
Article:
Banushkina, PV and Krivov, SV (2015) Fep1d: A Script for the Analysis of Reaction Coordinates. Journal of Computational Chemistry, 36 (12). pp. 878-882. ISSN: 0192-8651

https://doi.org/10.1002/jcc.23868

Fep1d: a script for the analysis of reaction coordinates

Polina V. Banushkina and Sergei V. Krivov*

Astbury Center for Structural Molecular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom

E-mail: s.krivov@leeds.ac.uk

Abstract

The dynamics of complex systems with many degrees of freedom can be analyzed by projecting it onto one or few coordinates (collective variables). The dynamics is often described then as diffusion on a free energy landscape associated with the coordinates. Fep1d is a script for the analysis of such one-dimensional coordinates. The script allows one to construct conventional and cut-based free energy profiles, to assess the optimality of a reaction coordinate, to inspect whether the dynamics projected on the coordinate is diffusive, to transform (re-scale) the reaction coordinate to more convenient ones, and to compute such quantities as the mean first passage time, the transition path times, the coordinate dependent diffusion coefficient, etc. Here we describe the implemented functionality together with the underlying theoretical framework.

Keywords: reaction coordinate, free energy profile, diffusion, cut free energy profile, dimesionality reduction.

*To whom correspondence should be addressed

Introduction

With the growing number of experiments and the steady progress in the simulations of different biological, physical, chemical processes the rigorous quantitative analysis of the obtained data becomes all the more important. The complex nature of the processes and the large size of data sets generated by simulation and experiment make the development of automated analysis tools very important, practically and conceptually.

One popular method of such a data analysis is the projection of multidimensional coordinate space to a one dimensional reaction coordinate (RC). During such a dimensionality reduction some information is inevitably lost. The optimal RC should absorb all the important dynamical information contained in the other degrees of freedom, so they can be neglected. Then the free energy profile (FEP) constructed along the optimal coordinate together with the determined diffusion coefficient can give an accurate description of the system dynamics as diffusion on this free energy surface. Examples of such dimensionality reduction can be found in different fields: order parameters in physics,[1,2] physicaly based reaction coordinate in single molecular experiments[3,4] and molecular dynamics simulations,[5–11] biomarkers in medicine,[12] chess game.[13]

In protein folding simulations the following RCs are frequently chosen: the root mean square distance from the native structure, the number of native contacts, the radius of gyration. In many cases, however, such coordinates can hide the complexity of the dynamics,[2,14] in particular the barriers between states and make the dynamics subdiffusive.[15] A number of methods to construct good, more sophisticated reaction coordinate have been suggested[8,16–22] and the analysis of the determined RC becomes as important as the description of the dynamics.

The following problems/questions often appears and are to be solved during such an analysis: determination of the associated free energy profile and the diffusion coefficient, which provide complete description of diffusive dynamics; establishing that the reaction coordinate is optimal and/or that the dynamics projected on such a coordinate is diffusive,

rather than sub or super diffusive; transformation of the putative reaction coordinate to another, more convenient one, for example, to the one with the constant diffusion coefficient; computation of various properties associated with stochastic dynamics, such as, the mean first passage time (mfpt), the mean transition path times (mtpt).

While some of the tasks are rather trivial (such as the computation of the conventional histogram free energy profile), other (such as the computation of the cut profile using transition path segments with different sampling intervals[23]) are not. Here we describe a python script fep1d.py, which allows one to readily perform most of the described tasks. Being a python script, the code can be easily read, adapted and modified to extend the functionality. Below we provide the summary of the implemented functionality together with the underlying theoretical framework. The fep1d.py program can be downloaded from http://sourceforge.net/projects/fep1d/

Fep1d

The command line

The command line contains the list of the data files with the RC time series (the mandatory arguments), and a list of optional arguments starting with --. For example,

fep1d.py rc1.dat rc2.dat --transformto=natural --hfep=1

will transform two different reaction coordinates rc1 and rc2 to the "natural" reaction coordinates with the constant diffusion coefficient (optional argument *transformto*) and will construct and plot the histogram based free energy profiles (the optional argument *hfep*) along the transformed RCs. For a protein folding trajectory these can be the root mean square distance from the native structure and the number of native contacts reaction coordinates. The command then can be used to compare which of the two putative coordinates is more optimal or better describes the dynamics: it is the one with the highest profile.

