F1000Research

← click for updates

RESEARCH ARTICLE

Molecular Dynamics Simulations of the Temperature Induced Unfolding of Crambin Follow the Arrhenius Equation. [version 1; referees: 1 approved, 2 approved with reservations]

Andrew Dalby[1], Mohd Shahir Shamsir[2]

[1]Faculty of Science and Technology, University of Westminster, London, W1W 6UW, UK
[2]Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Skudai, Johor, 81310, Malaysia

Abstract
Molecular dynamics simulations have been used extensively to model the folding and unfolding of proteins. The rates of folding and unfolding should follow the Arrhenius equation over a limited range of temperatures. This study shows that molecular dynamic simulations of the unfolding of crambin between 500K and 560K do follow the Arrhenius equation. They also show that while there is a large amount of variation between the simulations the average values for the rate show a very high degree of correlation.

Open Peer Review

Referee Status: ☑ ？ ？

	Invited Referees		
	1	**2**	**3**
version 1 published 20 Aug 2015	☑ report	？ report	？ report

1 **Melchor Sanchez-Martinez**, Mind the Byte Spain

2 **Dmitry Nerukh**, Aston University UK

3 **Adrian Mulholland**, University of Bristol UK

Discuss this article

Comments (0)

Competing interests: No competing interests were disclosed.

Introduction

Molecular dynamics (MD) simulations have become an important tool in understanding chemical and biochemical processes at the molecular level. Through the use of Newtonian mechanics and an empirically derived force-field, simulations have been used to investigate the interactions of drug molecules and their targets as well as to predict the behaviour of proteins and peptides (Vasquez *et al.*, 1994). Originally simulations were carried out *in vacuo*, but now with the increasing power of computers, simulations are usually carried out using periodic boundary conditions in water.

Increasingly realistic simulations lead to a better understanding of processes such as protein folding and unfolding at the molecular level (Lindorff-Larsen *et al.*, 2012; Scheraga *et al.*, 2007). The literature on molecular dynamics simulations for protein folding and unfolding is extensive. This can include very large simulations that can be simplified using coarse-graining, down to simulations of small proteins or peptides. One area where MD simulations have been particularly widely used is in the simulation of prion proteins and the protein mis-folding diseases associated with them (Shamsir & Dalby, 2005; Shamsir & Dalby, 2007; van der Kamp & Daggett, 2010; van der Kamp & Daggett, 2011).

Molecular dynamics simulations give us a detailed view of protein folding and unfolding pathways and their rates of unfolding (Daggett, 2002). Temperature is often used to accelerate protein unfolding and it has been shown that this does not affect the protein unfolding pathway (Day *et al.*, 2002). It is often difficult to relate the simulated results to experimental data. A review of simulated folding times has shown that the times predicted by MD and the experimental rates for thermal unfolding are in good agreement (Snow *et al.*, 2005). Atomic force microscopy data is another possibility and this has been used to investigate protein folding of T4 lysozyme (Peng & Li, 2008).

The rate of protein folding at increasing temperatures should be described by the Arrhenius equation over a limited range of temperatures (Alberty):

$$k = Ae^{E_a/RT}$$

Where k is the rate of the reaction, A is a constant (pre-exponential factor) E_a is the energy of activation, T is the Temperature in Kelvin and R is the Universal Gas Constant.

A rearrangement of the Arrhenius equation taking natural logarithms gives the linear function:

$$\ln(k) = \frac{-E_a}{R}\frac{1}{T} + \ln\ (A)$$

A plot of the natural logarithm of the rate ln(k) against 1/Temperature will be a straight line if the simulations obey the Arrhenius equation.

The Arrhenius equation has been used in solid-state chemistry calculations but currently no studies have tested whether it is valid in MD simulations of protein folding (Huwe *et al.*, 1999). This paper presents a MD simulation study using a small protein (crambin) to test whether the models do agree with the predicted linear behaviour.

Materials and methods

Crambin was chosen as the model protein for simulation of its size as it only has 46 amino acids (Caves *et al.*, 1998). A high resolution crystal structure of crambin (3NIR.pdb) was downloaded from the RCSB Protein Databank (Rose *et al.*, 2011; Schmidt *et al.*, 2011). Molecular dynamics simulations were carried out using Gromacs 4.6.4 on an Ubuntu 12.04 machine with GPU acceleration (Hess *et al.*, 2008).

