Why are MD simulated protein folding times wrong?

Dmitry Nerukh
Unilever Centre for Molecular Sciences Informatics,
Department of Chemistry, Cambridge University,
Cambridge CB2 1EW, UK
dn232@cam.ac.uk

The question of significant deviations of protein folding times simulated using molecular dynamics from experimental values is investigated. It is shown that, in the framework of Markov State Model describing the conformational dynamics of peptides and proteins, the folding time is very sensitive to the simulation model parameters, such as forcefield and temperature. Using two peptides as examples we show that the deviations in the folding times can reach an order of magnitude for modest variations of the molecular model. We, therefore, conclude that the folding rate values obtained in molecular dynamics simulations have to be treated with care.

Modern computational power is enough to simulate small proteins up to the times when they fold into their native conformations [1–4]. This is a remarkable achievement because phenomenological, relatively simple interatomic interactions built into the model lead to the molecular structures that essentially coincide with the crystallographically determined native conformations.

In contrast to the structure of proteins, the results on folding times are not as optimistic. For the majority of successfully folded proteins there are significant discrepancies between simulated and experimental folding times [5–7]. This is taking into account that only the results when the trajectories approach the folded conformations sufficiently close are published. In few cases even complete failures to reach the folded state in silico in simulations significantly exceeding the experimental folding times are reported [8]. Indeed it is well known in the modelling community how difficult it is no fold a protein ab initio, that is without introducing any information on the intermediates.

By analysing the MD trajectories of peptides in explicit water we suggest an explanation for these discrepancies. We show that the folding rates are very sensitive to the details of the simulation model. The sensitivity is so high that the obtained values of folding times are meaningless and can not be compared between each other and with the experiment.

We use the Markov State Model (MSM) [9–12] to describe the folding process. The configurational states are defined by clustering the MD simulated trajectories. This is done by analysing the Ramachandran plots of the residues of the peptide, Fig. 1. Each Ramachandran plot is clustered independently and the molecule's configurations are defined by the cluster indices from each plot. Not all possible combinations of index values are realised in the trajectory. For example, for the peptide from Fig. 1 the conformation B_1C_2 was very scarcely populated and was, therefore, joined with A_1C_2 into one conformation, thus resulting in 5 total configurations of the molecule.

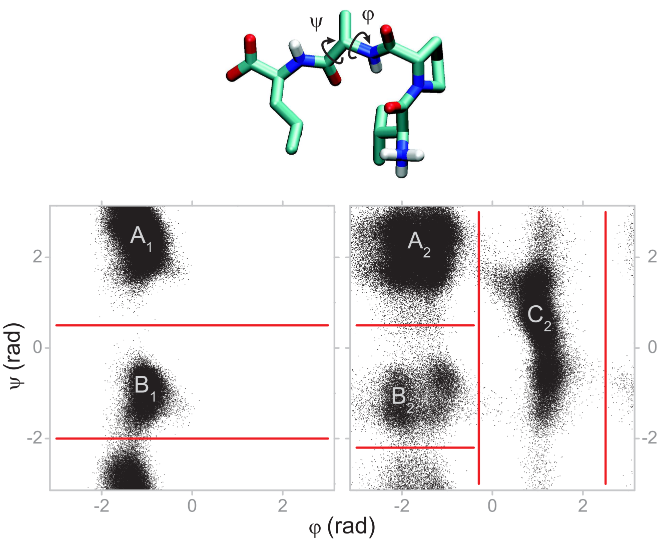

FIG. 1: Four residues peptide VPAL and the Ramachandran plots for the Proline (left) and Alanine (right). The clustering is marked by the boundaries that define the conformations as pairwise combinations of the indices from the sets $\{A_1, B_1\}$ and $\{A_2, B_2, C_2\}$

In the MSM framework the model is described by a state vector v, which holds probabilities of all the configurations at a given moment of time, and a transition matrix T. The total probability of the state vector has to sum to 100% since the peptide has to be in some configuration at any time. The transition matrix holds the probability that the system is transferred from one state to another at the next time step. Because the total probability of the state vector has to be conserved the requirement $\sum_i T_{ij} = 1$ is imposed, where i and j runs over all states. Given that the system has a state vector v_t at time t, the state vector at the time $t + \Delta t$ can be calculated as $v_{t+\Delta t} = Tv_t$. The property of a transition matrix is such that its eigenvalues λ_i are in the range from 0 to 1 with one eigenvalue being 1. In the following it is assumed that the eigenvalues are ordered in descending order so that $\lambda_0 = 1$.

