### **Indian Institute of Technology Kharagpur**



### **Computational Biophysics**

***Term Project:*** *Amino Acid Contact (AAC) matrix generation*

***Group id:*** *27*

***Details of group members:***

### Astitva (20CS30007)

### Saikat Moi (20CS10050)

### Rohit Kumar Prajapati (20CS30041)

### **Problem Statement**

Generate Amino Acid Contact Matrix.

Consider 20 amino acid types and one solvent contacting residues in protein surfaces. The Amino Acid Contact (AAC) matrix is obtained from the statistical analysis of residue-pairing frequencies in one protein–protein complex database.

Select complexes from the Protein Data Bank. These complexes are made up of two or more protein subunits and their structures are determined by X-rays with cutoff values of resolution 2.2 Å and sequence identity 30%.

Define a pair of residues from two subunits as a contact pair, if two atoms (one from each subunit) are within distance d (set to be 6).”

* **Abstract**

This project focuses on understanding how amino acids, the building blocks of proteins, interact with each other and with solvent molecules on the surface of proteins. We used a large database, the Protein Data Bank, to analyze patterns of how these amino acids come into contact. Our study included a detailed examination of more than 4000 protein complexes, which are groups of proteins that function together.

We chose only the highest-quality data from the database by selecting protein complexes with very clear images and unique characteristics. For our analysis, we looked closely at the distances between atoms in different amino acids to identify which ones are close enough to be considered "in contact" with each other.

The main output of our work is a matrix, or a grid-like representation, that shows how often each pair of amino acid types is found in contact. This matrix helps us see which amino acid interactions are common and could be important for how proteins work.

Our study makes it easier for other scientists to understand protein interactions. This information can be very useful for designing new experiments and developing drugs, as it gives insight into the ways proteins behave in the body, which is important in health and disease.

### **Introduction**

Proteins are essential molecules in our bodies that carry out a vast array of functions, from speeding up chemical reactions to supporting the immune system. To perform these functions, proteins often work together by interacting at their surfaces. These interactions are like a special language spoken between proteins, where the letters are amino acids, the building blocks of proteins.

In this project, we wanted to understand this language better by looking at how often different amino acids are found close to each other, which we call “contacts,” when proteins interact. We were particularly interested in how these amino acids interact with water molecules, which are abundant in the body and play a crucial role in many biological processes.

To do this, we turned to a large, public collection of protein structures known as the Protein Data Bank (PDB). We carefully chose over 4000 protein complexes with high-quality data, meaning they had clear, detailed images and varied enough sequences to give us a broad look at many types of protein interactions.

Our goal was to create a detailed map of amino acid contacts. This map, or AAC matrix, could help us and other scientists predict how proteins will interact and understand the principles that govern these interactions. This knowledge is not just academically interesting—it could also be used to design new drugs or to figure out how diseases that involve protein malfunctions develop.

### **Materials and Methods**

**Data Collection**: To study protein interactions, we first needed a source of protein data. We used the Protein Data Bank (PDB), which is like a huge online library that stores 3D shapes of proteins. PDB IDs were read from a text file, and a total of 4172 complexes were considered for analysis.

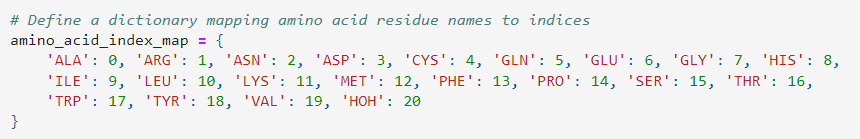
**Selection Criteria for Protein Complexes**: For our analysis, we selected protein complexes that consist of two or more protein subunits. The structures of these complexes were determined using X-ray crystallography, with cutoff values of resolution at 2.2 Ångströms and sequence identity at 30%. These criteria ensure that we focus on high-quality, relevant structural data for our study.

**Definition of Contact Pairs**: In our study, a pair of residues from two different subunits is defined as a contact pair if the distance between any two atoms (one from each subunit) is within 6 Ångströms. This definition is critical as it determines which amino acid interactions are considered significant and influences our analysis of protein interactions.

**Data Processing**: Each protein is made up of different amino acids, and we wanted to track how often each amino acid is in contact with others. To do this, we made a list that gave each amino acid a unique number, kind of like a name tag, so we could keep track of them easily. We included water in our list because it's not just the amino acids that are important in protein interactions—water plays a big part too.

