Variational Scheme to Compute Protein Reaction Pathways Using Atomistic Force Fields with Explicit Solvent

S. a Beccara,[1,3] L. Fant,[2] and P. Faccioli[2,3,*]
[1]*European Centre for Theoretical Nuclear Physics and Related Areas (ECT*-FBK), Strada delle Tabarelle 287, Villazzano (Trento) 38123, Italy*
[2]*Dipartimento di Fisica, Università degli Studi di Trento, Via Sommarive 14, Povo (Trento) 38123, Italy*
[3]*Trento Institute for Fundamental Physics and Applications (INFN-TIFPA), Via Sommarive 14, Povo (Trento) 38123, Italy*

(Received 23 May 2014; revised manuscript received 29 October 2014; published 4 March 2015)

We introduce a variational approximation to the microscopic dynamics of rare conformational transitions of macromolecules. Within this framework it is possible to simulate on a small computer cluster reactions as complex as protein folding, using state of the art all-atom force fields in explicit solvent. We test this method against MD simulations of the folding of an α and a β protein performed with the same all-atom force field on the Anton supercomputer. We find that our approach yields results consistent with those of MD simulations, at a computational cost orders of magnitude smaller.

DOI: 10.1103/PhysRevLett.114.098103 PACS numbers: 87.15.ap, 87.10.Tf, 87.14.E-

The development of the special-purpose Anton supercomputer has recently opened the way to MD simulations of biomolecules consisting of several hundred atoms, covering time intervals in the millisecond range [1]. By using this facility, Shaw and co-workers characterized the reversible folding of several small proteins, showing that the existing all-atom force fields are able to attain the correct protein native structures [1–3]. Unfortunately, many biologically important conformational reactions occur at time scales many orders of magnitude larger than the millisecond. Hence, it is important to continue the development of more efficient algorithms to sample the reactive pathways space (see, e.g., Ref. [4] and references therein).

In particular, in the dominant reaction pathways (DRP) approach [5–8], microscopic trajectories $X(\tau)$, connecting given initial and final molecular configurations $X_i = X(0)$ and $X_f = X(t)$, are determined by maximizing their probability density $\mathcal{P}[X]$ in the Langevin dynamics. This algorithm was first validated against MD using both simplified and realistic atomistic force fields (see, e.g., Ref. [8]). Next, it was applied to characterize in atomistic detail conformational reactions far too slow to be investigated by means of plain MD. Notable examples include the folding of a knotted protein [9] and the latency transition of several serpins [10].

One crucial limitation of the DRP method is that it can only be applied in implicit solvent simulations. In this work we overcome this limitation by introducing a new variational approximation suitable also for atomistic simulations in an explicit solvent.

Let (X, Y) represent a point of the system's configuration space, where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_{N'})$ denote the solute and solvent coordinates, respectively. The Langevin equations for the solvent and solute are

$$m_i\ddot{\mathbf{x}}_i = -m_i\gamma_i\dot{\mathbf{x}}_i - \nabla_i U + \eta_i(t),$$
$$m_j\ddot{\mathbf{y}}_j = -m_j\gamma_j\dot{\mathbf{y}}_j - \nabla_j U + \eta_j(t), \qquad (1)$$

where $U(X, Y)$ is the potential energy, η_i is a white noise, and m_i and γ_i denote mass and viscosity, respectively.

We are interested in the probability density for the solute to make a transition from X_i to X_f in a time t, along a given path $X(\tau)$. This is given by the path integral (PI),

$$\mathcal{P}[X] = \int \mathcal{D}Y e^{-S_{\mathrm{OM}}[X,Y]-U(X_i,Y_i)/k_BT}, \qquad (2)$$

where $S_{\mathrm{OM}}[X, Y]$ is the Onsager-Machlup functional, to be defined below. Maximizing $\mathcal{P}[X]$ with respect to the path X yields the DRP optimum condition [5–7]: $(\delta/\delta X)\langle S_{\mathrm{OM}}[X, Y]\rangle_Y = 0$, where the average $\langle\cdot\rangle_Y$ refers to the PI over $Y(\tau)$.