The script, invoked without arguments, lists all the options. The gnuplot library is used

to plot the graphs. The plots can be saved as postscripts files, or as data files to be plotted using other software. The transformed coordinates can be saved as well.

Construction of the profiles

Below we denote a time series of reaction coordinate x as $x(i\Delta t)$, while a point on this coordinate is denoted simply as x. Note, that while the profiles depend on sampling interval Δt, we omit this dependence in notation, i.e. we write $F_C(x)$ rather than $F_C(x, \Delta t)$.

hfep - the histogram based free energy profile $F_H(x)$ along the reaction coordinate x is calculated in conventional way as $F_H(x)/k_BT = -\ln(Z_H(x))$, where k_B is the Boltzmann constant, T is the temperature, and $Z_H(x)$ is the partition function estimated as $Z_H(x) = N_x/\Delta x$. N_x is the number of time-series points in bin x and Δx is the size of the bin which can be specified as *--dx=0.01*.

cfep - cut based free energy profile $F_C(x)$ is calculated as $F_C(x)/k_BT = -\ln(Z_C(x))$, where $Z_C(x)$ is the corresponding partition function defined as half the total number of transitions through point x[15]

$$Z_C(x) = 1/2 \sum_i \theta[-(x(i\Delta t) - x)(x(i\Delta t + \Delta t) - x)],$$

where $\theta[x]$ is the Heaviside step function and Δt is the time interval. The important property of the F_C is its invariance (in contrast to the conventional F_H) with respect to an arbitrary monotonous transformation of the reaction coordinate which makes possible partition of configuration space in an invariant way.[20]

cfep1 - the cut based free energy profile $F_{C,1}(x)$ is constructed using generalized partition function $Z_{C,1}(x)$ which is calculated by summing up the transitions through points x with weights equal to the length of the transition[23]

$$Z_{C,1}(x) = 1/2 \sum_i |x(i\Delta t) - x(i\Delta t + \Delta t)| \theta[-(x(i\Delta t) - x)(x(i\Delta t + \Delta t) - x)] \qquad (1)$$

$Z_C(x)$ and $Z_{C,1}(x)$ can be related to $Z_H(x)$ and $D(x)$ (the position dependent diffusion coefficient) as[20,22]

$$Z_C(x) = \sqrt{D(x)\Delta t/\pi} Z_H(x), \tag{2}$$

and

$$Z_{C,1}(x) = \Delta t D(x) Z_H(x). \tag{3}$$

The relations are derived using the following assumptions. First, the displacements are Gaussian distributed $P(\Delta x) \sim \exp(-\Delta x^2/2D\Delta t)$, i.e., the dynamics is diffusive. Second, $Z_H(x)$ and $D(x)$ are assumed to be approximately constant on the average displacement distance of $\sqrt{2D\Delta t}$ during Δt. The assumptions are generally valid for relatively small Δt and/or around the transitions states and local minima. Corrections due to a nonzero free energy gradient are of the second order.[13]

If the reaction coordinate is rescaled to the natural coordinate,[20] defined as the one with the diffusion coefficient $D(x) = 1$, then all the three profiles along the rescaled coordinate are differ only by a constant. However, the cut profile $F_C(x)$ is less prone to statistical noise and hence, is constructed by default by fep1d script. It should also be noted that all the profiles are reported in dimensionless units as $F_H(x)/k_BT$, $F_C(x)/k_BT$ and $F_{C,1}(x)/k_BT$. Fig. 1 shows two FEPs along an optimized RC constructed for the folding trajectory of FIP35 protein[6,21] and transformed to the natural coordinate with the constant diffusion coefficient. All examples below are given for this coordinate which is denoted as RC_{FIP35}.

