Identifying Crucial Moments and Hotspots in Protein Dynamics

November 5, 2024

1 Introduction

The dynamic existence of a protein can be tracked through snapshots, or "frames," captured at specific time steps. At each frame, residues are found in unique positions, with some cooperating towards protein folding, others resisting the fold and creating tension, while the role of others remains enigmatic.

This work operates under the premise that a protein's destiny is influenced by its constituent residues. By employing the energy gap method, we identify pivotal snapshots that display significant energy shifts. Tracking the energy contributions (eigenvector components) of each residue reveals their roles during these crucial moments.

2 The Energy Gap and Its Standard Deviation

A peak into when major energy shifts take place. The energy gap is defined as the difference between the least and second least eigenvalues, normalized by the average difference between consecutive eigenvalues. Each frame will have an Energy Gap value "ENG(t)" and its standard deviation "SDENG(t)". Frames with larger ENG(t) and smaller SDENG(t) may reflect a native protein state.

$$\mathrm{ENG}(t) = \frac{\Delta\lambda_{1-2}(t)}{\langle\Delta\lambda(t)\rangle}$$

The protein under study is the Trp cage. Figure 1 below reveals a time interval (200-270 ps) where there is a significant rise in the ENG(t) value and a corresponding drop in the SDENG(t) value. In this work, we capture 10 frames spaced at 40 ps. The frame found to have the maximum ENG(t) value is frame 7, and that with the minimum SDENG(t) is frame 6, at time points 160 and 190 respectively. No frame was found to have both maximum and minimum values for the ENG(t) and SDENG(t) simultaneously.

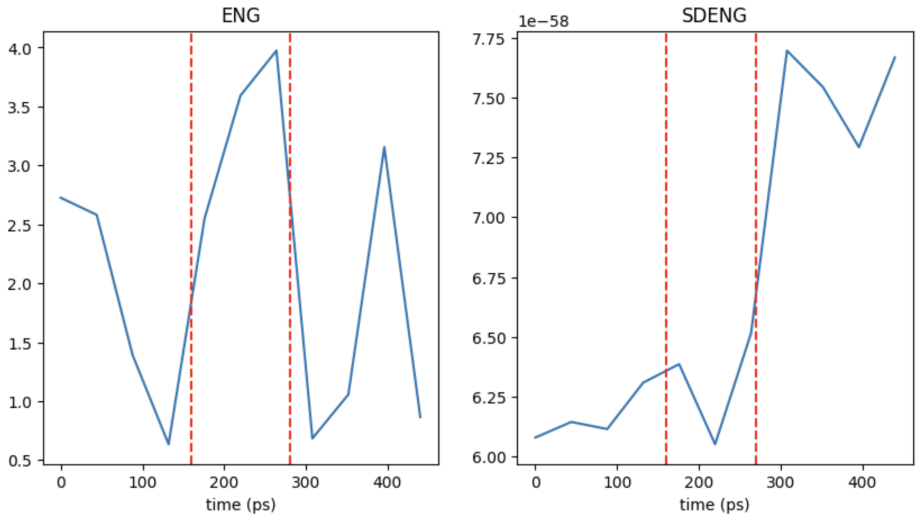


Figure 1: Plots of the ENG(t) and SDENG(t) of each frame over the total simulation time.

3 Which Residues Contribute to Folding, Unfolding or Neutral: Correlation Distance

The eigenvector components of each residue are collected over all the pivotal frames and plotted against time to visualize their contribution trends. Figure 2 identifies ASP9 and ARG16 as having major roles, consistent with the paper, which also highlights Gln5 and Gly15.

To further assess the type of contribution by each residue, the paper calculates the correlation distance, a quantity derived from the Pearson correlation coefficient as follows:

$$\mathrm{CD}_i = \sqrt{2(1-\rho_i)}$$

where ρ_i is defined by:

$$\rho_i = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

This parameter expresses a similarity measure between the two time series (ENG(t) and components). It is suggested that residues driving folding–unfolding can be identified by high correlation (minimum CDi value) between their energy component and the ENG. Non-native contacts, which decrease their stability contribution upon folding, are likely to be less correlated and thus exhibit a high CDi value. In this framework, regions of maximal correlation associated with minimal values of CDi predict the folding core. Regions of maximal anticorrelation correspond to residues involved in non-native con-