Unfortunately, computing this average with the accuracy required for the path optimization is computationally unfeasible, because of large statistical fluctuations. To overcome this problem, we need to derive an optimum criterion that does not involve any average over the solvent dynamics.

We begin by considering a modified stochastic dynamics, defined by introducing into Eq. (1) an *external* (possibly time-dependent) biasing force $\mathbf{F}_i^{\mathrm{bias}}(X, t)$, acting on the solute atoms only and accelerating the transition to the product. The probability of a given reactive pathway $X(\tau)$ in the biased dynamics is given by

$$\mathcal{P}_{\text{bias}}[X] = \int \mathcal{D}Y e^{-S_{\text{bias}}[X,Y]-U(X_i,Y_i)/k_BT}, \tag{3}$$

where the functional $S_{\text{bias}}[X,Y]$ is defined as

$$S_{\text{bias}} \equiv \frac{1}{4k_BT}\int_0^t d\tau \left[\sum_{i=1}^{N}\frac{1}{\gamma_i m_i}(m_i\ddot{\mathbf{x}}_i + m_i\gamma_i\dot{\mathbf{x}}_i + \nabla_i U - \mathbf{F}_i^{\text{bias}})^2 + \sum_{j=1}^{N'}\frac{1}{\gamma_j m_j}(m_j\ddot{\mathbf{y}}_i + m_j\gamma_j\dot{\mathbf{y}}_j + \nabla_j U)^2\right]. \tag{4}$$

The Onsager-Machlup functional $S_{\text{OM}}[X,Y]$ entering Eq. (2) is recovered, setting $\mathbf{F}_i^{\text{bias}} = 0$ in Eq. (4).

Let us now return to the problem of computing the reaction pathways in the *unbiased* Langevin dynamics [Eq. (1)]. Using the standard reweighting trick we can write the variational condition $(\delta/\delta X)\mathcal{P}[X] = 0$ as

$$\frac{\delta}{\delta X}[\mathcal{P}_{\text{bias}}[X]\langle e^{-(S_{\text{OM}}[X,Y]-S_{\text{bias}}[X,Y;t])}\rangle_{\text{bias}}] = 0. \tag{5}$$

We now introduce our main approximation, by restricting the search for the optimum path $X(\tau)$ within an ensemble of trajectories generated by integrating the *biased* Langevin equation. By definition, these paths have a large statistical weight in the biased dynamics; i.e., they lie in the functional vicinity of some path $\bar{X}(\tau)$ which satisfies $(\delta/\delta\bar{X})\mathcal{P}[\bar{X}] = 0$. Thus, the typical biased paths approximatively satisfy the stationary condition

$$\frac{\delta}{\delta X}\mathcal{P}[X] \simeq 0 \tag{6}$$

and obey the corresponding saddle point equations of motion:

$$\begin{aligned} m_i\ddot{\mathbf{x}}_i + m_i\gamma_i\dot{\mathbf{x}}_i + \nabla_i U - \mathbf{F}_i^{\text{bias}} &\simeq 0, \\ m_j\ddot{\mathbf{y}}_j + m_i\gamma_j\dot{\mathbf{y}}_j + \nabla_j U &\simeq 0. \end{aligned} \tag{7}$$

We emphasize that Eqs. (6) and (7) are only satisfied by paths generated by integrating the biased Langevin equation. Using Eq. (6) in Eq. (5), we find

$$0 \simeq \frac{\delta}{\delta X}\langle e^{-(S_{\text{OM}}[X,Y]-S_{\text{bias}}[X,Y])}\rangle_{\text{bias}}. \tag{8}$$

The crucial point to observe is that, since the biasing force $\mathbf{F}_i^{\text{bias}}$ acts on the solute atoms only, the difference $\Delta S[X] \equiv S_{\text{OM}}[X,Y] - S_{\text{bias}}[X,Y]$ does not depend on the solvent paths $Y(t)$. Thus, Eq. (8) reduces to $(\delta/\delta X)\Delta S[X] = 0$. Finally, we use the saddle point approximation [Eq. (7)] again, in order to eliminate the cross product between the terms $(m_i\ddot{\mathbf{x}}_i + m_i\gamma_i\dot{\mathbf{x}}_i + \nabla_i U)$ and $\mathbf{F}_i^{\text{bias}}$ in the expression for ΔS, yielding one more term $\propto |\mathbf{F}_i^{\text{bias}}|^2$. This leads to our final variational condition:

$$\frac{\delta}{\delta X}\int_0^t d\tau \sum_{i=1}^{N}\frac{1}{\gamma_i m_i}|\mathbf{F}_i^{\text{bias}}(X;\tau)|^2 \simeq 0. \tag{9}$$