Transformations of the RC

The reaction coordinate can be transformed to the one of several coordinate types listed below by adding argument *--transformto=type-of-RC* to the command line. The transformation can be performed "on-the-fly" (before determining the profiles), or the transformed coordinate time series can be saved to a file for further analysis.

natural coordinate. A reaction coordinate x with a variable diffusion coefficient can be

transformed to coordinate y, called the natural coordinate[20] by numerically integrating

$$y = \pi^{-1/2} \int_{-\infty}^{x} Z_H(w)dw/Z_C(w)$$

i.e. that $Z_C(y) = \pi^{-1/2} Z_H(y)$, so that $D(y) = 1$.[15] In this case the dynamics is completely specified by the FEP.

Za coordinate. $Z_A(x)$ is the reaction coordinate which measures the relative partition function of the coordinate segment between -∞ and x

$$y = Z_A(x) = \int_{-\infty}^{x} Z_H(w)dw/ \int_{-\infty}^{\infty} Z_H(w)dw$$

$Z_A(x)$ is invariant to coordinate transformation, and can be used to compare different coordinates.[15]

pfold coordinate. The folding probability (pfold) equals to the probability of the system being at position x to reach one (boundary) state x_1 before it reaches another state x_0.[24] Assuming that dynamics is diffusive, the folding probability is computed using partition functions of cfep and hfep[21]

$$y = p_{fold}(x) = \frac{\int_{x_0}^{x} Z_C^{-2}(x) Z_H(x)dx}{\int_{x_0}^{x_1} Z_C^{-2}(x) Z_H(x)dx} \tag{4}$$

pfoldMSM coordinate. If the assumption that the dynamics is diffusive and $Z_H(x)$ is relatively flat are not valid while the dynamics is Markovian, the transformation to the pfold coordinate is computed by using a Markov state model. The model is obtained by clustering the initial reaction coordinate into the bins of size Δx and computing the transition probabilities as $p_{ji} = n_{ji}/\sum_j n_{ji}$, where n_{ji} is the number of transitions from bin i to bin j. The folding probabilities then can be determined by solving the system of equations[13]

$$y_i = p_{fold}(x_i) = \sum_j p_{ji} p_{fold}(x_j), \tag{5}$$

with the boundary conditions $p_{fold}(x_0) = 0$, $p_{fold}(x_1) = 1$.

For pfold and pfoldMSM transformations x_0 and x_1 should be specified as *--x0=value-of-*x_0 and *--x1=value-of-*x_1. These values usually correspond to the positions of the minima of the FEP. Optimal value of Δx used for reaction coordinate discretization to construct a Markov state model should be choosen neither too small (leads to a very large model with poor statistics) nor too large (leads to a very coarse-grained model).

Optimality of the RC

Diffusivity of the dynamics. It is often assumed that dynamics along a putative coordinate is diffusive. In this case the free energy profile $F(x)$ determined as a function of the RC gives the equilibrium probabilities $p(x) \approx \exp(-F(x)/k_BT)$ and together with the position dependent diffusion coefficient $D(x)$ can provide a quantitative description of the dynamics as a diffusion on the free energy surface. However, dynamics is diffusive when projected on an optimal reaction coordinate and is sub-diffusive when projected on a sub-optimal coordinate.[23] In the latter case the dynamics should be described by a generalized Langevin equation with a memory kernel, indicating non-Markovian dynamics. An accurate determination of the kernel is a highly non-trivial task.[25]