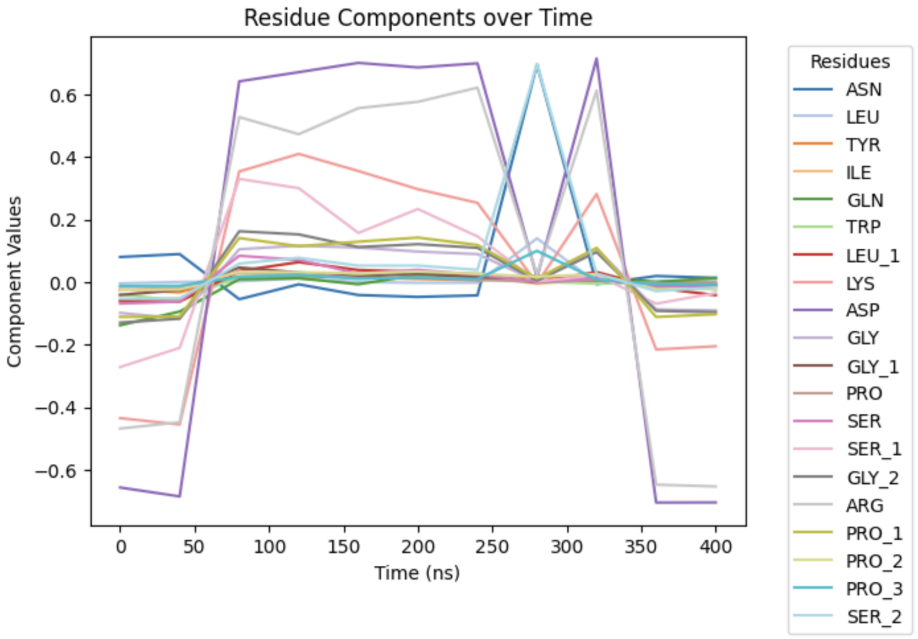


Figure 2: Energetic contribution per residue over the simulation time.

tacts that need to be disrupted to allow the protein to proceed to the folded ensemble.

4 Folded State or Unfolded State?

The precise thresholds defining a folded state versus a non-folded state can be identified as follows:
1) Identifying the highest peak (ENG_max) and the deepest valley (SDENG_min) in the energy landscape of the protein. This defines the range within which the protein's conformational changes are considered significant.
2) Calculate the standard deviation of the frames energies from the ENG_max. This determines the energy step size. (Msd)

Psudo code:

Define the range for percentage n from 1 to 100.
Initialise arrays to hold counts of conformations:

- `eng_counts`
- `sdeng_counts`

Loop through each percentage to calculate thresholds and count conformations:

- FOR each n in percentages:
- ENG threshold = ENGmax - $\left(\frac{n}{100}\right) \times$ MSD
- SDENG threshold = SDENGmin + $\left(\frac{n}{100}\right) \times$ MSD
- Count conformations where ENG is above ENG threshold
- Count conformations where SDENG is below SDENG threshold
- Append counts to `eng_counts` and `sdeng_counts`

The output of this algorithm should provide us the thresholds (the horizontal dashed lins on the paper) that specify the ENG(t) belonging to folding or otherwise.

The graphs in figure 4 had to look like sigmoid functions but I think it's due to the small number of data points (10 frames only). Figure 5 is how they looked like in the paper.

	Residue	Correlation Coefficient	Correlation Distance
0	TRP	0.095479	1.345007
1	ARG	0.020521	1.399628
2	PRO_1	-0.015837	1.425368
3	ASP	-0.022440	1.429993
4	GLY	-0.080936	1.470331
5	GLY_2	-0.088960	1.475778
6	PRO_2	-0.106012	1.487287
7	ILE	-0.126208	1.500805
8	SER_1	-0.129922	1.503278
9	LYS	-0.131437	1.504285
10	GLY_1	-0.143242	1.512113
11	LEU_1	-0.153215	1.518693
12	GLN	-0.178438	1.535212
13	SER	-0.229633	1.568205
14	PRO	-0.265331	1.590806
15	TYR	-0.316243	1.622494
16	ASN	-0.382859	1.663045
17	SER_2	-0.416370	1.683074
18	PRO_3	-0.424690	1.688011
19	LEU	-0.448599	1.702115

Figure 3: A table of Pearson correlation and correlation distance values per residue.