The protein was solvated using periodic boundary conditions and a surrounding distance of 4nm around the protein. Simulations were run at 500K to 560K at 10K intervals using the OPLS forcefield. At temperatures of 570K and above the simulations fail to complete. The models were initially equilibrated using the canonical (nvt) and isothermal-isobaric (npt) ensembles. At 500K the simulations were run for 10ns with a time-step of 2fs. At 560K the simulations were run for 1ns with a time-step of 2fs. Secondary structure was calculated using DSSP (Dodge *et al.*, 1998) (the data is available in dssp_files.zip). RMSD deviations from the original crystal structure were calculated in Gromacs and displayed in Grace (the data is available in rmsd_grace_datafiles.zip). All of the scripts for equilibration dynamics and analysis can be found in the accompanying data files (the Gromacs files are in gromacs_files.zip and the shell script to run all the simulations and analysis is gromacs_runs_complete_analysis.sh).

The statistical analysis of the rate data was carried out using SPSS version 22 (IBM_Corp., 2013). The line of best fit to the mean data was fitted using linear regression (the data files are available as md_arrhenius_crambin_averaged.sav, md_arrhenius_crambin_averaged.spv). The line of best fit for the complete dataset was calculated using the general linear model (the data files are available as md_arrhenius_crambin.sav, md_arrhenius_crambin.spv). All of the scripts and data for the statistical analysis are available in the supplementary materials.

Results

Over these time periods crambin did not unfold as much as had been expected. As it is such a small protein it seems to be particularly stable to rises in temperature. There were three possible end points that could be used in determining the rate:

1) The unfolding of the C-terminal final bend (Figure 1, the green region from residues 36-38).

2) The loss of the beta sheet (Figure 1, the two red regions residues 2-4 and 33-34).

3) The increase of the root mean squared deviation (RMSD) to 0.4nm from the crystal structure.

It was not clear which of the measures would be the most reliable and so time was taken for all three. There were however missing values because the end points were not reached during the simulations. This was particularly apparent in the 540K simulations, which seem to be anomalous as can be seen in the boxplots for the reaction times from the simulations (Figure 2A–C). The boxplots show the expected downward trend, although there is considerable variation between the times taken for the repeated simulations. This variability declines at

