Sensitivity of peptide conformational dynamics on clustering of a classical molecular dynamics trajectory

Christian H. Jensen, Dmitry Nerukh and Robert C. Glen

December 13, 2007

Abstract

We investigate the sensitivity of a Markov model with states and transition probabilities obtained from clustering a Molecular Dynamics trajectory. We have examined a 500ns Molecular Dynamics trajectory of the peptide VPAL (Valine - Proline - Alanine - Leucine) in explicit water. The sensitivity is quantified by varying the boundaries of the clusters and investigating the resulting variation in transition probabilities and the average transition time between states. In this way we represent the effect of clustering using different clustering algorithms. It is found that in terms of the investigated quantities the peptide dynamics described by the Markov model is sensitive to the clustering, in particular the average transition times are found to vary up to 46%. Moreover, inclusion of nonphysical sparsely populated clusters can lead to serious errors of up to 814%. In the investigation the time step used in the transition matrix is determined by the minimum time scale on which the system behaves approximately Markovian. This time step is found to be about 100ps. It is concluded that the description of peptide dynamics with transition matrices should be performed with care, and that using standard clustering algorithms to obtain states and transition probabilities may not always produce reliable results.

1 Introduction

There are many methods which seek to simulate the folding of a peptide or protein. They range from very course-grained approaches like the HP model [1] to models with atomic detail like Molecular Dynamics [2]. While the course-grained method gives results which can be useful as guidelines when designing proteins they do not describe exactly how a specific protein folds. To do this a model with the detail of Molecular Dynamics is needed. However, for the system sizes of interest the computational task of performing a Molecular Dynamics simulation which shows protein folding is unfeasible. Therefore, there have been developments of algorithms which modify standard Molecular Dynamics to allow for simulations of these larger systems [3, 4, 5, 6, 7, 8, 9, 10]. These methods range from modifying the potential energy landscape of the protein, to simulating several replicas of the same system at different temperatures, to constructing Markov models from a large number of Molecular Dynamics simulations.

A method which combines several Molecular Dynamics simulations by using clustering and a Markov model for the state transitions has recently been proposed. Using this method it is possible to reconstruct the overall dynamics of a peptide from thousands of individual simulations. This is done by counting the number of transitions between the different states from all the simulations. The Markov model can be described by a state vector v which holds probabilities for the different configurations and a transition matrix T. Given that the system has state vector v_t at time t, the state vector at time $t + \Delta t$ can be calculated as $v_{t+\Delta t} = T v_t$.

A source of error in this approach could be the clustering of configurational states. In the present paper we investigate how the state transition probabilities and folding dynamics varies with slightly different clustering. The total number of clusters is kept constant and only the boundaries between clusters are varied. This is done to try and mimic the effect of different clustering algorithms. The

investigation is carried out on a small peptide ensuring that possible transitions are sufficiently sampled. To the best of our knowledge there is no systematic analysis of the sensitivity of the clustering to the resulting dynamical characteristics. However, we have found that the results are sensitive to the clustering. Firstly, if the clustering is done in dihedral space, that could lead to the appearance of nonphysical sparsely populated states resulting in a variation in average transition times of up to 814%. This shows a likely effect of clustering incorrectly. Secondly, if the clusters are defined correctly, the sensitivity of the average transition times to the variation in the boundaries is up to 46%. This shows the likely variation in results obtained with a clustering algorithm that performs well.

Another source of error in the transition matrix approach is whether the transitions between the states can be described accurately with a Markov model. It is found that at short time scales the transitions does not behave Markovian, however, at longer time scales it does become Markovian. This is in line with previous work reported in the literature [11, 12]. The problem is addressed by choosing a sufficiently long time scale in the construction of the Markov model.

2 Methods

In this investigation we analyse a Molecular Dynamics trajectory. The simulation was performed using the software package GROMACS 3.2 [13]. The system examined was the four residue peptide VPAL (Valine - Proline - Alanine - Leucine) solvated in 874 water molecules. The peptide is shown in Figure 1. The simulation box was 3.0x3.0x3.0Å. The force field was 53a6 [14, 15, 16]. This is optimized for bimolecular systems interacting with water. Periodic boundary conditions were used. The temperature was kept at 300K using the Berendsen thermostat [17]. Atomic positions were recorded every 0.5ps. The integration algorithm was a Verlet type and the integration step was 0.002ps. The system was equilibrated before it was sampled for 500ns. This produced a total of 10^6

data points.