Fep1d can be used to test the diffusivity of the dynamics along a putative RC. Diffusive dynamics is characterized by the mean square displacement growing linearly with time $\langle \Delta x^2(\Delta t) \rangle = 2D\Delta t \sim \Delta t^{2\alpha}$, where D is the diffusion coefficient and $\alpha = 0.5$. For sub-diffusive dynamics $\alpha < 0.5$. The coordinate dependent exponent $\alpha(x)$ can be determined from the distance between two cut profiles $F_C(x, \Delta t)$ computed at two different sampling intervals Δt_1 and Δt_2[15] as

$$\alpha(x) = 1 + \frac{F_C(x, \Delta t_1)/k_BT - F_C(x, \Delta t_2)/k_BT}{\ln \Delta t_1 - \ln \Delta t_2}.$$

Fig. 2 shows $F_C(x)$ and calculated $\alpha(x)$ along the RC_{FIP35}. $\alpha(x)$ is equal 0.5 around the

both barriers that indicates diffusive dynamics in these regions.

Test of the RC optimality with the Bayesian criterion. A Bayesian criterion has been suggested to quantify the quality of an RC by calculating the probability $P(TP|x)$ of being on the transition path (TP)[8,26]

$$p(TP|X) = p(X|TP)p(TP)/p_{eq}(X),$$

where $p(X|TP)$ is the probability density of X on the TPs, $p(TP)$ is the fraction of time spent on the TPs, and $p_{eq}(X)$ is the equilibrium probability density at point X of configuration space. For diffusive dynamics, the maximum of $p(TP|X) = 0.5$ is attained at the points of the stochastic separatrix $p_{fold}(X) = 0.5$.[26] For a good reaction coordinate $x = R(X)$, $p(TP|x)$ should have a single sharp and high peak, collapsing the transition states with high values of $p(TP|X)$ into a single value of x.[8] The stochastic separatrix is a reasonable definition of the transition state ensemble for the systems with one dominant barrier. For more complex systems, e.g., systems with two approximately equal barriers,[21] where $p_{fold}(X) = 0.5$ describes an intermediate state, rather than transitions states, the criterion is not a sufficient condition of optimality and a different criteria should be used. To calculate $P(TP|x)$ the argument *ptpx* should be added to command line together with the positions of the boundaries x_0 and x_1 and the bin size Δx. Fig. 3 (A) shows that when the $P(TP|x)$ criteria applied to both transition states, it indicates that neither of the transition states are optimal. However, it shows that the first transition state is optimal, if $P(TP|x)$ is computed just around a single TS (Fig. 3 B).

The $F_{C,1}(x, \Delta t)$=const test of RC optimality. To test the optimality of a putative RC for more complex systems (i.e., with more than one barrier) another criterion has been suggested recently.[23] The criterion assesses whether a putative reaction coordinate accurately approximates the p_{fold} coordinate (which is considered to be an optimal coordinate[23,27,28]).

It states that for such a coordinate, the cut-based free energy profile $F_{C,1}(x, \Delta t)$, computed from an ensemble of transition path segments, is x and Δt independent.

The function *testoptimality* first transforms the reaction coordinate (the time series) x to the p_{fold} coordinate, by invoking the **pfoldMSM** internal function (Eq. 5). Then cut profiles $F_{C,1}$ along this coordinate are constructed at different sampling time intervals Δt=1, 2, 4, 8, 16, ... using the transition paths.[23] Fig. 4 shows cfep1 profiles constructed along $p_{foldMSM}(RC_{FIP35})$ with different Δt. All the profiles are relatively parallel lines with small fluctuations converging to the limiting profile, which means that chosen RC is close to the p_{fold} coordinate, however is still slightly sub-optimal.[21,23]

Diffusion coefficient and other quantities

Fep1d allows one to compute the diffusion coefficient according to Eq. 2 or according to Eq. 3 by adding arguments *--D=1* or *--D=2*, correspondingly.

The mean first passage time can be estimated (a) from the trajectory by estimating the mean waiting time between the transitions and (b) with the Kramer's equation[21,29]

$$mfpt = \int_{x_0}^{x_1} e^{\beta F_H(x)}/D(x)dx \int_{x}^{x_1} e^{-\beta F_H(y)}dy,$$

where $\beta = 1/k_BT$. The mean transition pass times mtpt is estimated by direct counting from trajectories. For computing these quantities the only argument *--mfpt=1* should be added.