Secondary structure
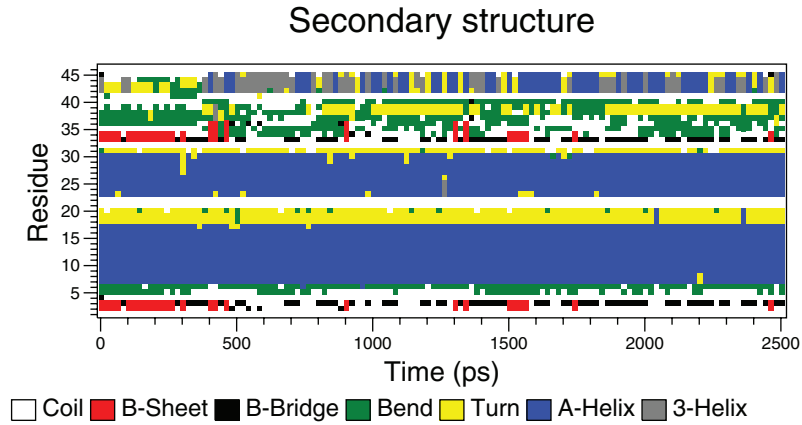
Figure 1. An Example DSSP Secondary Structure Plot (from 520K run 1).
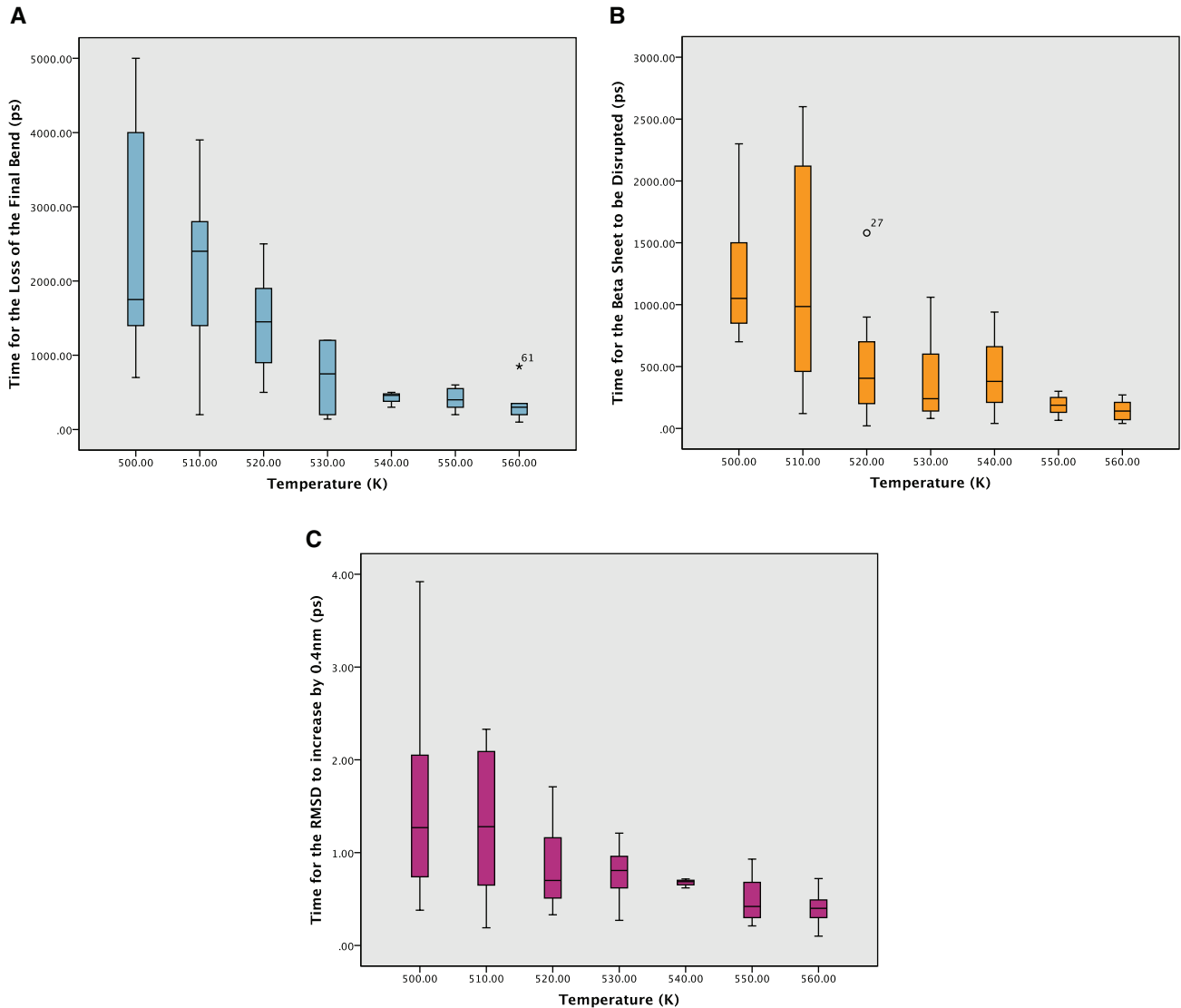

Figure 2. **A**: Boxplot of the times for the loss of the final bend from the secondary structure in picoseconds. Outliers are labelled. **B**: Boxplot of the times for the loss of the beta sheet in picoseconds. Outliers are labelled as circles or numbers if they are extreme. **C**: Boxplot of the times for the RMSD to go above 0.4nm from the crystal structure in picoseconds.

higher temperatures as rates become faster. A summary of the times to the three different end points are given in Table 1–Table 3.

The Arrhenius plot can be constructed using either the mean values for the different temperatures or all of the values for the repeats. In both cases a linear model is produced but the correlation is much stronger for the averaged values, which also give a narrower confidence interval for the line of best fit (Figure 3A–C). The wider confidence intervals for all of the data and the extent of the scatter can be seen in Figure 4A–C.

Table 1. Times until the loss of the bend from residue 36-38 in the protein in picoseconds.

Temperature	Mean (ps)	Standard Error (ps)	95% Confidence Interval (ps)
500	2340	480	1252 to 3427
510	2190	341	1418 to 2961
520	1470	219	974 to 1965
530	679	142	357 to 1001
540	420	61	157 to 683
550	415	42	319 to 510
560	328	71	163 to 492

Table 2. Times for loss of the beta sheet from the protein in picoseconds.

Temperature	Mean (ps)	Standard Error (ps)	95% Confidence Interval (ps)
500	1235	163	865 to 1604
510	1195	274	575 to 1815
520	517	151	175 to 859
530	400	111	149 to 651
540	577	122	278 to 876
550	186	25	129 to 243
560	135	25	78 to 192

Table 3. Time for the RMSD to go above 0.4nm from the initial crystal structure in picoseconds.