In our investigation we need to be able to vary the clusters. Therefore, the clustering is done by choosing dihedral angles as cutoff angles between the different regions. We only use the two central pairs of dihedrals because the terminal residues are too flexible and do not define the overall structure of the molecule. The initial clustering is represented by the solid lines in Figure 2. The dotted lines represent the interval in which the cutoff angles are varied. By varying each angle in turn it is possible to investigate the transition matrix as a function of different cutoff angles. Each angle is varied ± 0.5 rad around the initial cutoff. By plotting the variation in the transition matrix elements with the dihedral angles cutoff positions it is possible to inspect how sensitive the transition matrix is to clustering. By the method given in section 2.1 it is also possible to calculate how the variation in clustering affects the average transition time. The latter is a clear physical measure which characterises the folding routes directly. It can also be used to describe the folding pathways when there are multiple initial and final states.

To apply the Markov model transition matrix approach we need to find the time scale at which the system behaves Markovian. The Markovian assumption is that $v_{t+\Delta t} = T_{\Delta t} v_t$, where $T_{\Delta t}$ is the transition matrix constructed for a time step of Δt. For a transition matrix constructed at a time step of $n\Delta t$ we must have $T_{n\Delta t} = T^n_{\Delta t}$ where $n = 1, 2, 3 \ldots$. By expanding each transition matrix in eigenvalues and eigenvectors it can be shown that a necessary condition for the Markovian assumption to be valid is that $\lambda_{n\Delta t,i} = \lambda^n_{\Delta t,i}$ where λ denotes an eigenvalue and i runs over the number of eigenvalues. From this we find that $\lambda^{\frac{1}{n}}_{n\Delta t,i}$ has to be constant for $n = 1, 2, 3 \ldots$. This constant is the eigenvalue of a transition matrix with a time step of Δt which does satisfy the Markovian assumption. Given an eigenvalue it is possible to calculate a decay time (e.g. the half life) for the corresponding eigenvector. Using the constant eigenvalue we therefore get that the time $\tau_i = -\frac{n\Delta t}{\ln(\lambda_{n\Delta t,i})}$ has to be constant if the Markov description is correct. To find the time scale at which the system behaves

Markovian, we can therefore construct transition matrices for the time steps $n\Delta t$, $n = 1, 2, 3 \ldots$ and calculate τ_i for each matrix. The time step at which τ_i for all i become constants is the time step at which the system behaves Markovian.

2.1 Calculating the Average Transition Time

To calculate the average transition time of a Markov model we need to define initial and final states. Each of these can either be one state or a set of states. Assuming that we have a set of initial states I and a set of final states F the average transition time can be written as:

$$t_{IF} = \sum_{n=1}^{\infty} n P_{IF}(n) \tag{1}$$

Here $P_{IF}(n)$ is the probability for all paths of length n which start in I and end on F. We assume that the Markov process is described by a transition matrix T and that there is a total of N states. The first problem in the calculation is to find an expression for $P_{IF}(n)$. By introducing $\tilde{T}$,L,o and v we construct the following algorithm.

- From the transition matrix T remove the rows and columns for all states in F to form a new matrix $\tilde{T}$. This new matrix will have a dimension of $(N - d) \times (N - d)$, where d is the number of states in F.
- Form the matrix L which is of dimension $d \times (N - d)$ and holds the matrix elements of T that give the probabilities for entering F from all other states.
- Form the row vector o which is of dimension $1 \times d$ and holds ones in all places.
- Form the vector v of dimension $(N - d) \times 1$. The elements of v must describe the initial distribution of states in I. If each starting state is equally likely then their elements must be equal. For the states not in I

the initial value in v must be zero. The total sum of all elements in v must be 1.

