readmeENGapCV

28 December 2024

The following document includes two sections:

- Theory behind the code.
- How to use the code.

1 The Theory behind the code

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, the ENGap Collective Variable was created to be used as a biasing potential integrated into PLUMED.

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. Meli et al. (1).

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

where the spectral standard deviation is:

$$\mathrm{SDENG}(t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2} \tag{2}$$

ENGap Collective Variable Computation

The CV is computed as:

$$\mathrm{CV}(t) = \alpha(t)\cdot\mathrm{ENG}(t) - \beta(t)\cdot\mathrm{SDENG}(t) \tag{3}$$

where $\alpha(t)$ and $\beta(t)$ are weights modulating the ENG(t) and SDENG(t), respectively. The weights α and β are dynamically adjusted based on computed ENG and SDENG values:

- $\alpha = \alpha \times 1.1 \quad \text{if} \quad \text{ENG}(t) > \text{threshold}_{\text{ENG}}$
- $\beta = \beta \times 0.9 \quad \text{if} \quad \text{SDENG}(t) < \text{threshold}_{\text{SDENG}}$

The threshold values are defined in Meli et al. (1).

To integrate this CV into a biasing potential in plumed, it first needs to be differentiated with respect to the atomic positions r as follows:

$$\frac{\partial \text{CV}}{\partial \mathbf{r}} = \alpha(t) \cdot \frac{\partial \text{ENG}}{\partial \lambda_i} \cdot \frac{\partial \lambda_i}{\partial M} \cdot \frac{\partial M}{\partial D} \cdot \frac{\partial D}{\partial \mathbf{r}} - \beta(t) \cdot \frac{\partial \text{SDENG}}{\partial \lambda_i} \cdot \frac{\partial \lambda_i}{\partial M} \cdot \frac{\partial M}{\partial D} \cdot \frac{\partial D}{\partial \mathbf{r}} \quad (4)$$

Where:

- α: weight modulating the influence of ENG indicating the sensitivity of ENG to eigenvalues.
- β: weight modulating the influence of SDENG indicating the sensitivity of SDENG to eigenvalues.
- λ_i: is the eigenvalue.
- M: the energy matrix computed using Lennard-Jones and Coulombic interaction models.
- D: Standard Euclidean distances matrix

2 How to use the code

This code is designed to analyze molecular dynamics simulation data by processing GROMACS-generated TPR and XTC files using the MDAnalysis library. The core functionality is based on extracting protein structure and atomic properties from these files and converting them into arrays suitable for efficient computation with JAX.

2.1 Code's input files

To use the code, the following files are needed:

```
tpr_file = "/path/to/your/tpr/file"
xtc_file = "/path/to/your/xtc/file"
lj_contents = "/path/to/your/tpr_contents.txt"
```

The *lj_contents* file is the topology file converted into a txt so it can be read using ""regular expression Python Module.

2.2 Code's Output

The output of the code provides values computed for each frame of a molecular dynamics trajectory. For each frame, the following quantities are displayed:

- **CV(T)**: The value of the Collective Variable (CV) at the given frame. This is a linear combination of the Energy Gap (ENG) and the Spectral Standard Deviation (SDENG), modulated by the weights α and β.
- **ENG(T)**: The Energy Gap between the two lowest eigenvalues of the interaction energy matrix for the current frame.
- **SDENG(T)**: The Spectral Standard Deviation of the eigenvalues for the current frame.
- **Alpha** (α): The weight applied to the ENG term in the CV calculation. It is dynamically adjusted based on the computed ENG value.
- **Beta** (β): The weight applied to the SDENG term in the CV calculation. It is dynamically adjusted based on the computed SDENG value.
- **Gradient Norm**: The norm of the gradient of the CV with respect to atomic positions for the current frame. This indicates how sensitive the CV is to changes in atomic positions.

References

[1] Meli M, Morra G, Colombo G. Simple Model of Protein Energetics To Identify Ab Initio Folding Transitions from All-Atom MD Simulations of Proteins. *J Chem Theory Comput.* 2020;16(9):5960-5971. doi:10.1021/acs.jctc.0c00524.