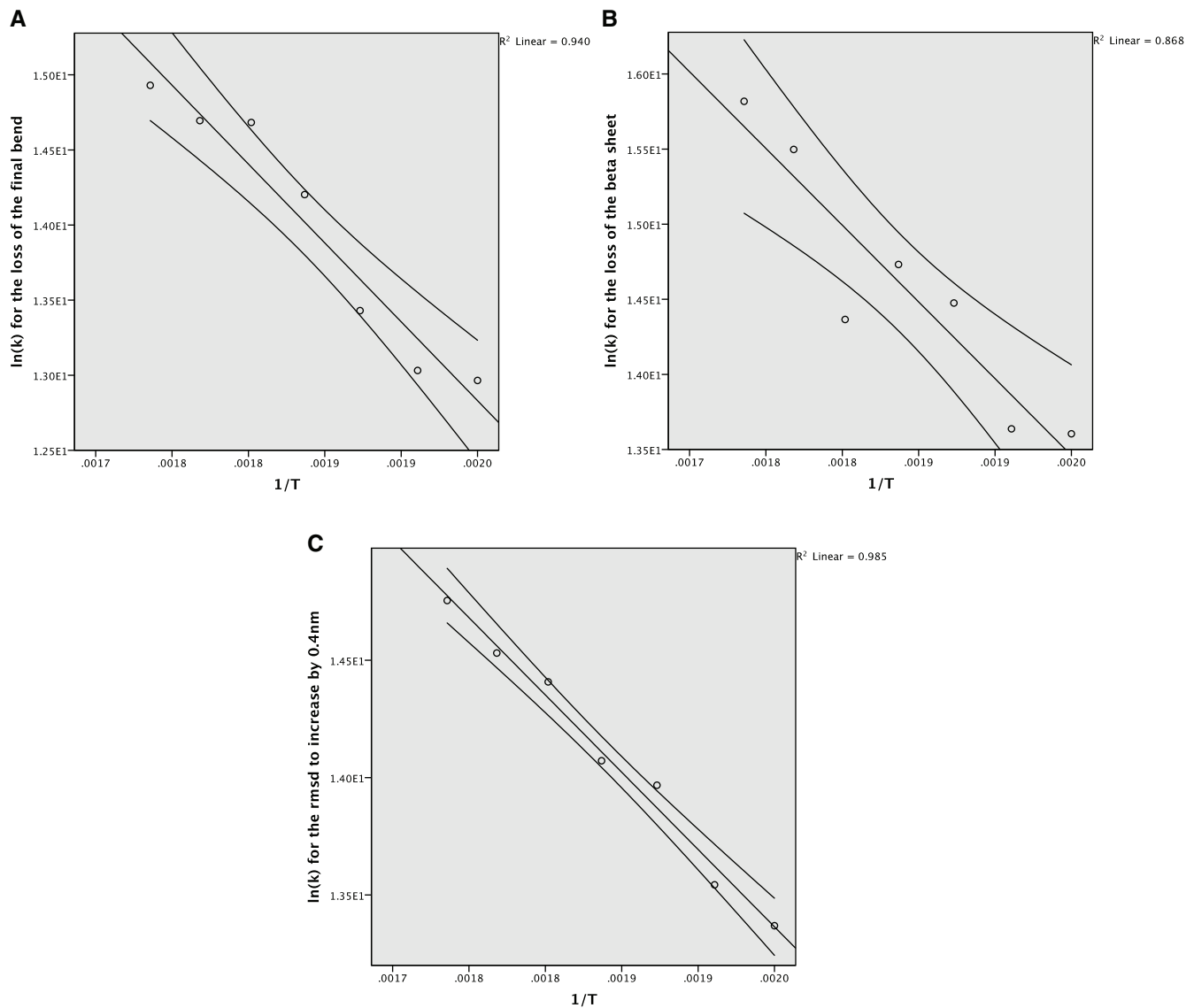| Temperature	Mean (ps)	Standard Error (ps)	95% Confidence Interval (ps)
500	1563	361	746 to 2380
510	1313	253	741 to 1885
520	859	147	528 to 1190
530	774	101	544 to 1003
540	553	76	341 to 764
550	489	77	314 to 664
560	391	64	242 to 540

A

B

C

Figure 3. **A**: The Arrhenius plot for the unfolding of the final bend using the averaged data. **B**: The Arrhenius plot for the unfolding of the beta sheet using the averaged data. **C**: The Arrhenius plot for the increase of the RMSD by 0.4nm from the crystal structure using the averaged data.

The parameters for the lines of best fit using all of the data and only the mean values are given in Table 4 and Table 5.

Discussion

All three of the end points used produce a linear model in the Arrhenius plot with a high degree of correlation. This suggests that in these cases the MD simulations are following Arrhenius behaviour and that these models can be used to predict rates of unfolding.

