying the saddle-point approximation to the stochastic path integral generated by the Langevin equation, we derive a least-action principle, which allows us to rigorously determine directly the most probable reaction pathways, bypassing the long-standing computational problems associated with the decoupling of time-scales in the problem. We show the results of number validation studies, in which the accuracy of the DRP approach was assessed, studying molecular transitions. In all cases, the DRP predictions are found to be consistent with the MD results, but extremely less computationally expensive.

Keywords: polymer dynamics; protein folding; stochastic path integrals

1. Introduction

The problem of characterizing thermally activated transitions of bio-molecules is crucial for the understanding of many phenomena at the interface of physics, chemistry and biology. In particular, the so-called ‘Part II’ of the Protein Folding Problem consists in identifying the sequence of conformational transitions that proteins assume on their way from the denatured state to the native state (‘Part I’: understanding the sequence structure relationship).

From a theoretical perspective, the most natural strategy to approach this problem would be to integrate numerically the equations of motion, e.g. via Molecular Dynamics (MD) simulations. In fact, from a statistically significant sample of MD transitions it is possible to identify the transition state and compute folding rates, which can be matched against experimental data.

Unfortunately, state-of-the-art all-atom MD simulations cover time intervals in the range $10^{-10^2}$ ns (see e.g. [1] and references therein), while folding transitions take place in the ms–s time range. Hence, MD can be useful only in investigating the reactions leading to local secondary structure formation; i.e. $\alpha$-helices and $\beta$-sheets. The reason for such
a decoupling of time-scales is that the folding transitions involve overcoming free-energy barriers of several units of \( k_B T \). As result, the computational time is entirely invested in describing the motion of the molecule in the molten globule state (for folding reactions) or in the native state (for unfolding reactions). On the other hand, one is interested in the information encoded in the reaction pathways joining these two states.

Given the limitations of the available MD simulations, several alternative approaches have been recently developed, to deal with the problem of characterizing the entire reaction in a high-dimensional space, using available computers. A drawback of these methods is that they rely either on ad hoc assumptions about the dynamics obeyed by the system [2–4] or on a specific choice of reaction coordinates [5,6].

In this contribution, we shall present the results of the DRP method [7–10], which allows us to rigorously identify the statistically significant protein folding pathways, without relying on any choice of reaction coordinates and without involving any uncontrolled ad hoc approximation.

The major advantage of the DRP approach is that it avoids investing computational time in simulating the local thermal motion in the metastable configurations. This is possible because the key equation can be formulated in a form that does not depend explicitly on the time variable. As a result, the computational difficulties associated with the existence of different time-scales in rare thermally activated reactions are naturally and rigorously bypassed and studying conformational transitions of large molecules in atomistic details becomes feasible on available computers.

2. Brief review of the DRP approach

The starting point of the DRP approach is the path integral representation of the Fokker-Planck conditional probability to perform a transition from an initial configuration \( x_i \) to a final configuration \( x_f \) in a time \( t \); i.e.

\[
P(x_f, t|x_i, 0) = \exp\left(-\frac{(U(x_f) - U(x_i))}{2k_B T}\right) \int_{x_i}^{x_f} \mathcal{D}x(\tau) \exp\left(-\int_0^t d\tau \left(\frac{x^2(\tau)}{4D} + V_{\text{eff}}[x(\tau)]\right)\right).
\]

(1)

\( D \) is a constant diffusion coefficient, and \( V_{\text{eff}}(x) \) is an effective potential defined as

\[
V_{\text{eff}}(x) = \frac{D}{4(k_B T)^2} \left( (\nabla U(x))^2 - 2k_B T \nabla^2 U(x) \right),
\]

(2)

\( x = (x_1, \ldots, x_{N_p}) \) is a vector specifying the position of all the \( N_p \) constituents of the molecule (atoms or amino-acids). In the specific case of the study of the protein folding reaction, the initial configuration \( x_i \) may be taken in the denatured state, while the final configuration \( x_f \equiv x_N \) lies in the native state. The representation (1) is analogous to a quantum mechanical path integral in imaginary time, in a system specified by the action \( S_{\text{eff}}[x] = \int_0^t d\tau \left(\frac{x^2(\tau)}{4D} + V_{\text{eff}}[x(\tau)]\right)\).