The plots of folding probability as a function of the reaction coordinate, estimated using either Eq. 4 or Eq. 5, can by obtained by using command line arguments *--pfold=1* or *--pfoldMSM=1*.

Conclusion

The popular approaches to analyze the results of atomistic simulations can be roughly divided into Markov state models (MSMs),[31] conformation network analysis[30] and the free-energy landscape framework.[2] In the former, trajectory snapshots are combined into microstates or clusters using a variety of clustering algorithms. Construction of such a model, which accurately describes the dynamics encoded by the trajectory is a very difficult problem and a number of software packages are being developed for its solution.[32,33] Once such a model has been constructed, the properties of the dynamics are inferred by analyzing the model. In particular, one can determine the free energy basins (metastable states), transition states, pathways.[2,30,31,34] The p_{fold} optimal coordinate between any two boundary nodes can be easily determined, and hence such properties as the cut profiles and diffusion coefficient. In particular, the popular wordom program provides such a workflow.[33] In our experience, such an approach works very well for relatively small systems (e.g., peptides with up to 15-20 residues in length).[2,20,35] For larger systems, it appears that construction of a good quality Markov state model, which accurately describes the dynamics of a system, is much more difficult due to the infamous "curse of dimensionality"; especially in the transition state regions, where the sampling is limited. Because of that, it is difficult to determine, for example, the major properties of the folding dynamics - the folding free-energy barriers and the pre-exponential factor, the structure of the transition states (TS), the transition path times and the diffusion coefficient. The free energy profiles as functions of optimal reaction coordinates constitute an alternative framework, which allows one to determine these properties in relatively straightforward manner.

Fep1d is a simple tool for the analysis of such reaction coordinates and the projected dynamics. Via command line options Fep1d allows one to determine free energy and cut based free energy profiles; to transform a putative reaction coordinate to another, more convenient one, for example, to the one with the constant diffusion coefficient; to compute various properties associated with stochastic dynamics, such as, diffusion coefficient, the

mean first passage times, the mean transition path times. However, the unique functionality of the script is the ability to assess the optimality of a putative reaction coordinate and to inspect whether the dynamics projected on such a coordinate is diffusive, rather than sub or super diffusive. Being a python script, the code can be easily read, adapted and modified to extend the functionality. We hope that it will make a rigorous analysis of increasingly available large amount of complex data less involved.

ACKNOWLEDGMENTS

The work has been supported in part by an RCUK fellowship and BBSRC grant (No. BB/J016055/1).

References

(1) Hedges, L. O.; Jack, R. L.; Garrahan, J. P.; Chandler, D. *Science* **2009**, *323*, 1309–1313.

(2) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. USA.* **2004**, *101*, 14766–14770.

(3) Yu, H.; Gupta, A. N.; Liu, X.; Neupane, K.; Brigley, A. M.; Sosova, I.; Woodside, M. T. *Proc. Natl. Acad. Sci. USA.* **2012**, *109*, 14452–14457 .

(4) Schuetz, P.; Wuttke, R.; Schuler, B.; Caflisch, A. *J. Phys. Chem. B* **2010**, *114*, 15227–15235.

(5) Jungblut, S.; Singraber, A.; Dellago, C. *Mol. Phys.* **2013**, *111*, 3527–3533.

(6) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

(7) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. USA.* **2000**, *97*, 5877–5882.

(8) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. USA*. **2005**, *102*, 6732–6737.

(9) Berezovska, G.; Prada-Gracia, D.; Rao, F. *J. Chem. Phys.* **2013**, *139*, 035102.

(10) Ceriotti, M.; Tribello, G. A.; Parrinello, M. *Proc. Natl. Acad. Sci. USA*. **2011**, *108*, 13023–13028.