This equation states that the optimum reaction trajectory is that for which the time-averaged square modulus of the bias force is least. Interestingly, a similar condition was recently derived in the context of optimal control theory [11]. We emphasize that the functional in Eq. (9) is not affected by solvent-induced fluctuations.

Let us now extend this discussion to include the case of a history-dependent biasing force. In particular, we focus on the ratchet-and-pawl molecular dynamics (RMD) algorithm developed in Refs. [12,13]. The advantage of this formalism is that the bias only sets in whenever the system attempts to backtrack towards the reactant—defined in terms of some position-dependent reaction coordinate (RC) z. Conversely, no bias is applied whenever the system spontaneously takes a step towards the product.

To define the RMD we consider the Langevin equations (1) with an additional biasing force $\mathbf{F}_i^{\text{RMD}}$, defined as

$$\begin{array}{ll} -\frac{k_R}{2}\nabla_i z(X)\cdot(z(X) - z_m(t)) & z(X) > z_m(t) \\ 0 & z(X) \leq z_m(t). \end{array} \tag{10}$$

$z_m(t)$ denotes the smallest value assumed by the RC z up to time t (we assume that z is minimum in the target); hence, it obeys the equation of motion $\dot{z}_m = \dot{z}\cdot\theta(z_m - z)$.

Let us now derive the PI expression for the path probability density $\mathcal{P}_{\text{RMD}}[X]$. To this end, we add a small stochastic noise to turn the equation of motion of z_m into an overdamped Langevin equation. The PI representation for the path probability density in the extended Langevin system (X, Y, z_m) is readily obtained. Finally, $\mathcal{P}_{\text{RMD}}[X]$ is recovered by taking the small-noise limit and is given by

$$\mathcal{P}_{\text{RMD}}[X] = \int_{z(X_i)}\mathcal{D}z_m\int\mathcal{D}Y e^{-S_{\text{RMD}}[X,Y,z_m]-U(X_i,Y_i)/k_BT} \cdot\delta[\dot{z}_m - \dot{z}[X]\theta(z_m[X] - z)], \tag{11}$$

where $S_{\text{RMD}}[X,Y,z_m]$ is obtained from Eq. (4) by setting $\mathbf{F}^i_{\text{bias}}(X,t) = \mathbf{F}^i_{\text{RMD}}(X,z_m)$. From here on, the derivation of the variational principle (9) is basically identical to the case of an external biasing force reported above [see Supplemental Material (SM) [14]].

Before presenting the results of atomistic protein folding simulations, it is instructive to illustrate and validate the present variational approximation on a simple toy model, which can be straightforwardly solved on a regular desktop computer. To this end, in the SM we present our study of a transition performed by a point particle diffusing on an asymmetric two-dimensional funnelled energy landscape. The diffusion from the top to the bottom of the funnel is thermally activated, due to the presence of an energy barrier. Plain MD simulations show that the particle reaches

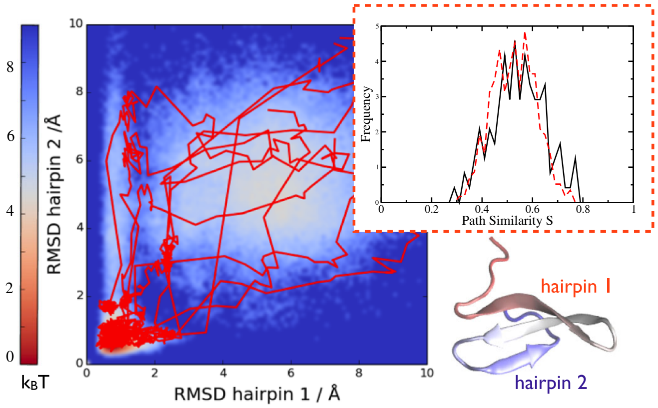

FIG. 1 (color online). Folding trajectories for the WW domain (crystal structure shown in the bottom right-hand corner) obtained in the variational approximation projected on the plain defined by RMSD to the native structure of the two hairpins. The color map in the background represents the free-energy landscape obtained from the frequency histogram of the Anton MD trajectories. Inset: Similarity distribution between variational and MD folding pathways (dashed line) compared with the intrinsic similarity of MD folding pathways (solid line).

