Analysis of ATP-interacting residues for their amino acid frequency

Animesh sahoo

ID-2017A8PS0976G

Introduction:

Amino acids have high or low binding with ATP depending on various factors (e.g. shape, charge, surface area). Various studies have been performed to understand the binding behaviour of different amino acids with the ATP. But here only Few analytical methods have been used to predict the preference of interacting and non-interacting amino acids with ligand such as ATP based on the frequency of interacting and non interacting residues

Implementation :

Workflow:

- Importing all the necessary libraries and modules
- Downloading the dataset
- Modifying the dataset into fasta format manually and stripping the unnecessary data
- Reading the dataset and extracting sequence name, sequence and corresponding labels("0'=non-interacting and "1'=interacting)
- Calculating the frequency for each residue for both interacting and non-interacting per sequence
- Converting the dictionaries into CSV format and displaying the bar graph plot and the table for each sequence
- Calculating the cumulative frequency for each interacting and non-interacting residue in the whole dataset and displaying the table and bar graph plot
- Calculating the AA composition percentage for interacting residues and plotting the bar graph and analysing it
- Calculating the Propensity score for each interacting residue and plotting the bar graph to analyse it.

Results:

The Cumulative interacting residue frequency for each residue from the whole dataset is computed and displayed in the code. Analysis is done from Figure 1 .As it is clearly visible that K AA has maximum interacting frequency followed by E,T and G . While C,W,Q,M being the least interactive .

[Fig 1 about here.]

I compared the AA residues based on the interacting amino acid composition computed using Equation (1) .It shows the composition of interacting residues by a bar graph. Figure 2 depicts the percentage of composition of different interacting AA residues with ATP It is clear that K amino acid has the highest frequency interaction with ATP followed by E, T and G amino acids. On the other hand, C, M, Q, A,F and W amino acids show the least frequency. P, D,V , R,H, I, K, L, N and Y amino acids show moderate interaction with ATP.

[Fig 2 about here.]

I compared all these nucleotides based on their propensity score obtained using Equation (2) . Figure 3 shows the propensity score of different AA residues as bar plot consists of all 20 amino acids. ATP has the least preference for C, A, M, Q, F, P,Y, Hand N amino acids while having a high preference or propensity score for G, D, V, T, S,I,R,L, K and E amino acids. The analysis suggests that the propensity score is not dependent on the chemical property or size of the amino acid.Rather,the propensity based prediction method assign the propensity score of the residue using equation 2. The region of amino acids having the highest propensity score has a higher probability to interact with the ATP. The propensities based method is recommended for analysis/understanding of interacting residues in a given sequence. For this dataset Propensity score hypothesis which I have applied is if the Propensity score is > 5 , the residue is preferred otherwise not.

[Fig 3 about here.]

Summary:

Analysis of ATP interacting residues is performed using a dataset of 5 sequences.To determine the which residues are mostlikely interactive some analytical approaches were taken .The total cummulative frequency of binding and nonbinding sites for each AA residue is calculated and bar graphs are plotted and analysed.Along with it percentage composition of interacting AA residue is calculated and propensity score statistic is used to predict the probable binding region for ATP. The propensity-based method can predict the probable interacting region for ATP. Simply, the regions having highest propensity scores have the highest probability to interact with ATP. These results may help biologist in better understanding ATP interacting regions. Considering that the analysis has been done on a very small dataset we cannot accurately predict the maximum interacting residue to be alway true but however , We can summarize from our analysis that among AA residues like(K,G,E,T)one will be having the maximum interacting frequency and will mostlikely be binding with ATP .

Materials and methods

Manipulating dataset:

- Many data types we want to work with in bioinformatics are stored as tabular plain text files. So therefore to work with the required dataset the data needs to be filtered and modified without manipulating the residues. For the assignment dataset was obtained from ATPbindingProtien.txt . To which some modifications were made like use of ">" symbol before the sequence name and removing unnecessary space characters .
- We extracted the all the context from the dataset where the first line is the protein ID, chain and primary sequence, and the next line is the label, where "0"stands for non-binding and "1" stands for binding.
- All the characters apart from AA single letter code are neglected from the dataset while doing the analysis .

Mathematical Equations:

The percent amino acids composition of interacting residues is calculated using equation

$$RCi = \frac{Ri * 100}{Ni} \tag{1}$$

Where RC_i is the percent composition of a residue of type i, R_i is the number of residues of type i, and **N** is the total the number of all twenty interacting residues.

