Understanding Energy Gap Method

October 17, 2024

0.1 Introduction

Protein structures appear to be extraordinarily tolerant to mutations at many positions, provided that they do not excessively destabilize the native state. However, there are some specific sites (hot sites) where even subtle mutations or perturbations result in severe changes in the stability of the folded structure. Not all amino acid positions along a chain are thus equally important in determining the native conformation during folding or in stabilizing it at equilibrium. The matrix described in the paper is an interaction matrix between the protein residues. In the paper, they performed the simulation on TRP Cage36 and Chignolin, PDB database code 1L2Y and 1UAO respectively. I attempted to reproduce the results only for the TRP-cage protein. My simulation identified 7 hot spot residues but only ASP9 is consistent with the 2 results in the paper. (*could be about the threshold. In the paper they used a 3 steps algorithm to automate the threshold choice. I am still working on it.*)

The protein has 20 residues therefore the interaction matrix M is of size 20 x 20 generated in my case using Gromax by first creating a contact (distance) Matrix.

0.2 My understanding of the theory

To simplify the pattern of interactions between the residue we can essentially transform the matrix into a new coordinates system where the eigenvectors would be the independent modes of interaction. Each eigenvector in the new coordinates system represents the residues working together and contributing to the particular mode of interaction. The components of the vector are the contributions per residue, the higher the component the higher the contribution of "Hot spot residue". The eigenvalues here represent the strength of each mode altogether. The smallest eigenvalue corresponds to the most stabilizing mode and vice versa. (*There is a note on the paper about graph theory that I am yet to understand.*)

0.3 Diagonalization: Simplifying the Interaction Space

The eigenvalue Decomposition of matrix M is expressed as follows:

$$M = V\Lambda V^{-1}$$

Where:
V is the matrix of eigenvectors.
Λ is the diagonal matrix of eigenvalues.
V^{-1} is the inverse of the matrix of eigenvectors.

I performed it using numpy.linalg.eig() function in Python.

0.4 Energy GAP

The energy gap compares the smallest and second smallest eigenvalues normalized by the average gap across all eigenvalues.
The energy gap ENG(t) is calculated as follows:

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

Where $\Delta\lambda_{1-2}(t)$ is the spectral gap between the first and second eigenvalue, and $\langle\Delta\lambda(t)\rangle$ is the average separation between consecutive eigenvalues.

Below is the plot of the ENG(t) and the SDENG(t) over time. The values are collected looping over all the frames.

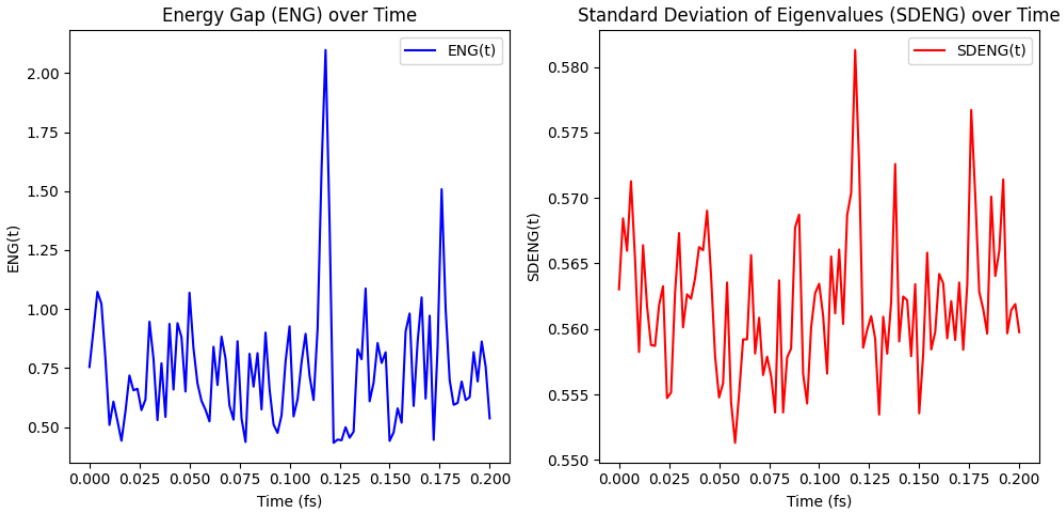

Figure 1: ENG(t) and SDENG(t)

```
Minimum Eigenvalue: -0.3497517499733523
ENG(t): 0.5371369618319116
Hotspot Residues, Names, and Eigenvector Components (corresponding to the minimum eigenvalue):
Residue 5 - GLN, Eigenvector component: 4.2179001051476225e-05
Residue 9 - ASP, Eigenvector component: -0.09093565230608261
Residue 10 - GLY, Eigenvector component: 0.5574969859592345
Residue 11 - GLY, Eigenvector component: -0.715252955090863
Residue 12 - PRO, Eigenvector component: 0.29617687751987326
Residue 13 - SER, Eigenvector component: -0.15919309990113445
Residue 15 - GLY, Eigenvector component: -0.05663010996971183
```

Figure 2: The identified Hot spot residues

```
List of all protein residues:
Residue 1 - ASN
Residue 2 - LEU
Residue 3 - TYR
Residue 4 - ILE
Residue 5 - GLN
Residue 6 - TRP
Residue 7 - LEU
Residue 8 - LYS
Residue 9 - ASP
Residue 10 - GLY
Residue 11 - GLY
Residue 12 - PRO
Residue 13 - SER
Residue 14 - SER
Residue 15 - GLY
Residue 16 - ARG
Residue 17 - PRO
Residue 18 - PRO
Residue 19 - PRO
Residue 20 - SER
```

Figure 3: The list of all 20 residues in Trp-cage protein