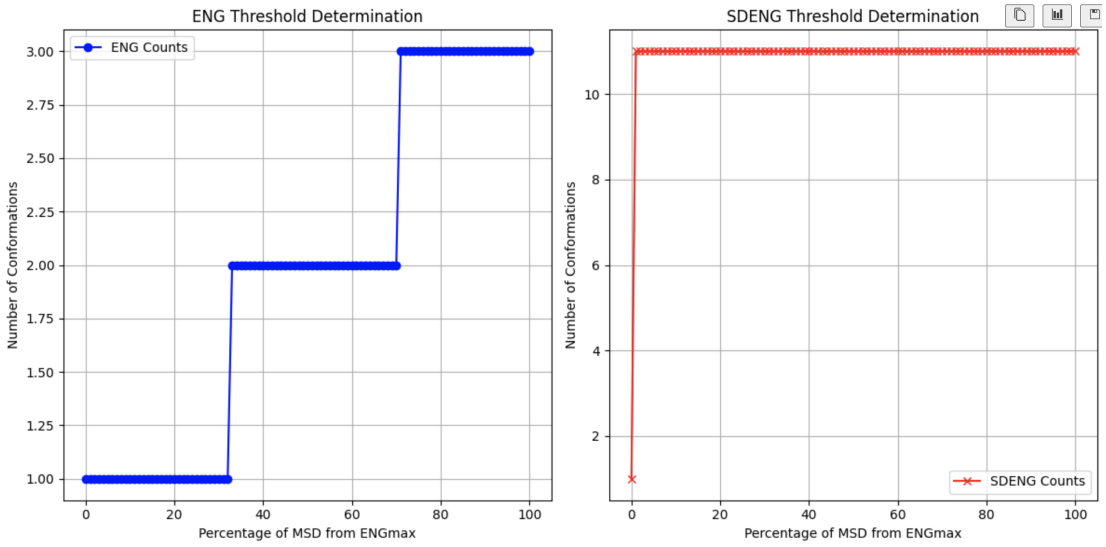
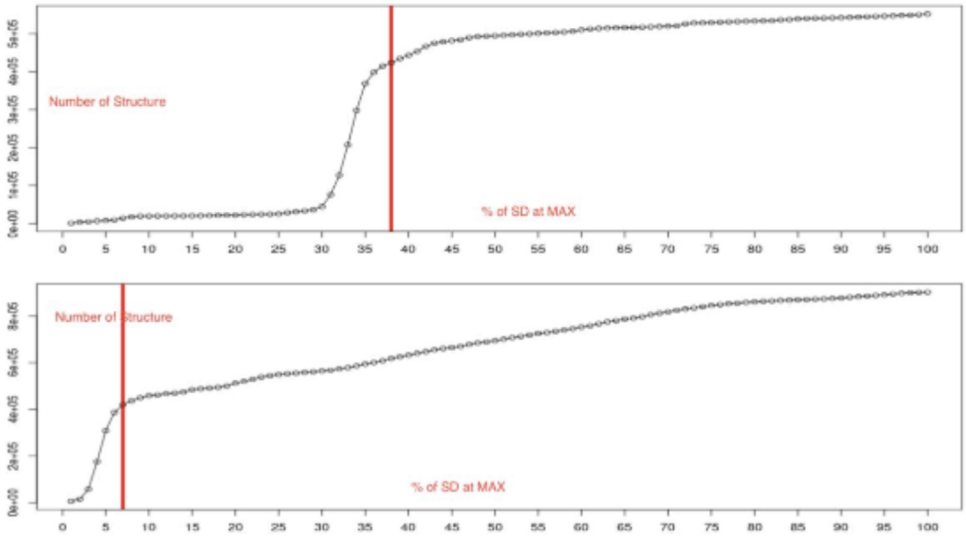


Figure 4: Population counts of frames residing below ENG_{max} and above SDENG_{min}



Supporting Information Figure S1: Example of population curve of MD snapshots above (below) threshold vs percentage of decrease (increase) from starting threshold. The decrease/increase steps are percentage multiples of the SD evaluated at ENGmax. Top. Number of conformations between ENGmax and [ENGmax – % of SD at MAX] Bottom. Number of conformations between SDENGmin and [SDENGmin + % of SD at MAX].

Figure 5: Threshold determinatin form the paper