The most probable paths contributing to (1) are those for which the exponential weight \( \exp(-S_{\text{eff}}) \) is maximum, hence for which \( S_{\text{eff}} \) is minimum. A trajectory that connects configurations that are not classically accessible in the absence of thermal fluctuations, corresponds to an instanton in quantum-mechanical language. In the context of diffusive
dynamics, we shall refer to these as to the dominant reaction pathways. Determining these trajectories for realistic proteins using conventional methods – such as Molecular Dynamics – is extremely challenging from the computational point of view. In addition to the numerical difficulties associated with the existence of very different time-scales, one has also to face the solution of boundary-value problems, which are considerably harder than initial-value problems.

Fortunately, a dramatic simplification is obtained upon observing that the dynamics described by the effective action $S_{\text{eff}}$ are energy-conserving and time-reversible. This property allows us to switch from the time-dependent Newtonian description to the energy-dependent Hamilton-Jacobi (HJ) description. We note that this could not be done at the level of the Langevin equations. In the HJ framework, the dominant pathways connecting, given initial and final positions, are obtained by minimizing numerically – e.g. via simulated annealing – a discretized version of the target function (HJ functional):

$$S_{\text{HJ}} = \int_{x_i}^{x_f} dl \sqrt{\frac{1}{D} (E_{\text{eff}} + V_{\text{eff}}[x(l)])},$$

(3)

where $dl$ is an infinitesimal displacement along the path trajectory. $E_{\text{eff}}$ is a free parameter that determines the total time elapsed during the transition, according to

$$t_f - t_i = \int_{x_i}^{x_f} dl \sqrt{\frac{1}{4D(E_{\text{eff}} + V_{\text{eff}}[x(l)])}}.$$  

(4)

In [8,9], we showed that for thermally activated transitions one needs to choose $E_{\text{eff}} = -V_{\text{eff}}(x_N)$.

The HJ formulation of the dynamics leads to an impressive computational simplification of this problem. In fact, the total Euclidean distance between the coil state and the native state of a typical protein is only 1–2 orders of magnitude larger than the most microscopic length-scale; i.e. the typical monomer (or atom) size. As a consequence, only $\sim$30–50 discretized displacement steps are usually sufficient for convergence. This number should be compared with $10^{12}$ time-steps required in the time-dependent Newtonian description. The physical reason why the HJ formulation is so much more efficient compared to the Newtonian formulation is the following: in traditional Molecular Dynamics simulations, proteins spend most of their time in meta-stable minima, trying to overcome free-energy barriers. The HJ formulation avoids investing computational times in such ‘waiting’ phases by considering intervals of fixed displacements, rather than fixed time-length.

Once the dominant pathways have been determined, it is also possible to systematically account for the effects of quadratic thermal fluctuations around them, by means of the Monte Carlo algorithm [9]. Let us denote by

$$\tilde{x}(n) = (\tilde{x}_1(n), \ldots, \tilde{x}_N(n)),$$

(5)

with $n = 1, \ldots, N$, the dominant pathway trajectory corresponding to a sequence of configurations of the molecule, and determined by minimizing numerically the HJ action (3). The time at which each configuration $\tilde{x}(n)$ is visited during the transition can be
obtained by computing the set of time intervals separating each of the path steps \(\tilde{x}(n)\) from \(\tilde{x}(n+1)\):

\[
\Delta \tau_{n,n+1} = \frac{\Delta l_{n,n+1}}{\sqrt{4D(V_{\text{eff}}(\tilde{x}(n)) - V_{\text{eff}}(\tilde{x}(1)))}}.
\]

The sequence of time intervals obtained this way can be used to write a discretized version of the transition probability (1):

\[
P(x_f, x_i; \tau(N)) = \int \prod_{n=1}^{N-1} [dx_1(n) \ldots dx_N(n)]
\times \exp \left( - \sum_{n=1}^{N-1} \Delta \tau_{n,n+1} \left[ \frac{1}{4D} \sum_{i=1}^{N_P} \left( \frac{x_i(n+1) - x_i(n)}{\Delta \tau_{n,n+1}} \right)^2 + V_{\text{eff}}(x_1(n), \ldots, x_{N_P}(n)) \right] \right).
\]

We stress the fact that, in Equation (7), the time intervals are chosen large (small), when the most probable trajectory at that time is evolving slowly (fast). In other words, the information encoded in the dominant pathway has been used to identify a particularly convenient discretized representation of the path integral (1), in which the sizes of the time steps are adapted according to the evolution of the most probable pathways.

3. Assessing the reliability of the DRP approach

The first validation study of the DRP approach was performed in [8], in which we studied the kinetics of alanine dipeptide, using the GROMOS96 force field.