Using the quantities given above $P_{IF}(n)$ can be written as $oL\tilde{T}^{n-1}v$ (an explanation is given in Appendix A). Let us assume that $\tilde{T}$ has eigenvectors e_i with corresponding eigenvalues λ_i. We then expand v in this basis. This gives $v = \sum_i \alpha_i e_i$. The average transition time (Equation 1) can then be written as:

$$
\begin{aligned}
t_{IF} &= \sum_{n=1}^{\infty} nP_{IF}(n) = \sum_{n=1}^{\infty} noL\tilde{T}^{n-1}v && (2)\\
&= \sum_{n=1}^{\infty} noL\tilde{T}^{n-1}\sum_i \alpha_i e_i = \sum_{n=1}^{\infty}\sum_i noL\alpha_i\lambda_i^{n-1}e_i && (3)\\
&= \sum_i \left(\sum_{n=1}^{\infty} n\lambda_i^{n-1}\right)\alpha_i oLe_i = \sum_i \frac{\alpha_i}{(1-\lambda_i)^2} oLe_i && (4)
\end{aligned}
$$

3 Results

In our investigation we partition the configurational space of the peptide in six different locations, Figure 2. In the plot for Proline we see that the two cutoff lines means that there are two states. In the Alanine plot there are four cutoff lines which gives three different states. This gives a total of six different states for the peptide. However, because one of the states found in this way is very sparsely populated we remove this state to form a total of five states. The average conformations in these states can be seen in Figure 4. To investigate if this clustering is correct we have compared it to clustering using Root Mean Square Deviation (RMSD). This is done by taking a representative configuration for each cluster and calculating the RMSD to all the configurations in each cluster. For cluster number 1 the result is shown in Figure 3. It can be seen that the RMSD is smallest for configurations which are also in cluster number 1, and that this cluster is well separated from the other clusters. Similar results are obtained when using the other clusters. Therefore clustering using cutoff angles in dihedral space is comparable to clustering using RMSD.

Using the states shown in Figure 4 allows the calculation of a transition

matrix. This is done by simply counting the number of transitions between the states in the Molecular Dynamics trajectory. This gives a frequency matrix which holds the number of transitions. By normalising the columns in this matrix to unity the transition matrix is obtained. To determine an appropriate time step to take when building the transition matrix we need to find the time step at which the system behaves in a Markovian manner. To to this we follow the procedure given in Section 2. Transition matrices are constructed with varying time steps. For each matrix the τ_i's are calculated for all i. The result of this can be seen in Figure 5. When the system behaves Markovian the τ_i's should be constant. From about 50ps it can be seen that the values become approximately constant, however, we chose a time step of 100ps to make sure that our system behaves sufficiently Markovian.

In Equation 5 the transition matrix for the initial clustering with a 100ps time step is given. It can be seen that once in a state there is a high probability of staying there in the next time step. From the transition probabilities it is possible to trace out the transition paths of the highest probabilities. These paths will be the conformational routes that the peptide will most likely follow during transitions. It is what is commonly know as the folding path. In Figure 6 the variation in transition probability between all pairs of states can be seen for the six different variations in cutoff angles. For some elements these variations are substantial. However, the variation of a single transition probability does not describe what happens for the peptide as a whole. Therefore to describe the sensitivity of the folding path of a peptide it is desirable to have a measure which describes how variations in the probabilities affect the folding path. This is exactly what is achieved by calculating the average transition time between states.

$$T_{100ps} = \begin{bmatrix} 0.8972 & 0.1006 & 0.0345 & 0.3502 & 0.0650 \\ 0.0407 & 0.7215 & 0.0055 & 0.0324 & 0.2129 \\ 0.0240 & 0.0136 & 0.9496 & 0.1289 & 0.0519 \\ 0.0295 & 0.0182 & 0.0091 & 0.3491 & 0.1475 \\ 0.0087 & 0.1461 & 0.0013 & 0.1394 & 0.5228 \end{bmatrix} \quad (5)$$

In Figure 7 the variation in average transition time between all pairs of states can be seen for the six different variations in cutoff angles. It is clear that the variation is much more significant compared to the variation of the transition matrix elements. This is because the variation in average transition time describes the variations in the folding path as a whole and not just a single transition. Since a deviation in cutoff angle from the initial cutoff angle will typically mean that clusters are connected by more transitions, the average folding time, between states, will generally tend to decrease. This causes a typical bell shaped variation in the average transition time as a function of the variation in cutoff angle. For the VPAL peptide we assume the unfolded state to be state 1 and the folded state, where the terminal residues of the peptide form a salt bridge, state 5. The average transition times betweens these two states are shown in Figure 8. In Figure 7 the average transition time between states are also shown for a transition matrix constructed with a time step of 0.5ps (red in the figure). As can be seen from Figure 5 this is not a correct description of the system since it does not behave Markovian at this time scale. However, it is still interesting to note that on this time scale the average transition times seem to be more sensitive to the clustering than at the longer time scale.