By expanding the transition matrix in terms of the left $|\lambda_i\rangle$ and right $\langle\lambda_i|$ eigenvectors the time evolution of the system is given by

$$v_{t+n\Delta t} = T^n v_t = \sum_i \lambda_i^n |\lambda_i\rangle\langle\lambda_i| v_t. \qquad (1)$$

This representation immediately provides information on the behaviour of the system under investigation.

First, by analysing how the eigenvalues vary with time step Δt it can be determined whether the dynamics of the system can actually be described by a Markov model [13]. We have shown [13] that the dynamics of a four residue peptide become Markovian (that is the next time step conformation depends on the current conformation only) at the time scale $\Delta t \approx 50$ps. For a larger peptide of 15 residues our recent estimations confirm the same time scale of Markovian behaviour. From general considerations it is reasonable to assume the same time period of "loosing memory" for the dynamics of larger peptides and small proteins.

Second, both the folded state and the folding time are readily obtained from (1). Indeed, at the limit $n \to \infty$ only the largest eigenvalue, equal to 1, survives while all other eigenvalues, being less then 1 tend to zero. Therefore, the eigenvector $|\lambda_0\rangle$ corresponds to the equilibrium distribution of conformations, the folded state. The speed at which the system approaches the equilibrium distribution is described by all other eigenvalues that are less than 1. Again, at $n \to \infty$ the second largest eigenvalue dominates since it describes the slowest convergence in the system, while all smaller valued eigenvalues become negligible. Thus at this limit λ_1 defines the folding rate.

The transition matrix T can vary in the simulation due to, for example, the differences in the forcefield or the variation in the macroscopic parameters of the system: temperature, simulation box size (number of molecules), etc. The forcefield differences are the result of phenomenological nature of the classical molecular dynamics potentials as well as different calibration criteria. The other source of the changes are various alterations of the interaction potentials that are aimed at speeding up the folding and become increasingly popular lately [14–17]. Similarly, temperature variations is the cornerstone of such widespread technique as Replica Exchange MD. These too can affect the transitions T.

Our goal is to investigate what happens to the folding rate if the transition matrix is changed. In our framework it means that the effect of the variation of T to λ_1 has to be determined. Here we assume that the matrix changes do not affect the folded state, that is the eigenvector $|\lambda_0\rangle$ remains the same [19].

To understand the meaning of the λ_1 variation, $\delta\lambda_1$, in more intuitive terms of folding times we introduce a "folding half time" measure as follows. Let us assume that after a large number of time steps, n, λ_1 reduces its value in half, Eq. (1): $\frac{\lambda_1^n}{\lambda_1} = \frac{1}{2}$. Then, we designate this time as a "folding half time" that can be calculated as

$$n_{1/2} = \frac{\ln 2}{\ln \lambda_1}. \qquad (2)$$

Suppose that we have changed the dynamics and, as a result, the eigenvalue λ_1 has changed by an amount $\delta\lambda_1$. The half time for this new eigenvalue is $n'_{1/2} = \frac{\ln 2}{\ln(\lambda_1 - \delta\lambda_1)}$ (for an accelerated folding $\delta\lambda_1$ is negative). The ratio $r = \frac{n_{1/2}}{n'_{1/2}}$ gives us a representative measure of the sensitivity of the folding time to the changes in the transition matrix T. In other words r is an amount by which the folding time is changed. Thus, the sensitivity in our description is a function of two parameters: the second largest eigenvalue λ_1 and its variations $\delta\lambda_1$ caused by the changes in the simulation model.

The rest of the paper is devoted to the estimations of the values of λ_1 and $\delta\lambda_1$ for representative protein systems simulated using MD.

Both parameters can be calculated directly from the transition matrices T obtained in the simulations of systems with varying parameters described above. We modelled the variations by performing the simulations of peptides and altering two parameters of the system: (i) scaling the masses of the atoms (corresponds to changing the forcefield) and (ii) varying the temperature in the simulation (resembles the Replica Exchange MD conditions). The variation in the masses was done by the introduction of a unified parameter α so that the new masses are αm: $(\alpha m)a = -\frac{\delta V}{\delta r}$ or $ma = -\frac{\delta \frac{V}{\alpha}}{\delta r}$. Therefore, varying the masses is equivalent to varying the potential energy, that is changing the forcefield of the model. The changes in temperature were achieved by standard methods simply setting the thermostat to different temperatures.