Of the three end points the RMSD variation is the easiest to calculate and least ambiguous, but it is also difficult to understand what this signifies at the protein level, when compared to using the

disruption of secondary structure as a metric. The other advantage of using the unfolding of the secondary structure is that experimental values are available for unfolding of proteins and so if an appropriate end point can be found in the simulations then the gradients of the Arrhenius plot can be used to calculate the activation energy for unfolding.

This study also highlights the high degree of variability between the trajectories of different simulations. This variability is very high at lower temperatures. Simulations were repeated ten times and this resulted in values for the standard errors for the time of unfolding that could be up to 25% of the time. These standard errors are large

A: ln(k) for the loss of the final bend vs 1/T — R^2 Linear = 0.5...

B: ln(k) for the loss of the beta sheet vs 1/T — R^2 Linear = 0.408

C: ln(k) for the rmsd to increase by 0.4nm vs 1/T — R^2 Linear = 0.348
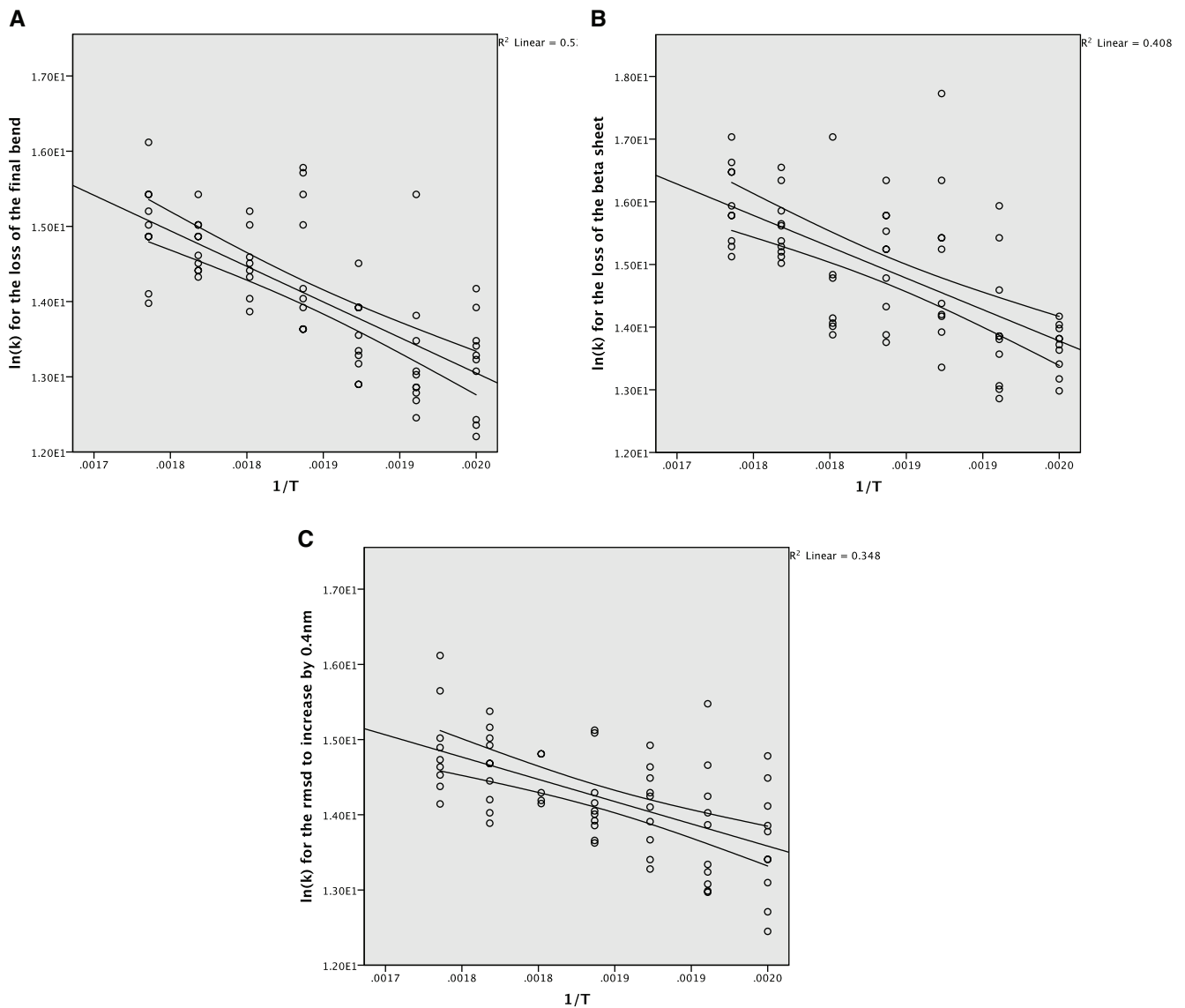
Figure 4. A: The Arrhenius plot for the unfolding of the final bend using the complete data. **B**: The Arrhenius plot for the unfolding of the beta sheet using the complete data. **C**: The Arrhenius plot for the increase of the RMSD by 0.4nm from the crystal structure using the complete data.