The transition probabilities for transitions directly between these two states are almost zero. This means that the variation in average transition time is caused by the variations of the transition probabilities between the intermediate states. The variation in average transition time between these two states is about 46%, which is significant. In the case where the sparsely populated state was included as a state on its own, the variations in average transition time

to and from this state was up to 814%. Examples of these large variations is shown in Figure 9. It can be seen that the variation mostly affected the average folding time between a few states. This is because the main path for transitions between other states does not include the scarcely populated state. For the VPAL peptide it can also be seen that t_{51} is generally larger than t_{15}, which means that the folded state is more stable than the unfolded state.

For a larger peptide the variation can be expected to be smaller, because there are many more paths by which the peptide can fold. However, assuming a given peptide has a folding path which passes though a few key states, then the average transition time could be very sensitive to the clustering of these states.

4 Conclusions

When constructing Markov models from Molecular Dynamics simulations care must be taken. Firstly, it is important that the Markov model is constructed with a sufficiently large time step so that the dynamics of the system are as close to Markovian as possible. In our investigation we found that the transitions behave sufficiently Markovian at 100ps time step. However, for the purpose of construction of reliable models we also found that this is not enough to ensure an accurate description of the dynamics. In particular we have found that transition probabilities and hence average transition times are sensitive to the specific clustering. By varying the boundaries between clusters we found that the variation in average transition time between representative initial and final states can reach 46%. When the transition matrix is constructed with a time step of 0.5ps (i.e. a non-Markovian time step) this variation increases to 100%. For a case where the initial clustering was miscalculated by inclusion of the nonphysical sparsely populated states we found the variations in average transition times between some of the states to be as much as 814%. The choice of clustering is a difficult one. On the one hand, if one chooses to use only clusters which are highly populated the transition probabilities and average transition

times will not be as sensitive. However, this may also mean that important information about the folding path is lost.

5 Acknowledgments

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

References

[1] K. F. Lau and K. A. Dill. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

[2] S. A. Adcock and J. A. McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106(5):1589–1615, 2006.

[3] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics*, 120(24):11919–11929, 2004.

[4] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, 1997.

[5] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, 1999.

[6] X. W. Wu and S. M. Wang. Self-guided molecular dynamics simulation for efficient conformational search. *Journal of Physical Chemistry B*, 102(37):7238–7250, 1998.

[7] X. W. Wu and S. M. Wang. Enhancing systematic motion in molecular dynamics simulation. *Journal of Chemical Physics*, 110(19):9401–9410, 1999.

[8] S. V. Krivov, S. F. Chekmarev, and M. Karplus. Potential energy surfaces and conformational transitions in biomolecules: A successive confinement approach applied to a solvated tetrapeptide. *Physical Review Letters*, 88(3), 2002.

[9] N. Singhal, C. D. Snow, and V. S. Pande. Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *Journal of Chemical Physics*, 121(1):415–425, 2004.

[10] G. Jayachandran, V. Vishal, and V. S. Pande. Using massively parallel simulation and markovian models to study protein folding: Examining the dynamics of the villin headpiece. *Journal of Chemical Physics*, 124(16), 2006.

[11] W. C. Swope, J. W. Pitera, and F. Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. theory. *Journal of Physical Chemistry B*, 108(21):6571–6581, 2004.

[12] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou. Describing protein folding kinetics by molecular dynamics simulations. 2. example applications to alanine dipeptide and beta-hairpin peptide. *Journal of Physical Chemistry B*, 108(21):6582–6594, 2004.

[13] D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. Gromacs: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, 2005.

[14] B. Hess and N. F. A. van der Vegt. Hydration thermodynamic properties of amino acid analogues: A systematic comparison of biomolecular force fields and water models. *Journal of Physical Chemistry B*, 110(35):17616–17626, 2006.

[15] C. Oostenbrink, T. A. Soares, N. F. A. van der Vegt, and W. F. van Gunsteren. Validation of the 53a6 gromos force field. *European Biophysics Journal with Biophysics Letters*, 34(4):273–284, 2005.

[16] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry*, 25(13):1656–1676, 2004.