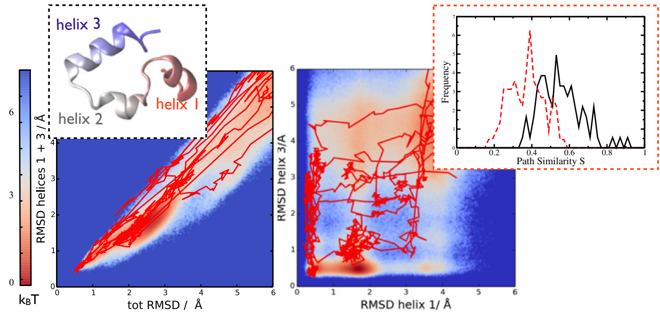

FIG. 2 (color online). Folding trajectories for villin (crystal structure shown in left-hand inset) obtained in the variational approximation, projected on the plain defined by the total RMSD to the native structure and by the RMSD to the native structure of the residues in helix I and helix III (left-hand panel) and on the plane defined by the RMSD to the native structure of helix I versus that of helix III (right-hand panel). In the background, the free-energy landscape obtained from the Anton MD simulations is shown. Right-hand inset: Distribution of similarity between variational and MD folding pathways (dashed line) compared with the intrinsic similarity of MD folding pathways (solid line).

the bottom of the funnel by passing through a gate, i.e., a spatially localized depression on the energy barrier (see Fig. 1 in the SM [14]).

We compared the results obtained using different algorithms to generate the trial paths (RMD and standard steered MD) and different values of the biasing force constant k_R. In all cases, we chose to bias the dynamics along a rather poor reaction coordinate, which does not take into account the presence of the gate.

We found that all of the trajectories generated by steered MD very closely follow the direction selected by the biasing coordinate, failing to predict the passage through the gate. Hence, in general, we expect a variational calculation based on steered MD trial paths to yield rather poor results unless the reaction coordinate is very accurately known.

Results obtained by using RMD trial paths are definitely better (see Fig. 2 of the SM [14]). In particular, even when choosing a large value for k_R, a significant fraction of the trial paths access the bottom of the funnel through the gate. This is because in RMD the biasing force is not continuously pushing the system, but only sets in to hinder backtracking. We also note that the variational principle systematically discards unphysical trial RMD trajectories and correctly predicts the essential qualitative features of the reaction. We conclude that the variational calculations based on RMD may yield reasonable results, even when the reaction coordinate is rather poorly known.

Let us now report our application to the folding transition of two globular proteins: the WW domain Fip35 (with β-type native secondary structures, see Fig. 1) and the villin headpiece subdomain (with α-type native secondary structures, see Fig. 2). In both cases, we have used the AMBER99SB-ILDN all-atom force field in TIP3P explicit water [15]. Several reversible folding-unfolding MD trajectories for these proteins generated on Anton by using the same force field have been made available by DES Research.

The RMD bias in Eq. (10) was based on the RC introduced in Ref. [13] (also reported in the SM [14]), defined as the distance between the instantaneous contact map and the native state's contact map. The k_R constant was set to 5×10^{-3} kJ/mol. With this value, the modulus of the total bias force was on average about 2 orders of magnitude smaller than that of the total physical force. We tested the robustness of our predictions by repeating the variational calculation with different values of K_R for a given initial condition (see Fig. 5 in the SM [14]).

For each test protein, we have used the RMD algorithm to produce in total about 1000 600-ps-long trial folding trajectories, started from 10 different denatured configurations $X_i^{(1)}, \ldots, X_i^{(10)}$. The 10 initial conditions were obtained by 1 ns of plain MD at the temperature $T = 800$ K, starting from the crystal native state and thermalized by 200 ps at 300 K. Folding events were defined as those attaining a final root-mean-square deviation (RMSD) to the native structure smaller than 2 Å. For each initial condition a single folding trajectory was selected out the ensemble of trial paths by applying condition (9).