In Figure 1, we present the results of the analysis relative to two specific transitions (\(C7_{\text{ax}} \rightarrow C7_{\text{eq}}\) and \(\alpha_L \rightarrow \alpha_R\)), compared with the free-energy landscape computed by direct integration. The values of the two \(\phi\) and \(\psi\) dihedrals along the paths obtained by minimizing the effective action are plotted on top of the relative free-energy map. These simulations were performed at temperature \(T = 300\) K, and assuming a diffusion constant of \(D = 0.02\) \(\text{Å}^2\ \text{ps}^{-1}\) for all atoms. An analysis of the residence time along the path shows that in each of the two dominant pathways, there are two points where the conformation of alanine dipeptide has shortest residence time. These points, indicated in Figure 1 with black symbols, very accurately locate the maxima of the free energy along the path. In all similar analyses performed on this system, we observed that the information encoded in the dominant pathway trajectories was in excellent agreement with the results of MD simulations, performed in the same model. On the other hand, the computational gain of adopting the DRP method was impressive: the characterization of the relevant transitions in atomistic detail took just a few minutes on a regular desktop, while about a week of calculations on the same machine was required, in order to extract the same amount of information, from the standard MD techniques.

The next step in our validation study consists of investigating the reliability of the DRP approach, when it is applied to characterize the folding reaction of a longer polypeptide chain. Since we are interested in comparing the results of the DRP method with those of traditional MD simulations, we chose a relatively small chain and adopted a coarse-grained model: we studied the folding of 16-residue \(\beta\)-hairpin GB1 [11–15], using
a Go-type force field, between the amino-acids. In fact, the dynamics of the Go model are sufficiently simple to allow several folding–unfolding trajectories to be generated by MD on a regular workstation, at an affordable computational cost. On the other hand, such a model displays some of the complexity that we expect to face, in more realistic and detailed all-atom simulations, of long polypeptide chains. For example, the molecule presents a unique stable fold and a ensemble of denatured configurations with large conformational entropy.

In the three panels of Figure 2 we present some preliminary results in which the predictions of the MD and DRP approaches are compared. In order to analyze the results, we project the dynamics on the three planes, selected by the following set of order parameters: the distance $d_{1-16}$ between the $C$ and $N$ termini of the chain, the angle $\phi$ at the turn of the hairpin (i.e. between the ASP-ALA and LYS residues) the fraction of native-like contacts and the radius of gyration $R_H$ of the group defined by the hydrophobic residues Phe, Tyr, and Trp.

The points correspond to configurations obtained by running five independent MD trajectories, of 15 ns each. The squares represent points obtained in the DRP approach, starting from 10 independent dominant pathways and sampling three different thermal fluctuations around each of them. The grey circle represents the value of the order parameters in the native state conformation, which was downloaded from the Protein Data Bank.

We can see that the DRP method correctly identifies the regions of configuration space, which are explored by the MD trajectories. This fact is observed in all the combinations of the selected set of reaction coordinates. More quantitative comparison between MD and
DRP, based on the study of the free energy landscape, is in progress [10]. It is important to stress that we do not expect the configurations visited by the DRP trajectories to accumulate in the regions where the density of MD configurations is highest. In fact, by definition, the density of MD configurations increases in the vicinity of the (meta-)stable states and becomes small (but finite) in the vicinity of the configurations visited during the folding–unfolding transitions. On the other hand, by construction, each dominant pathway trajectory leaves the native state and visits denatured conformations, by steps of equal $dl = \sqrt{\sum_i d\chi_i^2}$ and does not accumulate in the vicinity of stable and meta-stable states.

The analysis of the dominant pathway trajectories can be used to investigate the folding mechanism of the $\beta$-hairpin, within the present simple Go-type model. We have found support for a hydrophobic collapse mechanism [10]. In fact, in the initial (final) stage of the unfolding (folding) reaction, the size of the hydrophobic core remains constant and consistent with the value in the native state, while the number of native-like contacts rapidly drops (increases).

4. Conclusions and future developments

In this paper, we have presented some recent applications of the DRP method to study conformational transitions of macromolecules. In the analysis of transitions of alanine-dipeptide using the GROMOS94 force field, and in the study of the folding of the 16-residue $\beta$-hairpin terminus of protein GB1 using a coarse-grained model, the DRP predictions were confronted with the results of MD simulations. We have found that the two methods lead to consistent pictures of the transition. On the other hand, the DRP approach leads to a huge computational gain, relative to traditional MD simulations. This development opens the door to a characterization of thermally activated reactions involving large molecules such as proteins, in atomistic detail using available computers.

Note

1. Two residues are considered in contact if their distance is less than 6 Å. The fraction of native-like contacts in Figure 2 is defined as the number of non-consecutive residues in contact, divided by the number of non-consecutive residues in contact in the native state.
References