[17] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

A Calculation of $P_{IF}(n)$

To illustrate how $P_{IF}(n)$ is calculated let us consider a three state system. Let the initial state be 1 and the final state 3. The transition matrix for the system is given as:

$$T = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

First we form the matrices $\tilde{T}$, L, o and v:

$$\tilde{T} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad L = \begin{bmatrix} a_{31} & a_{32} \end{bmatrix}, \quad o = \begin{bmatrix} 1 \end{bmatrix} \quad v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

For $n = 1$ we get:

$$P_{31}(1) = oL\tilde{T}^0 v = a_{31}$$

Since $P_{31}(1)$ is the probability to go from state 1 to state 3 in one step there is only one possible path 1-3. The probability for this is simply a_{31}. For $n = 2$ we get:

$$P_{31}(2) = oL\tilde{T}^1 v = a_{31}a_{11} + a_{32}a_{21}$$

There are two possible paths 1-1-3 and 1-2-3. The probability for each of these is $a_{31}a_{11}$ and $a_{32}a_{21}$ respectively. The sum of these therefore gives the total probability. For $n = 3$ we get:

$$P_{31}(3) = oL\tilde{T}^2v = a_{31}a_{11}a_{11} + a_{31}a_{12}a_{21} + a_{32}a_{21}a_{11} + a_{32}a_{22}a_{21}$$

In this case there are four possible paths from state 1 to 3. These are 1-1-1-3, 1-2-1-3, 1-1-2-3 and 1-2-2-3. $P_{31}(3)$ is the sum of the probabilities for each of these paths.

Figure 1

The VPAL (Valine - Proline - Alanine - Leucine) peptide. Carbon atoms are light blue, oxygens are red, nitrogens are dark blue, and hydrogens are grey. The united atoms force field 53a6 was used.

Figure 2

The Ramachandran plots for the Proline (left) and Alanine (right) residues. The initial clustering is marked by solid lines, while the boundaries for the variation in the clustering are marked by dotted lines. The lines are placed at: 1) -2.0 rad, 2) 0.5 rad, 3) -2.2 rad, 4) 0.5 rad, 5) -0.3 rad and 6) 2.5 rad. The areas marked A_1, B_1, A_2, B_2 and C_2 correspond to the conformations in Figure 4.

Figure 3

The average RMSD for the molecular configurations from different clusters compared to a representative conformation from cluster number 1. The error bars indicate the standard deviation.

Figure 4

The average conformations of the VPAL molecule in the different states. Comparing to the clusters in Figure 2 the states correspond to: 1) A_1A_2 2) B_1A_2 3) $A_1C_2 + B_1C_2$ 4) A_1B_2 5) B_1B_2

Figure 5

The variations in the τ_i's (see text). Each curve corresponds to an eigenvalue. The curve for the eigenvalue 1 is not shown as this gives an infinite τ value.

Figure 6

The range of transition probabilities for the different matrix elements as the clustering is varied. k is the matrix element index defined as $k = 5(i-1) + j$

where i is the row number and j the column number. The range of the variation has been magnified five times for clarity.

Figure 7

The range of the average time required for transition between all pairs of states. $k = 5(i-1)+j$ where i is the index of the initial state and j the index of the final state. Plots 1 to 6 correspond to each of the boundary variations. In red the same is shown for a model constructed with time step of 0.5ps (non-Markovian). The numbering is the same as in Figure 2.

Figure 8

The average transition time from state 1 to 5 (left) and 5 to 1 (right). The numbers correspond to different cutoff angles. The numbering is the same as in Figure 2.

Figure 9

The variations in the average transition times between all pairs of states. The variation is obtained by varying boundary 5. left: Total of five states right: Total of six states. $k = (5 \text{ or } 6) \cdot (i-1) + j$ where i is the index of the initial state and j the index of the final state.