Table 4. Lines of best fit for the Arrhenius equation using all of the data.

End Point		Value	Confidence interval	Coefficient of determination
Bend loss	Gradient	-9458	-11760 to -7245	52%
	Intercept	32	27 to 36	
Beta sheet loss	Gradient	-10020	-13012 to -7027	41%
	Intercept	34	28 to 39	
RMSD > 0.4nm	Gradient	-5918	-7976 to -3860	35%
	Intercept	25	22 to 29	

Table 5. Lines of best fit for the Arrhenius equation using the mean rates.

End Point		Value	Confidence interval	Coefficient of determination
Bend loss	Gradient	-10500	-13555 to -7444	94%
	Intercept	34	28 to 40	
Beta sheet loss	Gradient	-10217	-14800 to -5635	87%
	Intercept	34	25 to 43	
RMSD > 0.4nm	Gradient	-6585	-7510 to -5660	98%

and so the number of simulations that are run needs to be increased in order to reduce them. As the standard errors falls with the square root of the sample size this would mean a 4-fold increase in the number of simulations is needed to reduce the standard error by half. This would suggest that more reliable results could be obtained by carrying out 40 simulations at each of the temperatures, which is a considerable additional computational burden.

The large amount of variation also affected the quality of the lines of best fit to the Arrhenius equation. The coefficient of determination was much lower when considering all of the data but nonetheless there is clear evidence of the simulations following Arrhenius behaviour. The confidence intervals for the linear models using all the data and the average data are similar. The variation in the gradient is too large for making comparisons with experimentally derived energies of unfolding and this is another reason why a larger number of simulations will be needed in future studies.

There was a surprising degree of agreement in the slopes and intercepts of the bend loss and beta sheet loss end-points. This suggests that the energies involved in stabilising the beta sheet and final bend are similar. This consistency is encouraging and suggests that detailed energy predictions will be possible from MD simulations.

Crambin was not an ideal case for using in this study. Although it is very small and allows the simulations to be run in a shorter time, the protein does not unfold very much and so longer time-scales are needed within the simulations. Prion protein is another possible model that could be used to test unfolding as this has been shown

to unfold over short simulation times (< 10ns) (Shamsir & Dalby, 2005). The other alternative is longer simulations times in order to produce clearer and less ambiguous end-points for the simulations.

Data and software availability

All of the code and data required for carrying out the simulations is available from http://dx.doi.org/10.5281/zenodo.20550. The shell script used to perform the simulations is available from http://dx.doi.org/10.5281/zenodo.20544. The dssp plots and the RMSD graphs for the simulations are available from http://dx.doi.org/10.5281/zenodo.20548. The SPSS data files and output files that include the details of how the analysis was carried out are available from http://dx.doi.org/10.5281/zenodo.20549. The software is released under a MIT license.

Author contributions

ARD conceived the study and designed the experiments. ARD and MSS carried out the simulations and the analysis. ARD and MSS were involved in all of the versions of the manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting the work.

References

Alberty R: **Physical Chemistry**. John Wiley and Sons: New York. 1987.
Reference Source

Caves LS, Evanseck JD, Karplus M: **Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin.** *Protein Sci.* 1998; **7**(3): 649–666.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Daggett V: **Molecular dynamics simulations of the protein unfolding/folding reaction.** *Acc Chem Res.* 2002; **35**(6): 422–429.
PubMed Abstract | **Publisher Full Text**

Day R, Bennion BJ, Ham S, *et al.*: **Increasing temperature accelerates protein unfolding without changing the pathway of unfolding.** *J Mol Biol.* 2002; **322**(1):

189–203.
PubMed Abstract | **Publisher Full Text**

Dodge C, Schneider R, Sander C: **The HSSP database of protein structure-sequence alignments and family profiles.** *Nucleic Acids Res.* 1998; **26**(1): 313–315.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Hess B, Kutzner C, Van Der Spoel D, *et al.*: **GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation.** *J Chem Theory Comput.* 2008; **4**(3): 435–447.
Publisher Full Text

Huwe A, Kremer F, Behrens P, *et al.*: **Molecular Dynamics in Confining Space:**