In order to define a convergence criterion for the variational search, we note that the least value of the functional (9) is non-negative and vanishes for spontaneous transitions. These events have a negligible probability to be observed in the short simulation time, $t \sim 200$ ps. Typically, we observed that the least value of the functional

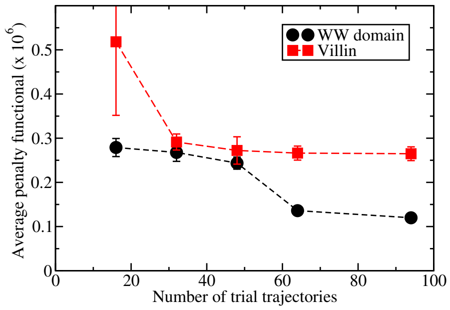

FIG. 3 (color online). The average value of the penalty functional given in Eq. (9) as a function of the number of trial trajectories. The average is performed over the different initial conditions.

(9) decreases on increasing the number of trial trajectories, until it reaches a plateau for more than ∼50 trial paths (see Fig. 3).

It is important to check that, once the plateau region is reached, the predicted folding mechanism does not change when increasing the number of trial trajectories. To this end, we adopted a simplified representation of the folding mechanism realized in a given trajectory: We define a matrix $\hat{M}$, which describes the order in which the native contacts are formed [13]. Namely, let i, j be two indexes running over all native contacts between C_α atoms, and let $t_i(k)$ and $t_j(k)$ be the times at which they are formed. The matrix element $M_{ij}(k)$ is defined to be 1 (0) when $t_i(k) < t_j(k)$ $[t_i(k) > t_j(k)]$ and 1/2 when $t_i(k) = t_j(k)$. A quantitative measure of the difference in the folding mechanism followed by two given trajectories k and k' is provided by their path similarity $s(k,k')$, defined as $s(k,k') = [1/N_c(N_c-1)]\sum_{i\neq j}\delta[M_{ij}(k) - M_{ij}(k')]$. Notice that $s(k,k') = 1$ if all native contacts are formed in the same order in k and in k', and it is 0 if they are formed in a completely different order.

The path similarity can be used to assess the stability of the predicted folding mechanism in the plateau region. For each given initial condition $X^{(i)}$, we computed the similarity between pairs of variational folding pathways, obtained using a different number of trial trajectories. Namely, we computed the similarity of the variational path obtained with 16 and 48, with 48 and 64, and with 64 and 96 trial trajectories. We found that the mechanism remains stable $[s(k,k') \gtrsim 0.9]$ above 48 trial paths, i.e., in the plateau region.

In Fig. 1 we project the folding trajectories for the WW domain obtained with our variational approach onto the plane defined by the RMSD of the two hairpins to the native state and we compare it with the free-energy landscape obtained from a frequency histogram of the long MD trajectories reported in Ref. [1].

Some comments on these results are in order. First, we note that the initial conditions used in the variational calculation are typically more denatured than the configurations in the equilibrium unfolded state obtained in the Anton simulation. In spite of this difference, the variational trajectories reach the native state by traveling along regions of low free energy. This fact indicates that the two methods yield the same folding mechanism, i.e., predict that the formation of the secondary structures predominantly occurs in a definite sequence and that in the most likely mechanism the N terminal hairpin folds before the C terminal [1,16], in agreement with the ϕ-values analysis of Ref. [17].

To provide a quantitative measure of the agreement between the variational and the MD paths, we again employed a path similarity analysis. First, we computed the distribution of $s(k,k')$ within the ensemble of MD folding trajectories (see dashed line in the inset of Fig. 1), to quantify the intrinsic degree of heterogeneity of the folding mechanism. Next, we computed the similarity between all MD and all variational paths, i.e., $s(k,k')$, where k and k' run over MD and variational trajectories, respectively (solid line). The overlap of the two curves indicates that the average difference between the folding mechanism obtained in the two methods lies within intrinsic statistical fluctuations.