(11) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins: Structure, Function, and Bioinformatics* **2005**, *58*, 45–52.

(12) Krivov, S. V.; Fenton, H.; Goldsmith, P. J.; Prasad, R. K.; Fisher, J.; Paci, E. *PLoS Comput. Biol.* **2014**, *10*, e1003685.

(13) Krivov, S. V. *Phys. Rev. E* **2011**, *84*, 011135.

(14) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751–758 .

(15) Krivov, S. V. *PLoS Comput. Biol.* **2010**, *6*, e1000921.

(16) Ferguson, A. L.; Panagiotopoulos, A. Z.; Kevrekidis, I. G.; Debenedetti, P. G. *Chem. Phys. Lett.* **2011**, *509*, 1–11.

(17) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*, 124116.

(18) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.

(19) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.

(20) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. USA*. **2008**, *105*, 13841–13846 .

(21) Krivov, S. V. *J. Phys. Chem. B.* **2011**, *115*, 12315–12324 .

(22) Krivov, S. V. *J. Phys. Chem. B.* **2011**, *115*, 11382–11388 .

(23) Krivov, S. V. *J. Chem. Theory Comput.* **2013**, *9*, 135–146.

(24) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.

(25) Darve, E.; Solomon, J.; Kia, A. *Proc. Natl. Acad. Sci. USA.* **2009**, *106*, 10884–10889.

(26) Hummer, G. *J. Chem. Phys.* **2004**, *120*, 516–523.

(27) Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102–174102–13.

(28) Berezhkovskii, A. M.; Szabo, A. *J. Phys. Chem. B* **2013**, *117*, 13115–13119.

(29) Kramers, H. A. *Physica* **1940**, *7*, 284–304.

(30) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.

(31) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413–18419.

(32) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

(33) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625–2627.

(34) No, F.; Schtte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. USA.* **2009**, *106*, 19011–19016.

(35) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.

Figure captions

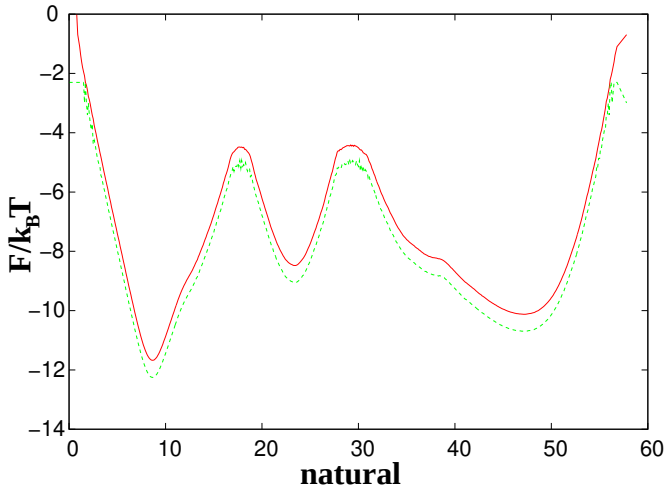

Figure 1: Free energy profiles $F_C(x)$ (red solid line) and $F_H(x)$ (green dashed line) determined for an optimized RC_{FIP35} constructed for the folding trajectory of FIP35 protein using command *fep1d.py rc.dat --transformto=natural --hfep=1*.