Residues propensity for AA is computed from the below formula

$$RPi = \frac{Ri \times 100}{Ni} \tag{2}$$

Where $\mathbf{RP}_i$ is AA propensity score for residue type **i**, $\mathbf{R_i}$ is number of interacting residues of type **i** and $\mathbf{N_i}$ is the total number of residues (interacting and non-interacting) of type **i**.

BarGraphs and Tables:

Table and AA frequency(interacting as well as non-interacting) vs AA is plotted for each sequence for detailed per sequence analysis. Along with it the AA composition, Propensity score , cumulative interactive and non interactive AA frequency data were also displayed along with respective bar graphs

Abbreviations:

AA-Amino Acids

[1–3]

References

1. Chauhan J, Mishra N, Raghava G. Identification of ATP binding residues of a protein from its primary sequence. BMC Bioinformatics;2009(10).

2. Singh H, Srivastava HK. Raghava GP. A web server for analysis, comparison and prediction of protein ligand binding sites. Biol Direct. 2016 3;2016(11):25.

3. Chen K, Mizianty MJ, Kurgan L. ATPsite: sequence-based prediction of ATP-binding residues. Proteome Sci; 2011. Suppl 1(Suppl 1):S4. Published.

Author biography

Animesh sahoo ID-2017A8PS0976G
BITS-pilani goa campus

List of Figures

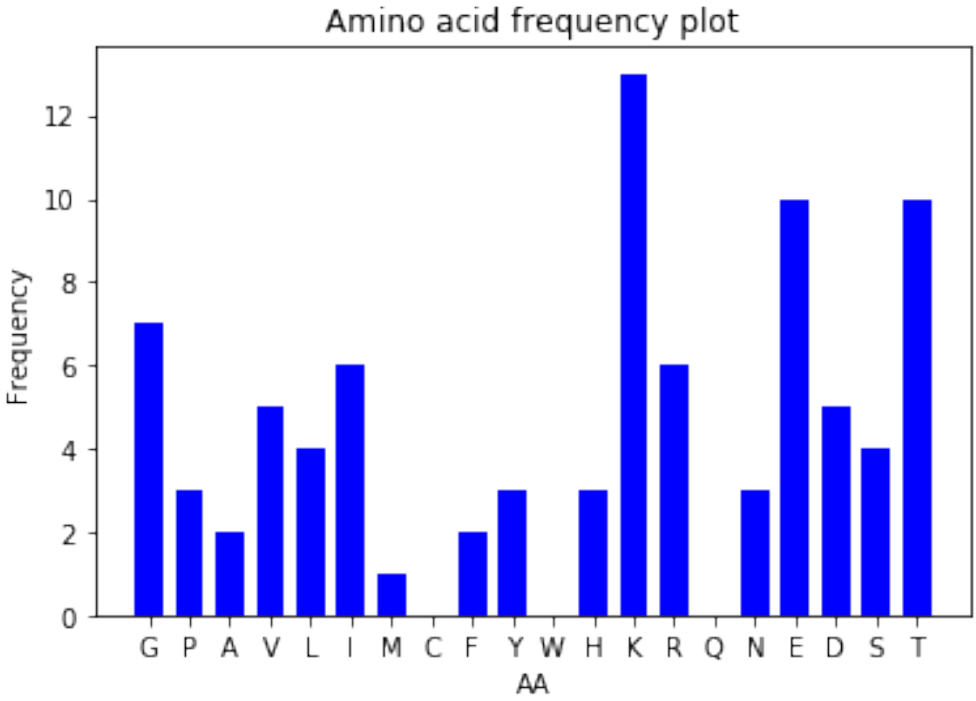

Fig 1. Bar graph showing the frequency of interacting amino acid residues

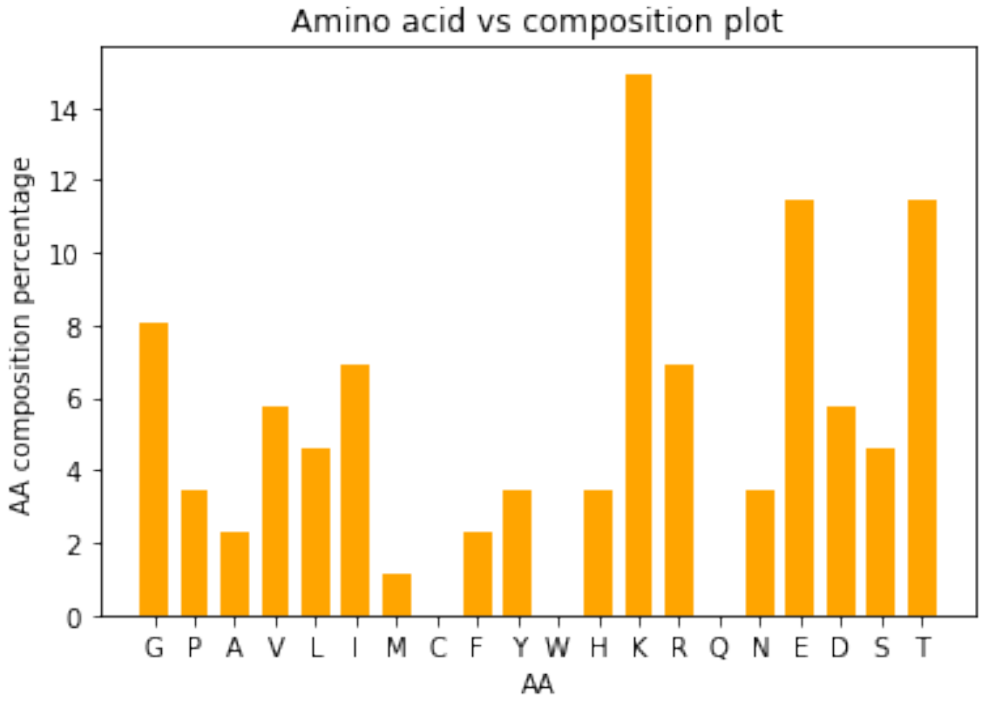

Fig 2. Bar graph representing the percentage of interacting amino acid composition.

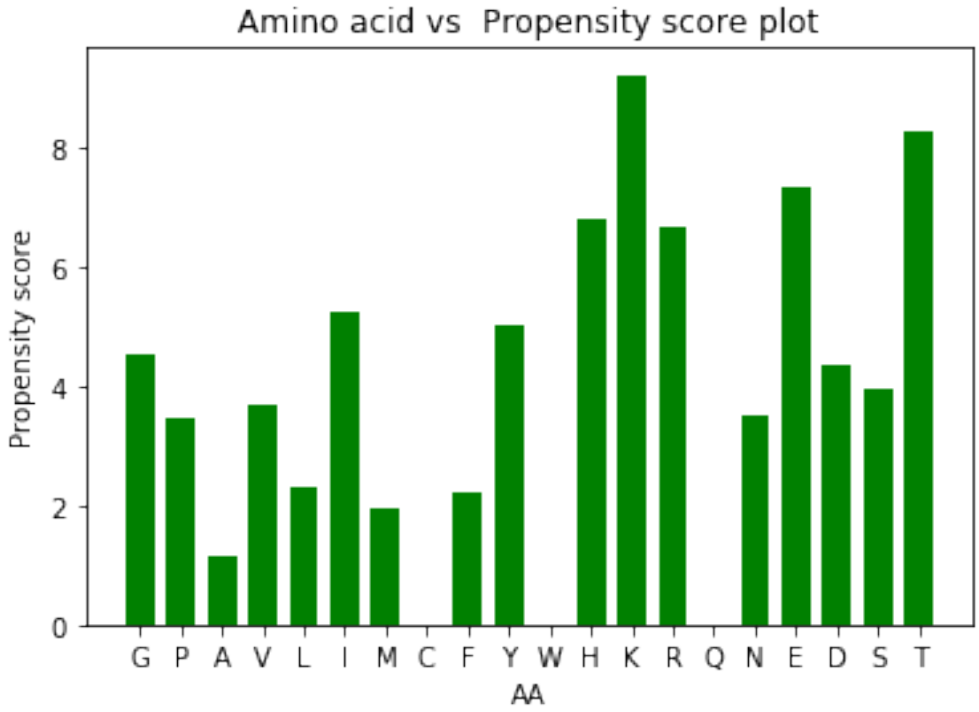

Fig 3. Bar graph representing Propensity score for each amino acid residue