A concern about the variational approach is that the bias may overpromote the rate of formation of local secondary structures, in particular, α helixes, relative to that of tertiary structures. In order to test if this is the case, we have studied the folding of a villin headpiece subdomain, which contains three α helixes. In the left-hand panel of Fig. 2, we report our variational folding trajectories projected onto the plane defined by the RMSD to the native structure of the two largest α helixes and the total RMSD to the native structure. We see that the two approaches give consistent results and predict that the formation of secondary and tertiary contacts is quite cooperative. Hence, we conclude that the bias force does not enhance the folding rate of α helixes.

In the right-hand panel of Fig. 2, we project the variational trajectories onto the plane defined by the RMSD to the native structure of the first and third helix, respectively, and we compare it with the corresponding equilibrium free-energy landscape. We note again that the variational trajectories travel along low free-energy regions, correctly predicting that the secondary structures form one after the other. However, we found that the preferential order of helix formation is different in the two calculations. This fact is reflected by a small discrepancy in the path similarity between MD and variational trajectories, of the order of the typical spread of the self-similarity distribution of the MD paths (see the inset in the top right-hand corner of Fig. 2). As a reference, the similarity distribution with random sequences of contact formation for the folding mechanism predicted by our variational method or by MD is sharply peaked around 0.3 (see Fig. 6 in the SM [14]).

In conclusion, the variational approach introduced in this work yields the microscopic mechanism for reactions as complex as protein folding, using realistic force fields in all-atom detail. In view of its computational efficiency, we foresee applications to many transitions that cannot be simulated by plain MD. The possibility of adopting the explicit solvent all-atom model opens the door to the simulation of conformational changes of other biomolecules, notably, nucleic acids.

We thank DES Research for making available their MD simulation data and acknowledge discussions with H. Orland and S. Piana. All calculations were performed on the Kore cluster at the FBK institute. S. a B. acknowledges support by Istituto Nazionale di Fisica Nucleare through the "Supercalcolo" agreement with Fondazione Bruno Kessler.

*faccioli@science.unitn.it

[1] D. E. Shaw *et al.*, Science **330**, 341 (2010).
[2] K. Lindorff-Larsen, S. Piana, R. Dror, and D. Shaw, Science **334**, 517 (2011).
[3] S. Piana, K. Lindorff-Larsen, and D. Shaw, Proc. Natl. Acad. Sci. U.S.A. **109**, 17845 (2012).
[4] W. E and E. Vanden-Eijnden, Annu. Rev. Phys. Chem. **61**, 391 (2010).
[5] R. Elber and D. Shalloway, J. Chem. Phys. **112**, 5539 (2000).
[6] P. Faccioli, M. Sega, F. Pederiva, and H. Orland, Phys. Rev. Lett. **97**, 108101 (2006); M. Sega, P. Faccioli, F. Pederiva, G. Garberoglio, and H. Orland, Phys. Rev. Lett. **99**, 118102 (2007).
[7] P. Eastman, N. Gronbech-Jensen, and S. Doniach, J. Chem. Phys. **114**, 3823 (2001).
[8] S. a Beccara, T. Škrbić, R. Covino, and P. Faccioli, Proc. Natl. Acad. Sci. U.S.A. **109**, 2330 (2012).
[9] S. a Beccara, T. Škrbić, R. Covino, C. Micheletti, and P. Faccioli, PLoS Comput. Biol. **9**, e1003002 (2013).
[10] G. Cazzolli, F. Wang, S. a Beccara, A. Gershenson, P. Faccioli, and P. L. Wintrode, Proc. Natl. Acad. Sci. U.S.A. **111**, 15414 (2014).
[11] C. Schütte, S. Winkelmann, and C. Hartmann, Math. Program. **134**, 259 (2012).
[12] E. Paci and M. Karplus, J. Mol. Biol. **288**, 441 (1999).
[13] C. Camilloni, R. Broglia, and G. Tiana, J. Chem. Phys. **134**, 045105 (2011).
[14] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.098103 for details on the illustrative toy model and related numerical results.
[15] K. Lindorff-Larsen *et al.*, Proteins **78**, 1950 (2010).
[16] S. V. Krivov, J. Phys. Chem. B **115**, 6 (2011).
[17] T. R. Weikl, Biophys. J. **94**, 929 (2008).