From the Single Molecule to the Liquid State. *Phys Rev Lett.* 1999; **82**(11): 2338.
Publisher Full Text

IBM_Corp: **IBM SPSS Statistics for Windows, Version 22.0**. Armonk, NY: IBM Corp. 2013.
Reference Source

Lindorff-Larsen K, Trbovic N, Maragakis P, *et al.*: **Structure and dynamics of an unfolded protein examined by molecular dynamics simulation.** *J Am Chem Soc.* 2012; **134**(8): 3787–3791.
PubMed Abstract | **Publisher Full Text**

Peng Q, Li H: **Atomic force microscopy reveals parallel mechanical unfolding pathways of T4 lysozyme: evidence for a kinetic partitioning mechanism.** *Proc Natl Acad Sci U S A.* 2008; **105**(6): 1885–1890.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rose PW, Beran B, Bi C, *et al.*: **The RCSB Protein Data Bank: redesigned web site and web services.** *Nucleic Acids Res.* 2011; **39**(Database issue): D392–D401.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Scheraga HA, Khalili M, Liwo A: **Protein-folding dynamics: overview of molecular simulation techniques.** *Annu Rev Phys Chem.* 2007; **58**: 57–83.
PubMed Abstract | **Publisher Full Text**

Schmidt A, Teeter M, Weckert E, *et al.*: **Crystal structure of small protein crambin at 0.48 Å resolution.** *Acta Crystallogr Sect F Struct Biol Cryst Commun.* 2011; **67**(Pt 4): 424–428.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Shamsir MS, Dalby AR: **One gene, two diseases and three conformations: molecular dynamics simulations of mutants of human prion protein at room temperature and elevated temperatures.** *Proteins.* 2005; **59**(2): 275–290.
PubMed Abstract | **Publisher Full Text**

Shamsir MS, Dalby AR: **Beta-Sheet containment by flanking prolines: Molecular dynamic simulations of the inhibition of beta-sheet elongation by proline residues in human prion protein.** *Biophys J.* 2007; **92**(6): 2080–2089.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Snow CD, Sorin EJ, Rhee YM, *et al.*: **How well can simulation predict protein folding kinetics and thermodynamics?** *Annu Rev Biophys Biomol Struct.* 2005; **34**: 43–69.
PubMed Abstract | **Publisher Full Text**

van der Kamp MW, Daggett V: **Pathogenic mutations in the hydrophobic core of the human prion protein can promote structural instability and misfolding.** *J Mol Biol.* 2010; **404**(4): 732–748.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

van der Kamp MW, Daggett V: **Molecular dynamics as an approach to study prion protein misfolding and the effect of pathogenic mutations.** *Top Curr Chem.* 2011; **305**: 169–197.
PubMed Abstract | **Publisher Full Text**

Vasquez M, Nemethy G, Scheraga HA: **Conformational energy calculations on polypeptides and proteins.** *Chem Rev.* 1994; **94**(8): 2183–2239.
Publisher Full Text

Open Peer Review

Current Referee Status: ☑ ? ?

Version 1

Referee Report 28 October 2015

doi:10.5256/f1000research.7345.r10967

Adrian Mulholland

Centre for Computational Chemistry, University of Bristol, Bristol, UK

I've looked at this paper and have some comments and concerns, some of which reflect the other reviewer's comments.

I believe that there have been previous demonstrations of Arrhenius behaviour in unfolding simulations, including for crambin (see e.g. Ferrara *et al*. J. Phys. Chem. B 2000, 104, 5000-5010) and other proteins (e.g. Piana *et al.* PNAS 109 17845–17850, doi: 10.1073/pnas.1201811109).

Consideration of protein stability is crucial here (see Scalley and Baker Proc Natl Acad Sci U S A. 1997 Sep 30; 94(20): 10636–10640) and as the current simulations are run at high temperatures where the protein is unstable, this is an issue. The dielectric behaviour of the water model at these high temperatures is also a concern for me.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 14 October 2015

doi:10.5256/f1000research.7345.r10438

Dmitry Nerukh

Institute of Systems Analytics, Aston University, Birmingham, UK

I have two main critical points:
1. The conceptual one is that the loss of the beta sheet or the bend could not be, strictly speaking, classified as "unfolding". There is very little in the text that describes how specifically the authors quantified the moment of this loss of structure. Looking at Fig. 1 it can be seen that both motifs come back for some periods of time. Is this a loss of structure or just equilibrium fluctuations? If the latter, then there are no observable unfolding event in the data.

