

Metadynamic sampling of the free energy landscapes of proteins coupled with a Monte Carlo algorithm

F. Marini ^{a,*}, C. Camilloni ^{a,b}, D. Provasi ^{a,b}, R.A. Broglia ^{a,b,c}, G. Tiana ^{a,b}

^a*Department of Physics, University of Milan, via Celoria 16, 20133 Milan, Italy*

^b*INFN, via Celoria 16, 20133 Milan, Italy*

^c*The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, DK 2100 Copenhagen, Denmark*

Abstract

Metadynamics is a powerful computational tool to obtain the free energy landscape of complex systems. The Monte Carlo algorithm has proven useful to calculate thermodynamic quantities associated with simplified models of proteins, and thus to gain an ever-increasing understanding on the general principles underlying the mechanism of protein folding. We show that it is possible to couple metadynamics and Monte Carlo algorithms to obtain the free energy of model proteins in a way which is computationally very economical.

Key words:

1. Introduction

Metadynamics is an algorithm which coupled to molecular dynamics provides an efficient tool to obtain the energy landscape of systems displaying large energy barriers, and thus whose sampling by standard tools is, at best, problematic. It is based on the knowledge of few slow collective variables s_i of the system and on the use of a non-Markovian potential $U(s_i)$ that disfavors the exploration of regions of the phase space already visited by the system (Laio and Parrinello (2002)). This algorithm has been successfully used to obtain the free energy of molecular systems at atomic detail (Babin et al. (2006)).

In the case of simplified protein models, where the atomic structure of each amino acid is coarse-grained, it is common to sample the conformational space with the help of Monte Carlo algorithms. Such an approach is computationally more economic and more simple to implement than the corresponding molecular dynamics algorithm (see, e.g. Shimada et al. (2001), Kussell et al. (2002), Shimada and Shakhnovich (2002)). It is then natural to try to extend metadynamics so as to make it possible to couple it to a Monte Carlo algorithm.

Of course, other modifications of the straight Monte Carlo sampling have been developed during the last tens of years, including simulated tempering, multicanonical sampling, parallel tempering, etc. All of them are aimed at preventing the system to get trapped in free energy minima. In the following we show that Monte Carlo metadynamics is efficient, accurate and particularly easy to implement.

We apply a scheme to the calculation of the free energy, as a function of the RMSD, of a small domain protein, namely Src-SH3. It is a widely studied domain (Grantcharova et al. (1998), Yi et al. (1998), Riddle et al. (1999)) of the *Proto-oncogene tyrosine-protein kinase Src*, a 536 residue protein that plays a multitude of roles in cell signalling. Src is involved in the control of many functions, including cell adhesion, growth, movement and differentiation. SH3 is a domain built out of 60 residues, displaying mainly β -strands (see Fig. 1). From calorimetry and fluorescence experiments, it is known to fold according to a two-state mechanism, that is, populating at biological temperature mainly two states (the native and the unfolded state) (Grantcharova and Baker (1997)). Consequently, we expect the free energy landscape to display two minima separated by a barrier.

In Section 2 we present the protein model used in the simulations along with a working description of the algorithm. We devise a formal proof of the correctness of the method in Section 3 and then test it in the specific case of

* Corresponding author.

Email address: franz.marini@mi.infn.it (F. Marini).

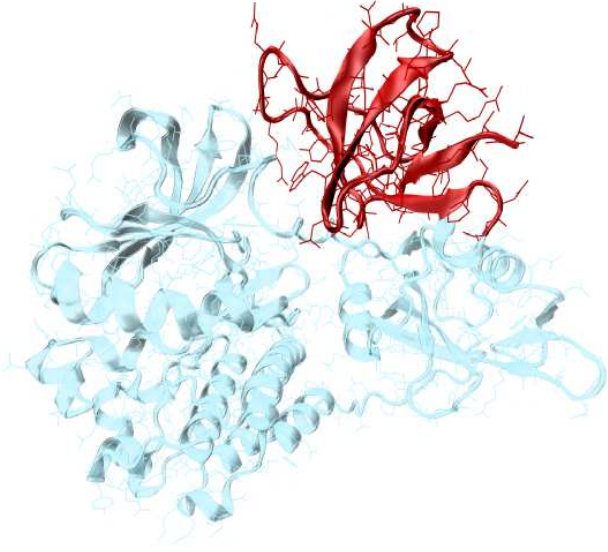


Fig. 1. Src protein (*Proto-oncogene tyrosine-protein kinase Src*). The upper right part (dark, red online) is the SH3 domain.

Src-SH3 in Section 4.

2. Method

The model employed in the simulations describes the protein as a chain of beads centered on the C_α of the protein backbone (see Fig. 2). The allowed moves are the *flip-move*

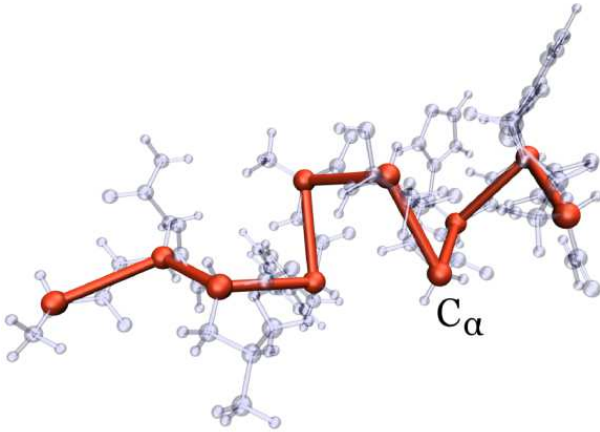


Fig. 2. Schematic protein model used in the simulations. The whole protein is shown, and in particular the linear C_α chain (dark, red online). Also visible is the sidechain (light blue online), which is although not used in the model.

and *tail-flip*. The interactions are described by a Gō-model (Gō (1975)), where the only contacts participating in the potential energy calculation are the native contacts.

Every τ steps of the Monte Carlo sampling (Metropolis and Ulam (1949)), the non-Markovian energy contribution is updated by adding a Gaussian hill with height W and spread δs , centered around the current values of the collective variables.

Each Monte Carlo step, we apply a *Metropolis* algorithm (Metropolis et al. (1953)) where the transition probability is given by

$$w(\mathbf{x}_n \rightarrow \mathbf{x}_{n'}) = w_0 p_{ap}(n \rightarrow n') \times \min \left[1, e^{-\frac{E' - E + U(s_i)' - U(s_i)}{k_B T}} \right], \quad (1)$$

that is, the probability with which the next Monte Carlo move is accepted is calculated on the variation of the energy of the system, plus the variation of the metadynamics potential.

3. Theory

During each fragment of trajectory after the update of the non-Markovian potential at each time T , the collective variable s explores a region $A(T)$. If one makes the critical assumption that the dynamics has been able to visit this region so extensively that ergodicity holds, then the probability distribution of the collective variable is

$$P(s, T) = \frac{\exp[-\beta(F(s) + U(s, T))]}{\int_{A(T)} ds' \exp[-\beta(F(s') + U(s', T))]} \cdot \quad (2)$$

After the end of this sampling, the non-Markovian potential is updated, and the new potential reads

$$U(s, T + \tau) = U(s, T) + W\tau P(s, T), \quad (3)$$

where W is the height of the energy added to the non-Markovian potential. Further assuming that W is small, that is that the new term does not perturb in an important way the shape of the potential U , then the previous equation can be rewritten as

$$\frac{dU(s, T)}{dT} = W \frac{\exp[-\beta(F(s) + U(s, T))]}{\int_{A(T)} ds' \exp[-\beta(F(s') + U(s', T))]} \cdot \quad (4)$$

Once the free energy landscape is completely filled by the non-Markovian energy, then the growth of this non-Markovian energy will be independent on s , that is $dU/dT = W/A$, or equivalently

$$\exp[-\beta(F(s) + U(s, \infty))] = \frac{1}{A} \int_A ds' \exp[-\beta(F(s') + U(s', \infty))], \quad (5)$$

where A is the whole interval spanned by the collective variable. Integrating by saddle-point evaluation, leads to

$$F(s) = -U(s, \infty) - \beta^{-1} \log \left(\frac{1}{A} \sqrt{\frac{2\pi}{\beta|F''(s_0) + U''(s_0)|}} \right) + F(s_0) + U(s_0), \quad (6)$$

where s_0 is defined by $U'(s_0) = -F'(s_0)$. The last equation states that the free energy of the system is, except

for an additive constant, equal to the opposite of the non-Markovian potential. A nice property of this algorithm is that the obtained free energy depends logarithmically on any additive error in the determination of $U(s, T + \tau)$ (i.e., if one adds $\epsilon(s)\tau$ to Eq. (4), one obtains an additive term $\log \epsilon(s)$ in $F(s)$).

4. Results

In order to obtain a reference free energy landscape as a function of the RMSD for comparison to the metadynamics reconstructed landscapes, we first carried out a fairly long classical Monte Carlo simulation (90 billions of Monte Carlo Steps (MCS)). The free energy calculated at temperature $\theta = 0.625$ (slightly below the folding temperature, defined as the temperature at which the volume of the native basin is equal to that of the denatured basin) is displayed in Fig. 3. After 80 billion steps the root mean square difference between the landscape at time T and at time $T - \Delta T$, where $\Delta T = 10000$ MCS, was constantly below 10^{-2} Å, indicating that the free energy is likely to have reached its equilibrium shape.

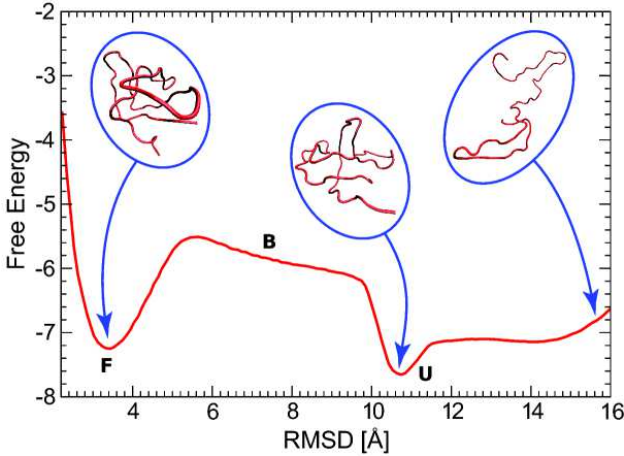


Fig. 3. Reference energy landscape as a function of the RMSD, obtained from a 90 billions of step long Monte Carlo simulation. Shown in the insets are a typical folded configuration (**F**), a compact unfolded one (**U**) and an elongated unfolded one (right).

The landscape presents a fairly broad barrier between the folded and the unfolded states (marked with **B** in Fig. 3). Also shown in Fig. 3 are a typical folded configuration (the shown configuration has a RMSD of 3.11 Å, less than the distance between two consecutive C_α in the protein sequence), a compact unfolded one, and an elongated unfolded configuration.

To be able to quantify the degree of convergence of the free energy landscapes reconstructed by Monte Carlo metadynamics, we calculate the standard deviation between the reconstructed landscape after T steps and the reference one

$$\sigma = \sqrt{\frac{1}{R_{max} - R_{min}} \int_{R_{min}}^{R_{max}} (f_{meta}(x) - f_{MC}(x))^2 dx}, \quad (7)$$

where $f_{meta}(x)$ and $f_{MC}(x)$ are the reconstructed and reference landscapes respectively, while R_{min} and R_{max} define the range of the RMSD over which σ is calculated. They are chosen so as to englobe in the calculation the most significant fraction of the landscape, the corresponding values being 2 and 16 Å, respectively. The reason why the edges of the landscape are not included in the calculation is that they are both noisy, as they are seldom visited by the system, aside from corresponding to high values of the free energy ($\gg k\theta$, where k is the Boltzmann constant and θ is the temperature), and consequently not interesting from the thermodynamic point of view.

The two collective variables used are the RMSD and the radius of gyration R_g . We then proceed to integrate out the radius of gyration

$$F(RMSD) = -\theta \log \int dR_g \exp \left[\frac{-F(RMSD, R_g)}{\theta} \right], \quad (8)$$

where θ is the simulation temperature, so as to have a simpler, one-dimensional, visualization of the free energy landscape. The simulations are carried out at different values of the height W of the Gaussian terms added to the non-Markovian potential and of their deposition time τ , with each Gaussian having a fixed standard deviation $\delta s = 0.25$.

In Fig. 4 we show the reconstructed free-energy landscape as a function of the RMSD at different values of the number of MCS elapsed in the simulation (with $W = 0.01$ and $\tau = 200000$). At the beginning (see the inset) the pro-

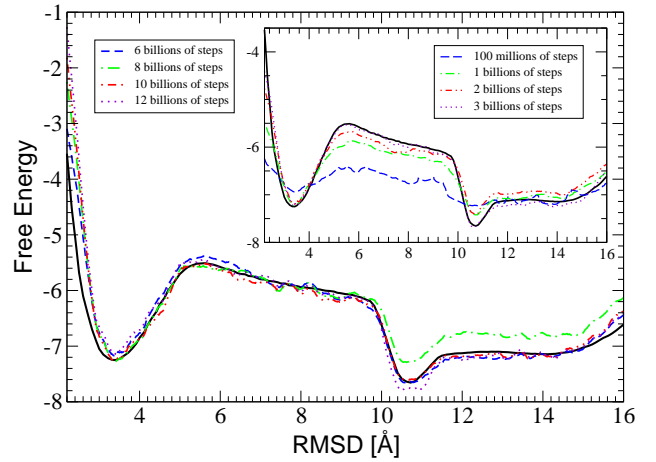


Fig. 4. Expected and reconstructed energy landscapes for a typical simulation. Shown in continuous line is the expected energy landscape. It can be appreciated how, after having reached the minimum value for σ , the reconstructed landscapes (here shown one every 2 billions of steps, starting from 6 billions) have small oscillations around the expected one. (Inset) Expected and reconstructed energy landscapes, ranging from the reconstructed landscape after about 100 millions of steps, to the one at the minimum σ , after 3 billions of steps.

tein explores mainly the regions around 3 and 11 Å, producing the two minima associated with the native and the denatured state. After these have been filled by the non-Markovian term, the rest of the landscape is refined and

converges to the reference one within 3×10^9 MCS (to be compared with the 8×10^{10} MCS needed by the standard Monte Carlo simulation).

The corresponding values of σ are displayed in Fig. 5 as a function of the number of MCS. The simulation reaches

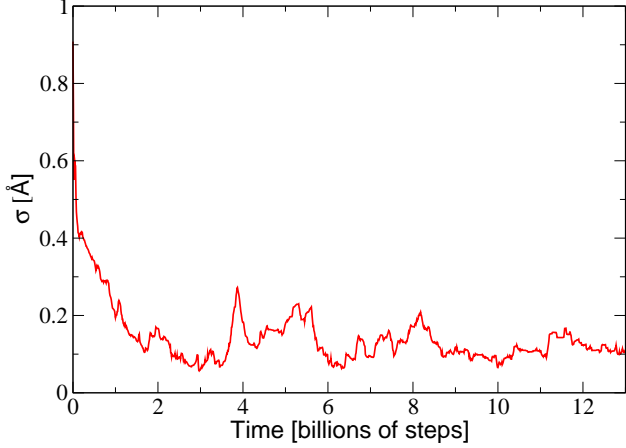


Fig. 5. σ as a function of time. It can easily be seen how it goes down really fast, and then starts to oscillate in a quite small zone.

a fairly low σ (under 0.2 \AA) in a few billions MCS, and then oscillates (with a spread of less than $\sim 0.15 \text{ \AA}$) around $\sim 0.1 \text{ \AA}$.

Fig. 6 shows the dependence of σ on W , at fixed $\tau = 200000$. The value of σ plotted here corresponds to the average of the last 2×10^6 MCS of the simulation (cf. Fig. 5). The plot shows a steep increase of σ with respect to the height of the hills (note the logarithmic axis scale), indicating that only a fine-grained deposition of the non-Markovian term is able to drive the system to equilibrium. This is consistent with the fact that Eq. (4) is derived by Eq. (3) as an expansion for small W , and consequently fails when W is increased.

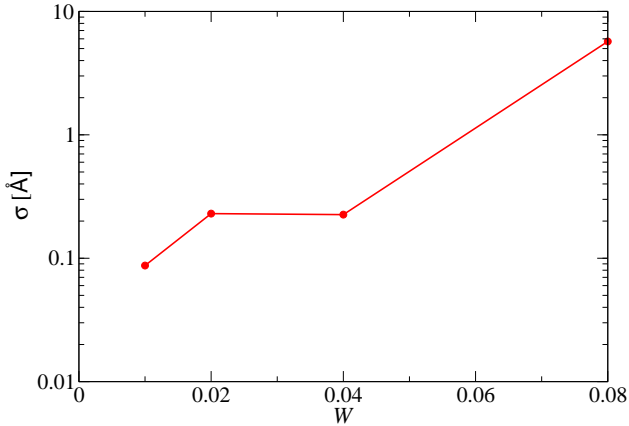


Fig. 6. σ as a function of W for a fixed τ .

In Fig. 7 the dependence of σ on W/τ is shown. It was shown by Laio and coworkers (Laio et al. (2005)) that the lower is W/τ , the higher the accuracy of standard metadynamics is. The data shown indicate that also in Monte Carlo metadynamics the accuracy of the reconstructed landscape

increases when W/τ is decreased. In particular, it seems that σ is related to this ratio by a linear function (the linear fit indicates a slope of 5.3×10^5 , with a correlation of 0.955).

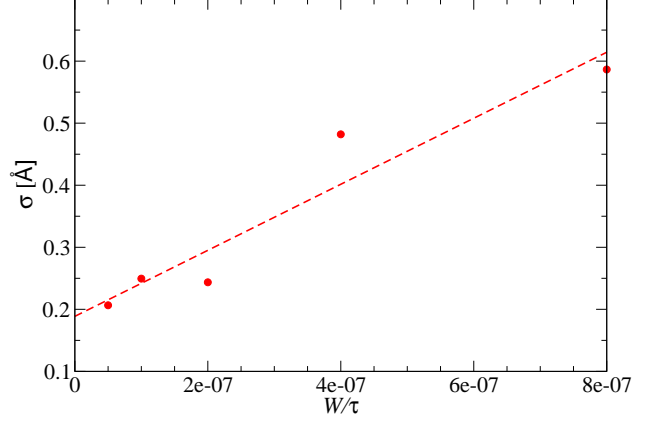


Fig. 7. σ as a function of W/τ . (dashed, red online) Linear fit, with intercept 0.189, slope 5.319×10^5 , standard deviation 0.170 and correlation coefficient 0.955.

Fig. 8 shows how σ varies for different values of τ , at fixed W . As τ increases, σ becomes smaller, as expected from the theoretical discussion carried out in Section 3. In fact, the larger is τ , the more likely it is for the system to having explored a region $A(T)$ exhaustively and thus for Eq. (2) to be a good approximation of the actual probability distribution.

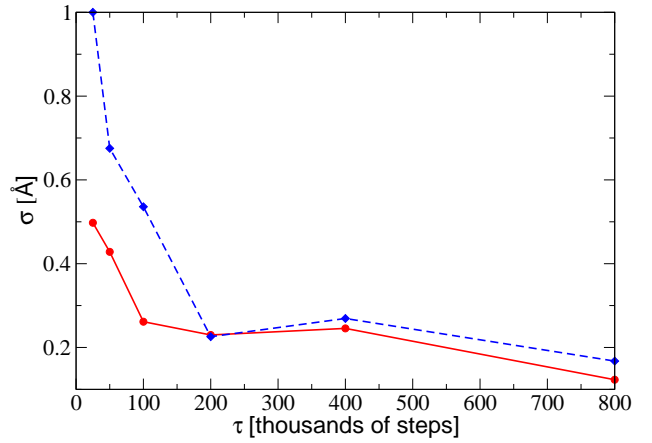


Fig. 8. σ as a function of τ for two different, fixed, W . In particular, $W = 0.02$, continuous line (red online) and $W = 0.04$, dashed line (blue online).

5. Conclusion

The metadynamics strategy has been implemented within a Monte Carlo scheme in order to take benefit from the positive aspects of both approaches. The algorithm is tested with a simplified protein model, and results particularly efficient and accurate in reconstructing the free energy landscape of the protein.

References

- Babin, V., Roland, C., Darden, T. A., Sagui, C., 2006. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *J. Chem. Phys.* 125 (20), 204909.
- Gō, N., 1975. Theory of reversible denaturation of globular proteins. *Int. J. Pept. Protein Res.* 7 (4), 313–323.
- Grantcharova, V., Baker, D., 1997. Folding dynamics of the src sh3 domain. *Biochemistry* 36 (50), 15685–92.
- Grantcharova, V. P., Riddle, D. S., Santiago, J. V., Baker, D., 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src sh3 domain. *Nat. Struct. Mol. Bio.* 5, 714–720.
- Kussell, E. L., Shimada, J., Shakhnovich, E. I., 2002. A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci.* 99 (8), 5343–5348.
- Laio, A., Parrinello, M., 2002. Escaping free energy minima. *Proc Natl Acad Sci USA* 99 (20), 12562–6.
- Laio, A., Rodriguez-Forte, A., Gervasio, F. L., Ceccarelli, M., Parrinello, M., 2005. Assessing the accuracy of metadynamics. *J. Phys. Chem. B* 109, 6714–6721.
- Metropolis, N., Rosenbluth, A. W., Teller, M. N., Teller, E., 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21 (6), 1087–1092.
- Metropolis, N., Ulam, S., 1949. The monte carlo method. *J. Amer. Statistical Assoc.* 44 (247), 335–341.
- Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., Baker, D., 1999. Experiment and theory highlight role of native state topology in sh3 folding. *Nat. Struct. Mol. Bio.* 6, 1016–1024.
- Shimada, J., Kussell, E. L., Shakhnovich, E. I., 2001. The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. *J. Mol. Bio.* 308 (1), 79–95.
- Shimada, J., Shakhnovich, E. I., 2002. The ensemble folding kinetics of protein g from an all-atom monte carlo simulation. *Proc. Natl. Acad. Sci.* 99 (17), 11175–11180.
- Yi, Q., Bystroff, C., Rajagopal, P., Klevit, R. E., Baker, D., 1998. Prediction and structural characterization of an independently folding substructure in the src sh3 domain. *J. Mol. Bio.* 283 (1), 293–300.