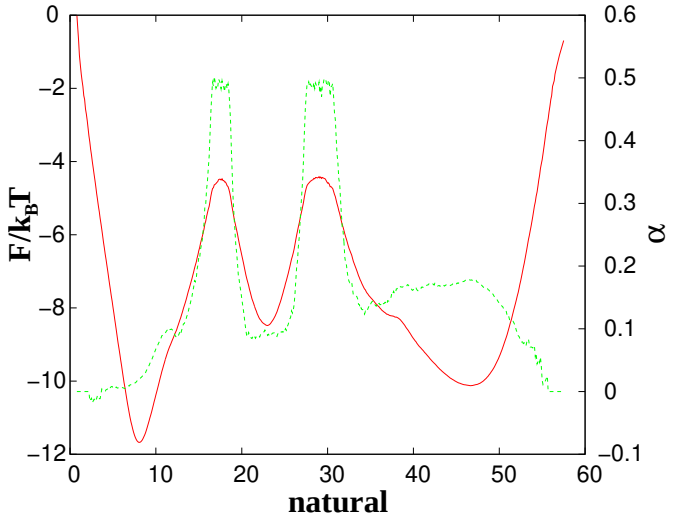

Figure 2: $F_C(x)$ (red solid line) and $\alpha(x)$ (green dashed line) determined for the RC_{FIP35}. To calculate $\alpha(x)$, the command line should contain *--alpha=* $[\Delta t_1, \Delta t_2]$. Here, by the time steps Δt_1 and Δt_2 we mean integer numbers referred to the steps of the trajectory: *fep1d.py rc.dat --alpha=[1,8]*

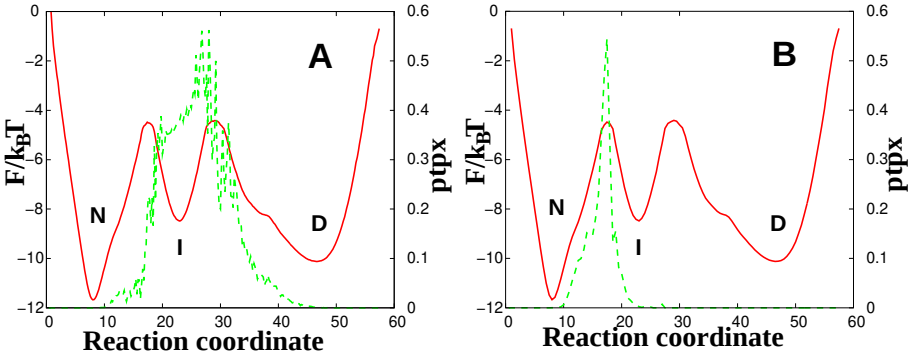

Figure 3: A) $P(TP|x)$ calculated for the RC_{FIP35} (green dashed lines). The command line is *fep1d rc.dat --ptpx=1 --x0=47 --x1=8.2 --dx=1.0* with x_0 and x_1, corresponding to the minima of the denatured state (D) and the native state (N); B) $P(TP|x)$ calculated with x_0 and x_1 taken in the intermediate state (I) and the native (N) state, correspondingly.

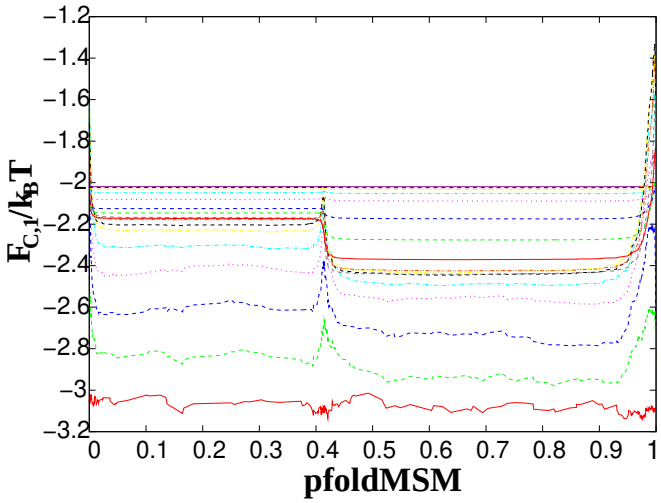

Figure 4: Testoptimality of RC_{FIP35}. $F_{C,1}$ are shown for different sampling intervals $\Delta t = 1$ (red line at the bottom),2, ...,2^{16} (magenta line at the top). The command line is *fep1d rc.dat --testoptimality=1 --x0=8.2 --x1=47*. For function *testoptimality* the positions of x_0 and x_1 should be given from left to right.