### **Amino Acid Index Mapping**

To analyze proteins effectively, we assigned each amino acid a unique number. This is similar to assigning each student in a school a locker number. Our list included all 20 standard amino acids, like Alanine (ALA) as number 0 and Valine (VAL) as number 19. We also gave water, often found in and around proteins, its own number: 20. This system turned complex protein structures into a simple numerical form that our computer programs could easily track and analyze, recording when and how often each amino acid came into contact with others.



### **Amino Acid Contact Calculation**

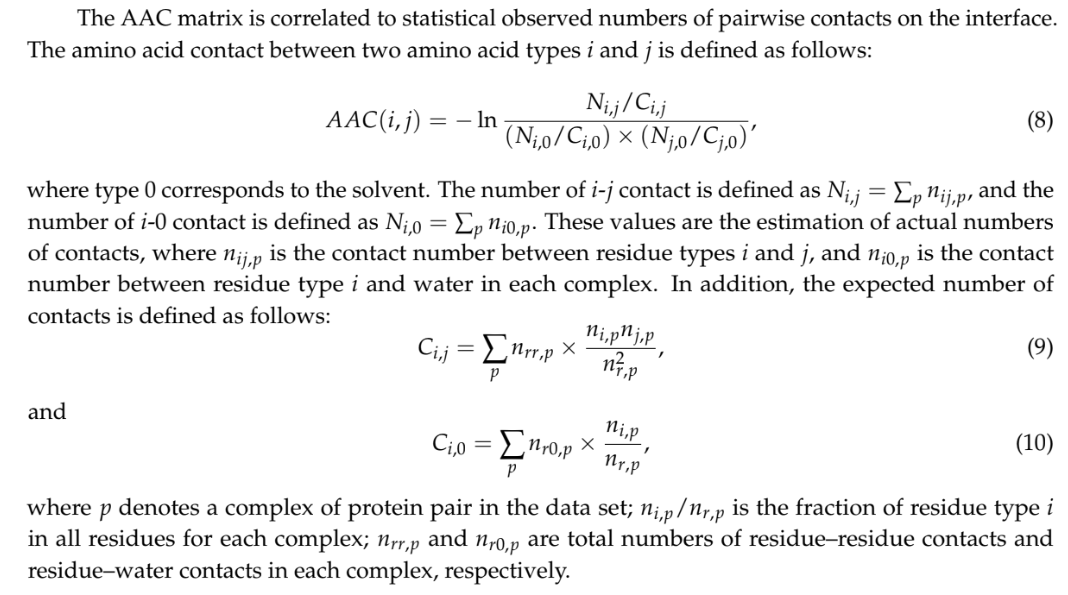
### To determine if two atoms from different amino acids are in contact, we use a cutoff distance of 6.0 Angstroms. This distance is like a boundary line—if the distance between the atoms is less than or equal to 6.0 Angstroms, we consider them to be in contact. We calculate the distance between two atoms using a mathematical formula called the Euclidean distance.

### Our method first calculates the distance between the coordinates of the two atoms using the numpy library's np.linalg.norm() function. If this distance is less than or equal to our cutoff distance, we conclude that the atoms are in contact and return True. Otherwise, we return False. This process allows us to efficiently determine which atom pairs are in contact within a protein structure.

### **Amino Acid Contact Matrix Calculation**

The AAC matrix quantifies how often different amino acid pairs interact within protein complexes. It's calculated using observed and expected contact numbers. Observed contact numbers (Ni,j) count actual interactions between amino acid types i and j, while expected contact numbers (Ci,j) consider amino acid fractions within complexes. This matrix helps us understand the frequency and significance of amino acid interactions in protein interfaces.

We used the below information from the Referenced Research Paper.



### 

### **Computing the Amino Acid Contact (AAC) Matrix**

### The AAC matrix quantifies the statistical significance of amino acid interactions within protein complexes. Our method, compute\_AAC\_matrix, calculates this matrix using observed and expected contact numbers.