We have simulated exhaustively a four residue peptide VPAL (Valine - Proline - Alanine - Leucine) and calculated both λ_1 and $\delta\lambda_1$ directly from the transition matrices varying the parameters in both ways described above. We have obtained the value of λ_1 equal to 0.785. For $\delta\lambda_1$, VPAL produces the values of 0.073 when varying the scaling α in the range 0.75 - 1.25 and 0.161 when varying the temperature in the 280 - 320K boundaries. These correspond to the ratios $r = 1.40$ for varying α and $r = 1.95$ for varying T. In other words, the folding times became almost twice as high in the different scenarios used.

It should be noted that for longer peptides λ_1 is normally larger. This is not surprising since larger peptides have lower folding rate, that is larger λ_1. Indeed, we have also analysed the folding trajectories of a fifteen residue peptide with the sequence SESYIDPDGTWTVTE and obtained $\lambda_1 = 0.9915$. Assuming approximately the same value for $\delta\lambda_1$, say $\delta\lambda_1 = 0.1$, we obtain $r = 13.45$, that is more than an order of magnitude increase in folding half time.

These results clearly demonstrate the high sensitivity of the folding times to the details of simulation models. It also seems reasonable to conclude that the sensitivity tends to be higher for larger peptides that fold slower. Therefore, the results on the folding times for larger realistic proteins would be even less reliable.

Since any force field is only approximately correct this means that calculated folding times are significantly inaccurate, even though the folded state reached in the simulation is correct. It is therefore not meaningful to make a comparison between a simulated folding time and the one determined experimentally especially for slowly folding proteins.

Acknowledgement The work is supported by Unilever and the European Commission (EC Contract Number 012835 - EMBIO).

[1] H. Lei and Y. Duan, Journal of Physical Chemistry B **111**, 5458 (2007), ISSN 1520-6106.

[2] H. Lei and Y. Duan, Journal of Molecular Biology **370**, 196 (2007).

[3] A. Suenaga, T. Narumi, N. Futatsugi, R. Yanai, Y. Ohno, N. Okimoto, and M. Taiji, Chemistry - An Asian Journal **2**, 591 (2007), 10.1002/asia.200600385.

[4] S. Gnanakaran, H. Nymeyer, J. Portman, K. Y. Sanbonmatsu, and A. E. Garca, Current Opinion in Structural Biology **13**, 168 (2003).

[5] J. Kubelka, J. Hofrichter, and W. A. Eaton, Current Opinion in Structural Biology **14**, 76 (2004).

[6] D. L. Ensign, P. M. Kasson, and V. S. Pande, J. Mol. Biol. **374**, 806 (2007).

[7] L. Tsai, H. Chen, T. Lin, W. Wang, and Y. Sun, J. Theor. Comput. Chem. **6**, 213 (2007).

[8] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, Biophys. J. **94**, L75 (2008).

[9] C. Schuette, A. Fischer, W. Huisinga, and P. Deuflhard, Journal of Computational Physics **151**, 146 (1999), ISSN 0021-9991.

[10] W. C. Swope, J. W. Pitera, and F. Suits, The Journal of Physical Chemistry B **108**, 6571 (2004).

[11] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, The Journal of Chemical Physics **126**, 155101 (pages 17) (2007).

[12] F. Noe and S. Fischer, Current Opinion in Structural Biology **18**, 154 (2008), ISSN 0959-440X.

[13] C. H. Jensen, D. Nerukh, and R. C. Glen, The Journal of Chemical Physics **128**, 115107 (2008).

[14] Y. Sugita and Y. Okamoto, Chemical Physics Letters **314**, 141 (1999).

[15] X. Periole and A. E. Mark, Journal of Chemical Physics **126**, 11 (2007).

[16] A. Baumketner and J. E. Shea, Theoretical Chemistry Accounts **116**, 262 (2006).

[17] K. P. Ravindranathan, E. Gallicchio, R. A. Friesner, A. E. McDermott, and R. M. Levy, Journal of the American Chemical Society **128**, 5786 (2006).

[18] C. H. Jensen, D. Nerukh, and R. C. Glen, The Journal of Chemical Physics **129**, 225102 (pages 6) (2008).

[19] To the best of our knowledge this requirement has never been checked in numerous methods aiming to accelerate the folding in MD (Accelerated Molecular Dynamics, Hyperdynamics, Replica-Exchange Molecular Dynamics, etc [14–17]). This almost inevitably leads to incorrect folded state obtained in these simulations (see [18] for details).