2. As authors rightfully admit, the uncertainty of these loss of structure times is very high, especially at low temperatures. This is most likely because the length of simulations, 1-10ns, is too short (as again, the authors mention this in the text) and also because the algorithm of determining these times is not statistically robust. For proteins, even small ones like crambin, typical time scales for structural rearrangements is of the order of tens of nanoseconds. Even dialanin takes hundreds of nanoseconds to accumulate statistically sound number of its very simple conformational transitions.

Taking into account these points, the main conclusion of the authors based on data shown on Fig. 2-4 appears unsubstantiated to me.

Overall, the paper is interesting and worth indexing if the authors more rigorously calculate the "loss of structure" times. I would consider building a statistical network on the major conformational states, in the spirit of Markov States Model. Then, the average transition times would faithfully represent the "unfolding" events.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 28 September 2015

Melchor Sanchez-Martinez
Mind the Byte, Barcelona, Spain

The research article entitled 'Molecular Dynamics Simulations of the Temperature Induced Unfolding of Crambin Follow the Arrhenius Equation' by Dalby and Shamshir shows how the Molecular Dynamics simulations of crambin unfolding do follow the Arrhenius equation. It is addressing an interesting point as that Molecular Dynamics simulations of protein folding is used to exhibiting an non-Arrhenius behaviour and protein unfolding follow an Arrhenius equation, it is widely accepted. Testing if it is true is a really nice and interesting work.

In the article, there is a comprehensive explanation of study design, methods and analysis. The conclusions are well explained and justified on the basis of the results, Furthermore, the authors have made the data fully available, including the scripts to reproduce their work, being able to be revised by the author. That point makes the work more trustable and robust.

As a consequence of that, the manuscript is recommended to be approved. However there are some minor changes and comments that the authors may consider to improving the study.

The authors stated that "The Arrhenius equation has been used in solid-state chemistry calculations but currently no studies have tested whether it is valid in MD simulations of protein folding". This phrase should be rewritten and clarified, because as it is written, it seems that this is the first time that the Arrhenius equation is used in protein folding simulations and that's not true because its usage is widely established in protein folding.

To gain statistical significance, the authors performed 10 replicas, which is a good number. However more replicas changing the force-field, for instance using Amberff99SB-ILDN, would be a good option to gain more statistics and at the same time ensure to avoid the force-field dependence of the results, as in folding/unfolding processes it is very relevant (see for instance Lindorff-Larsen *et al*., 2012). Simulations with one or two more force-fields, performing 10 replicas for each one, would be appreciated.

To explore if the protein is unfolded or not, three different metrics were employed. One of them is the increase of the root mean squared deviation (RMSD) to 0.4nm from the crystal structure. Why 0.4 and not another value? There are some references to that in the literature or it is just an observed trend in the simulations? If it is an observed trend, would change varying the force-field. In Shimada *et al.*, 2001, they identified important structural regions of Crambin. Some of them are used by the authors as unfolding indicators, although the aminoacid numbers are not exactly the same. Thus, I wondering if helix 1 (residues 6-18,sequence: SIVARSNFNVCRL) could be also a good indicator of Crambin unfolding with the employed PDB structure, and if it is, if there is a reason to not be used.

As the authors stated longer simulations of 40ns or even longer should show a more realistic and picture of Crambin unfolding. It is true that it constitutes a considerable additional computational time, but will probably result in stronger and more sound results. Thus I encourage the authors to do it, at least, in the future.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 30 Sep 2015
Andrew Dalby, University of Westminster

I would like to thank the referee for his helpful comments.

We definitely need to correct that omission in referring to other work using Arrhenius on proteins, although it has mostly been on short segments of protein or peptides rather than a complete protein. In writing the introduction I was surprised by the lack of literature on using Arrhenius in a biological simulation context, except for enzyme simulations. I had searched for Arrhenius behaviour in molecular dynamics and as stated I only found references to its use in simulations of materials. It seems that the problem was I should have searched just for Arrhenius and in using behaviour it needs to be the US spelling, behavior to return the protein modelling literature! The work of Baker, Karplus and Kuriyan, and Pande are particularly important in establishing the field. The more recent work by Best and Mittal is another excellent example which incorporates the effects of forcefields.

I agree with the point about the forcefield but this adds a confounding variables and also increases the variance of the simulations, which increases the uncertainty in the gradient and has a negative impact on the confidence intervals for predicted energy barriers. An alternative is to use the technique to calculate relative differences in energy barriers between protein variants using the same forcefield. As you are calculating a difference the forcefield effects can be considered to cancel out. This was the idea behind free energy perturbation calculations.

Competing Interests: I am the author