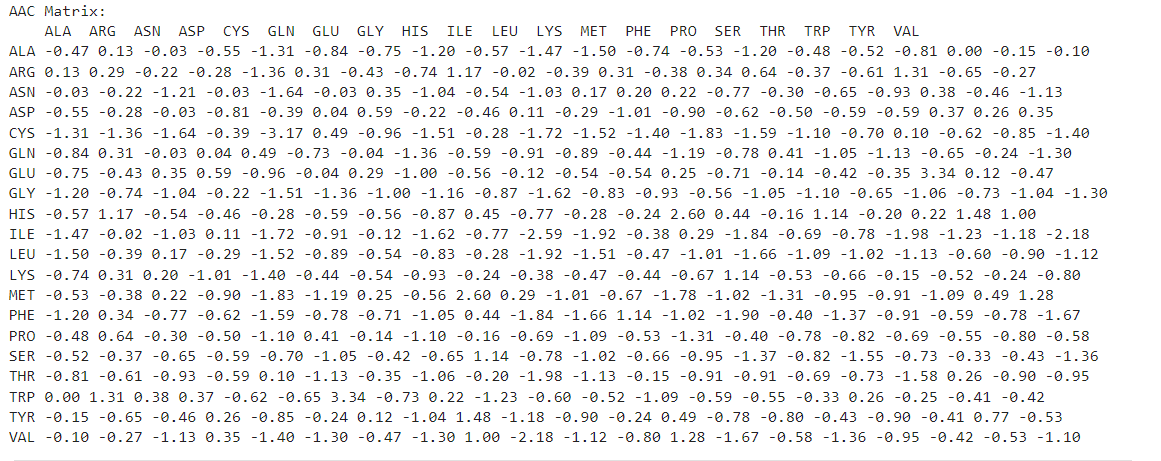
### Given the C matrix representing expected contact numbers between amino acid pairs, the C vector representing expected contact numbers with the solvent, and the matrix of observed contact frequencies, we iterate over each amino acid pair (i, j).

### Using a formula involving observed and expected contact numbers, we compute the AAC value for each pair. This value reflects the likelihood of interaction between amino acids, considering their frequency in the complex and their interactions with the solvent.

### The AAC matrix provides valuable insights into amino acid interactions within proteins, aiding in the understanding of protein-protein interfaces and their functional roles.

### 

### **Results**

**

### 

### **Discussion**

The **Amino Acid Contact (AAC) matrix** you mentioned provides valuable insights into the frequency and distribution of amino acid pairings in protein complexes. This information can be crucial in understanding protein structure and function, and it has significant implications in biophysical processes and drug design.

1. **Biophysical Processes**: The AAC matrix can help us understand how proteins fold into their 3D structures. Proteins are made up of amino acids, and the way these amino acids interact with each other influences how the protein folds. Understanding these interactions can help us predict protein structure from sequence, a fundamental problem in molecular biology.
2. **Drug Design**: In drug design, the AAC matrix can provide insights into which amino acid interactions are critical for a protein’s function. If a particular amino acid interaction is essential for the function of a disease-related protein, a drug could be designed to disrupt this interaction, thereby inhibiting the protein’s function and potentially treating the disease.
3. **Protein Engineering**: In protein engineering, where the goal is to design new proteins with desired functions, understanding these interaction patterns can guide the selection of amino acids at specific positions in the protein sequence.
4. **Protein-Protein Interactions**: The AAC matrix can also shed light on protein-protein interactions. If certain amino acid pairs are frequently in contact, it might indicate that these pairs play a key role in protein-protein interactions.

### **Conclusion**

In conclusion, this project has provided valuable insights into the intricate world of protein-protein interactions through the generation and analysis of the Amino Acid Contact (AAC) matrix.

By leveraging data from the Protein Data Bank (PDB) and employing rigorous computational methods, we have quantified the frequency and significance of amino acid interactions within protein complexes. The AAC matrix serves as a powerful tool for understanding the structural and functional dynamics of proteins at the molecular level.

Our findings shed light on the patterns and preferences of amino acid interactions, revealing both common and less frequent pairings that may play important roles in protein function and regulation. Moreover, the normalization of contact frequencies by expected contact numbers with the solvent provides a nuanced perspective on the biological relevance of these interactions.

The insights gained from this study have implications across various fields, including structural biology, drug discovery, and biotechnology. The AAC matrix can inform the design of experiments aimed at elucidating protein-protein interfaces and aid in the rational design of therapeutics targeting specific protein complexes.

Looking ahead, future research could explore more complex protein systems, incorporate dynamic aspects of protein interactions, and integrate experimental validation to further refine our understanding of protein-protein interactions.

In summary, this project contributes to the broader goal of deciphering the molecular mechanisms underlying biological processes and holds promise for advancing our understanding of protein function and disease mechanisms.

### **References**

*Ding, Yijie, Jijun Tang, and Fei Guo. "Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information." International journal of molecular sciences 17, no. 10 (2016): 1623.*