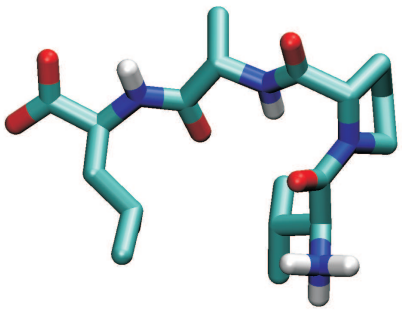


Figure 1:

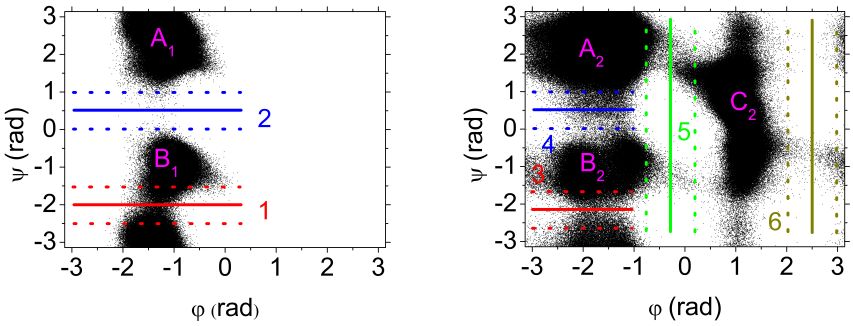


Figure 2:

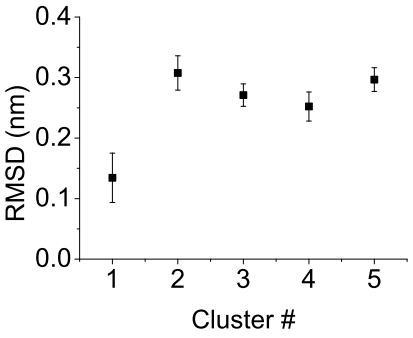
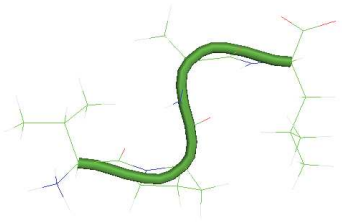
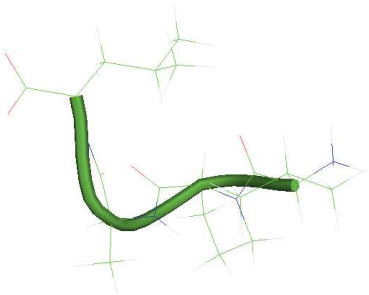


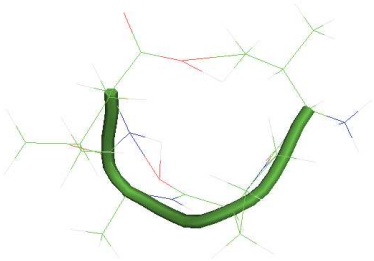
Figure 3:



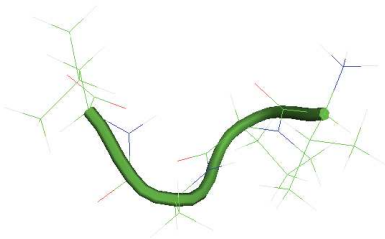
1)



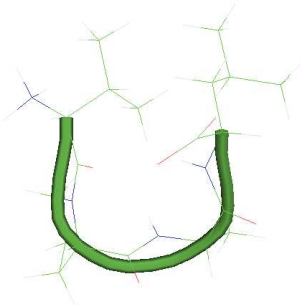
2)



3)



4)



5)

Figure 4:

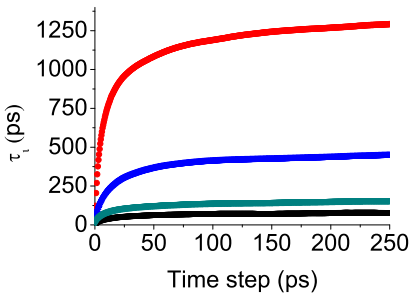


Figure 5:

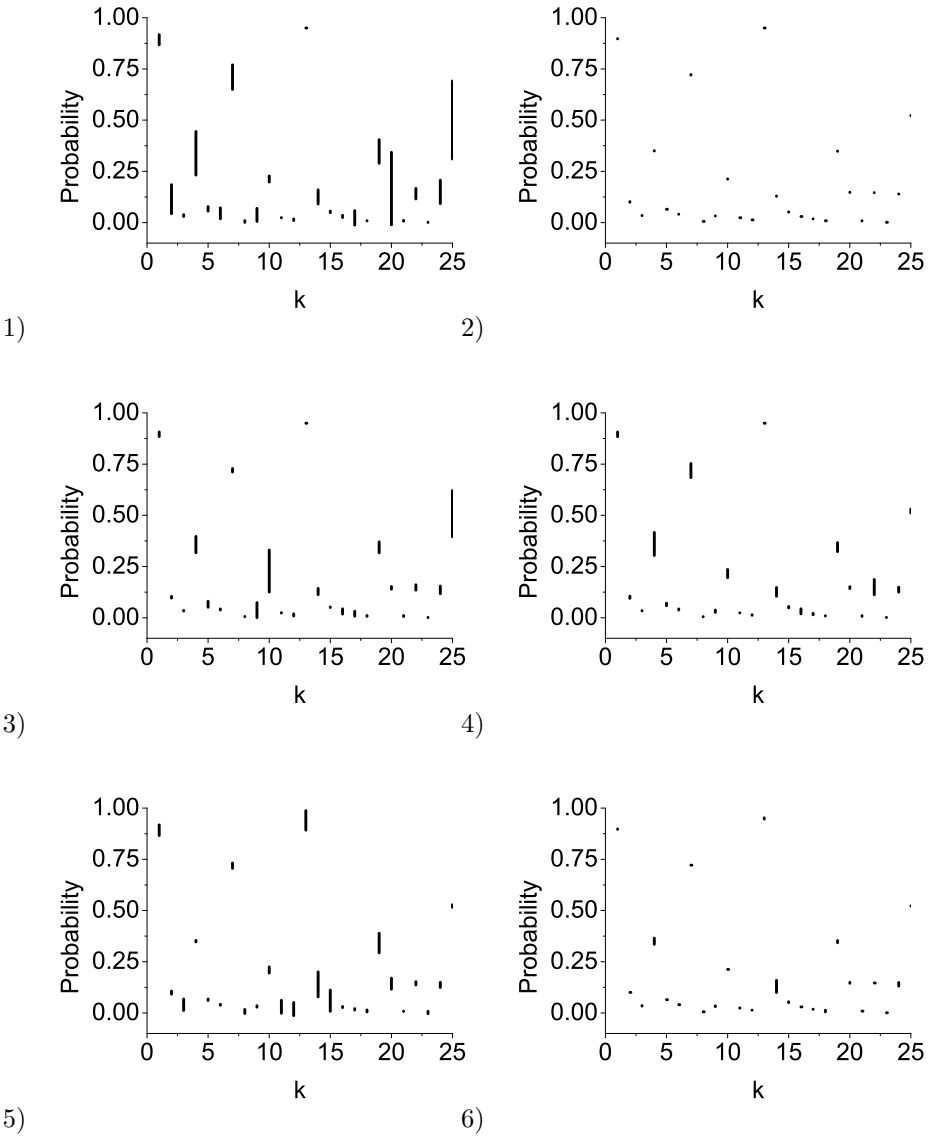


Figure 6:

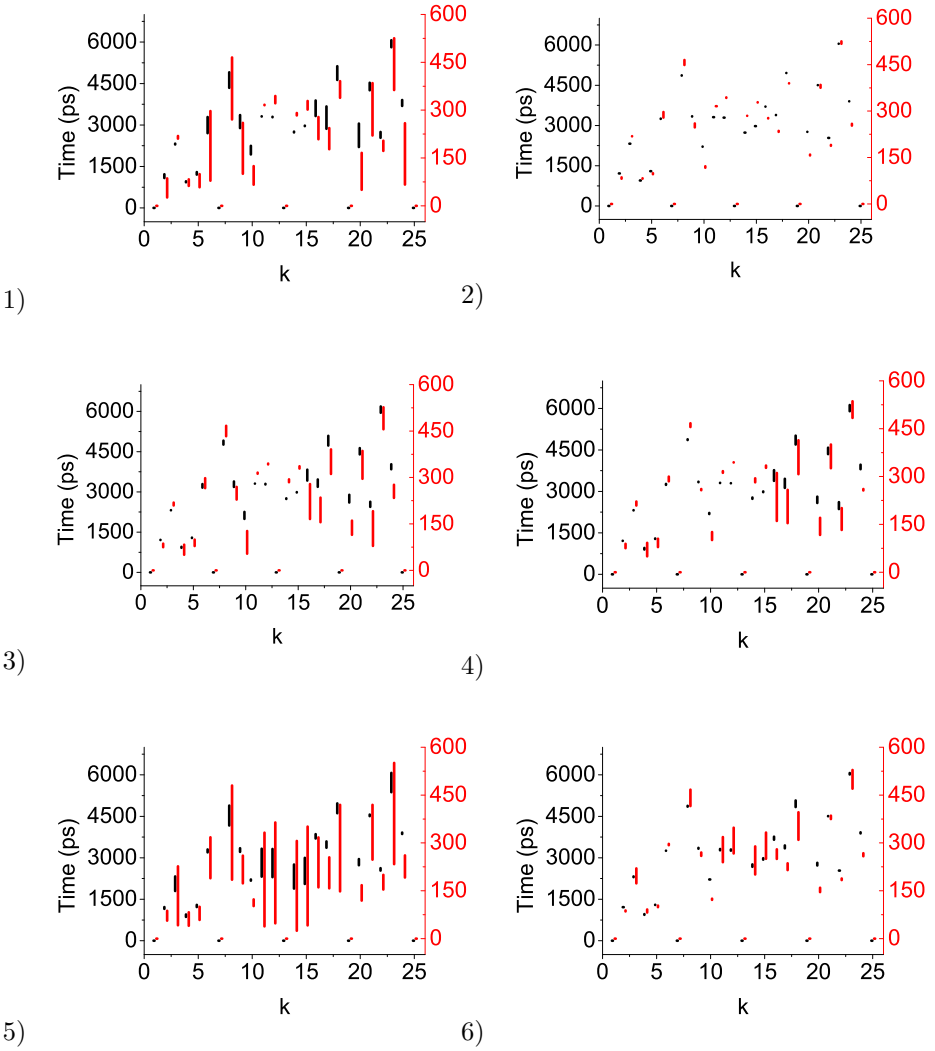


Figure 7:

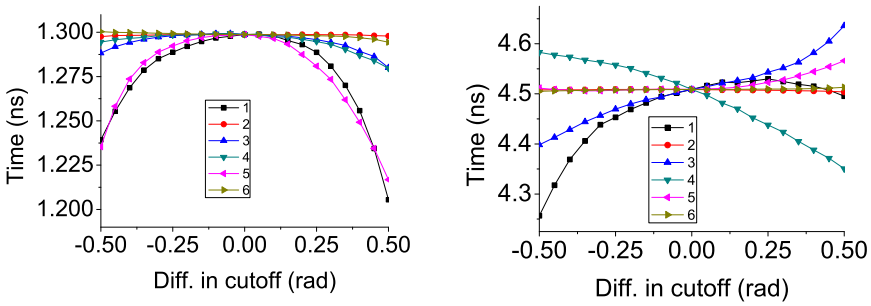

1.300
1.275
1.250
1.225
1.200
Time (ns)
-0.50
-0.25
0.00
0.25
0.50
Diff. in cutoff (rad)
1
2
3
4
5
6
4.6
4.5
4.4
4.3
Time (ns)
-0.50
-0.25
0.00
0.25
0.50
Diff. in cutoff (rad)
1
2
3
4
5
6

Figure 8:

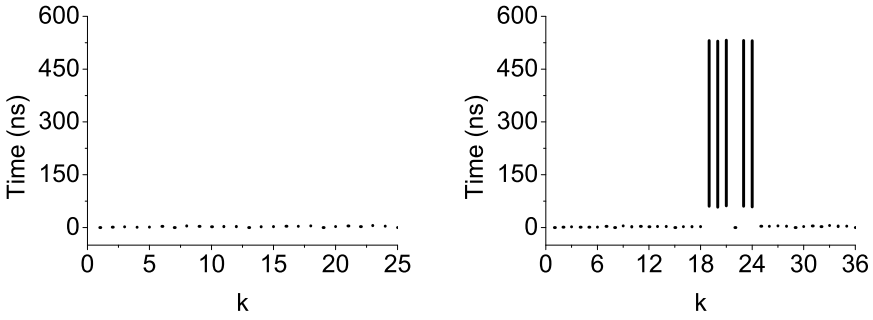

600
450
300
150
0
Time (ns)
0
5
10
15
20
25
k
600
450
300
150
0
Time (ns)
0
6
12
18
24
30
36
k

Figure 9: