CS529: Visual Data Science

Visualization of Amino acid protein structure

Venkata Laxmi Mounika Batchu (vbatch2@uic.edu)

Department of Computer Science

University of Illinois at Chicago

Client: Boris Igic, *Assistant Professor*, Department of Biological Sciences, UIC

Team members:

Venkata Laxmi Mounika Batchu

Junaid Mohammad Shaik

Harsh Mishra

Interviewer: Venkata Laxmi es Mounika Batchu, *Graduate Student*, Department of Computer Science, UIC

Location: 1168 ERF, UIC

Date: October 5, 2022

Duration: 40 minutes

Abstract

This project proposes an illustration of the novelties of amino acids. This project aims to develop a tool that enables users to investigate the regions of amino acids where variability is quite high.

This Visualization helps the faculty to identify the variations in the amino acids folds and be able to find what place of Protein will have the most variation in the amino acids folds.

The recommended solution will help the users to have knowledge on the location where pollination takes place as a result.

Introduction

There are male and female sex organs in every plant. These regions depend on the wind, water, or any other natural type of pollination for Polin exporting. This pollination takes place in areas with significant levels of diversity. We need to identify areas with minimal variability. Folded amino acids should be highlighted if they differ in any way from other amino acid components. Visualize the specific region in the given sequence by selecting the genome sequence of the given variation.

The data contains a variety of amino acid sequences that require various forms of visualization. The choice of the genome sequence can be made arbitrarily or programmatically, which builds a pattern for the specified sequence and displays the visualization of folding.

Aim of the project

The project's goal is to develop a method of visualization that will let the faculty see how the Protein folds are happening in the amino acids. This Visualization also helps in assisting the faculty in identifying what place of Protein will have the most variation in the amino acids folds.

Significance

The significance of this project will help the users to understand the area where pollination occurs as a result. This project emphasizes the research challenges from topics related to:

- Large Data Visualization, Scalable data representations
- Representation of data and knowledge
- Visual design process, Visual data mining and visual knowledge discovery
- Working with personal or social data

Requirement Analysis

I. **Humans**

- All faculty and students in the biology department.
- All scientists and students interested in working with plants from the Botany department.
- Users would access when necessary.

II. Tasks

High level tasks:

- Navigate the visualized graph of high novelty folded amino acids and analyze the variability of all parts of the amino acid.
- Locate a region with a high degree of variability and display some difference from the amino acid's other components in appearance.
- The amino acid sequence has a very long length. The triplets of the sequence should be gathered, which is exceedingly difficult.

Detailed tasks and activities:

- Cleaning the data set and getting the required columns
- First, we have to separate sequence data of all the amino acids from the given data.
- Finding the folds in the protein structure
- Coloring the protein structure based on the high variability.
- We can save all the sequences which are already done with visualization
- The data points of the sequence should be marked in the protein structure
- Making the triplets of the amino acids.
- Creating the folded amino acids variables to the desired folds
- Finding where there is more variability in the folded amino acids
- Based on the variability, designing the choropleth on the amino acid.
- Showing the 3-D structure of the amino acids.
- Adding the user interactive features to the visualization like Zoom in, Zoom out and more.
- There are many amino acids that have to do the above mentioned steps to all the amino acids.

III. Data

- The data of this project is from the biology department at UIC.
- The data in the project can be shared among the UIC staff and Students but not for the public use.
- We have Categorical and original data in this dataset.
- Since it is a huge dataset there might be some missing values.

IV. Flow

- In the main page the user sees all the proteins' icons in SVGs and have an option to select any of the displaying projects.
- When he clicks on a particular protein, all the different phases along will appear with their visualization.
- When hovering on a particular phase, the user can see the variability at each position of the amino acid.
- Users can read the summary below the visualization about the variation and pollination chances.

V. Non functional requirements

- ***Scalability:*** There is no scope of scalability in this project.
- ***Performance:*** The response times and the processing times should be fast.
- ***Privacy:*** This dataset is the collection of multiple datasets, and this collection is done by the professor Boris
- ***Accessibility:*** It should be accessible to the users who consult the professor
- ***Useability:*** Some biologists are working on this project so this can be used by all biologists.

VI. Probes

Is visualization appropriate for this project?

- Yes, visualization is definitely necessary for the representation of the folded amino acid. It gives an overview of the pollination details of the plant for biological researchers.

Any existing tools or research?

- A group of biologists are working on this project, and they are contributing to the python notebook.

Possible exploration

- The present visualization can also be extended by adding hybrid amino acids.

Work plan

The implementation process can be divided into the following processes:

- Analyze user requirements and data to figure out how to preprocess the data.
- Design the user interface and figure out what libraries we need to use.

- Start implementing front design first for the alpha release.
- Start implementing the visualization part.
- Implement other parts.
- Revisit the requirements and check everything is satisfied.
- Prepare for the final report and presentation/demo.
- Create documentation for users.

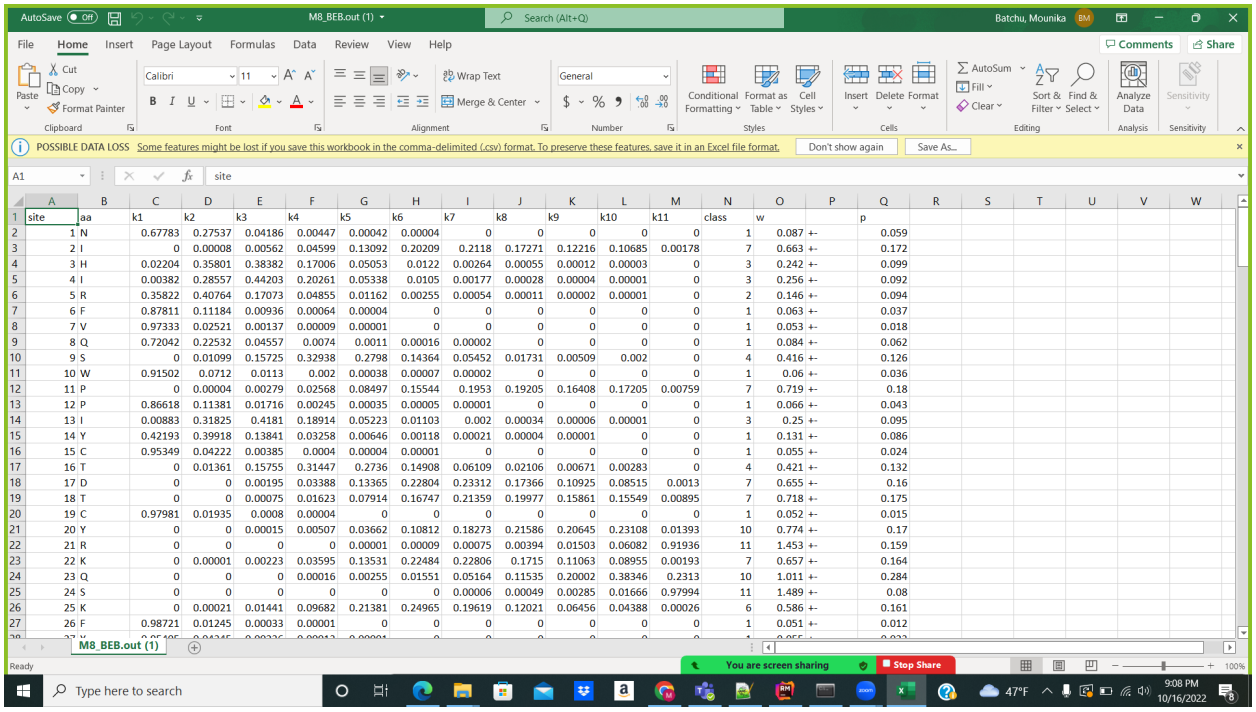
Facilities:

- A server to publish the web service (github.io or people.uic.edu can be used)
- Computer: team members will use their own laptops/desktops.
- Space: UIC campus

References

- AlphaFold2 Collab notebooks: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
- AlphaFold2 (protein folding): https://www.nature.com/articles/s41592-022-01488-1
- Viewer 1 (PyMol): https://pymol.org
- Viewer 2: https://www.rcsb.org/3d-view

Data Sample -



site	aa	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10	k11	class	w		p
1	N	0.67783	0.27537	0.04186	0.00447	0.00042	0.00004	0	0	0	0	0	1	0.087	+-	0.059
2	I	0	0.00008	0.00562	0.04599	0.13092	0.20209	0.2118	0.17271	0.12216	0.10685	0.00178	7	0.663	+-	0.172
3	H	0.02204	0.35801	0.38382	0.17006	0.05053	0.0122	0.00264	0.00055	0.00012	0.00003	0	3	0.242	+-	0.099
4	I	0.00382	0.28557	0.44203	0.20261	0.05338	0.0105	0.00177	0.00028	0.00004	0.00001	0	3	0.256	+-	0.092
5	R	0.35822	0.40764	0.17073	0.04855	0.01162	0.00255	0.00054	0.00011	0.00002	0.00001	0	2	0.146	+-	0.094
6	F	0.87811	0.11184	0.00936	0.00064	0.00004	0	0	0	0	0	0	1	0.063	+-	0.037
7	V	0.97333	0.02521	0.00137	0.00009	0.00001	0	0	0	0	0	0	1	0.053	+-	0.018
8	Q	0.72042	0.22532	0.04557	0.0074	0.0011	0.00016	0.00002	0	0	0	0	1	0.084	+-	0.062
9	S	0	0.01099	0.15725	0.32938	0.2798	0.14364	0.05452	0.01731	0.00509	0.002	0	4	0.416	+-	0.126
10	W	0.91502	0.0712	0.0113	0.002	0.00038	0.00007	0.00002	0	0	0	0	1	0.06	+-	0.036
11	P	0	0.00004	0.00279	0.02568	0.08497	0.15544	0.1953	0.19205	0.16408	0.17205	0.00759	7	0.719	+-	0.18
12	P	0.86618	0.11381	0.01716	0.00245	0.00035	0.00005	0.00001	0	0	0	0	1	0.066	+-	0.043
13	I	0.00883	0.31825	0.4181	0.18914	0.05223	0.01103	0.002	0.00034	0.00006	0.00001	0	3	0.25	+-	0.095
14	Y	0.42193	0.39918	0.13841	0.03258	0.00646	0.00118	0.00021	0.00004	0.00001	0	0	1	0.131	+-	0.086
15	C	0.95349	0.04222	0.00385	0.0004	0.00004	0.00001	0	0	0	0	0	1	0.055	+-	0.024
16	T	0	0.01361	0.15755	0.31447	0.2736	0.14908	0.06109	0.02106	0.00671	0.00283	0	4	0.421	+-	0.132
17	D	0	0	0.00195	0.03388	0.13365	0.22804	0.23312	0.17366	0.10925	0.08515	0.0013	7	0.655	+-	0.16
18	T	0	0	0.00075	0.01623	0.07914	0.16747	0.21359	0.19977	0.15861	0.15549	0.00895	7	0.718	+-	0.175
19	C	0.97981	0.01935	0.0008	0.00004	0	0	0	0	0	0	0	1	0.052	+-	0.015
20	Y	0	0	0.00015	0.00507	0.03662	0.10812	0.18273	0.21586	0.20645	0.23108	0.01393	10	0.774	+-	0.17
21	R	0	0	0	0	0.00001	0.00009	0.00075	0.00394	0.01503	0.06082	0.91936	11	1.453	+-	0.159
22	K	0	0.00001	0.00223	0.03595	0.13531	0.22484	0.22806	0.1715	0.11063	0.08955	0.00193	7	0.657	+-	0.164
23	Q	0	0	0	0.00016	0.00255	0.01551	0.05164	0.11535	0.20002	0.38346	0.2313	10	1.011	+-	0.284
24	S	0	0	0	0	0	0	0.00006	0.00049	0.00285	0.01666	0.97994	11	1.489	+-	0.08
25	K	0	0.00021	0.01441	0.09682	0.21381	0.24965	0.19619	0.12021	0.06456	0.04388	0.00026	6	0.586	+-	0.161
26	F	0.98721	0.01245	0.00033	0.00001	0	0	0	0	0	0	0	1	0.051	+-	0.012

Tools : Code repository & Bug tracker : https://github.com/harshm16/VDS_project

Data and Tasks Abstraction

Data Abstraction

- Our main data set is a CSV file that contains 200 records of amino acid protein data. The link to the sample data set can be found below.
    - https://drive.google.com/file/d/1D2A7bfH30xUeoLmuQGnuCzcFo1YAbWaq/view?usp=sharing

The dataset contains below data types:

- **Categorical:**
    - **Class:** This data attribute specifies under which class a particular amino acid structure falls and there are a total of 11 different classes. (possible categorical values for each amino acid structure's class are 1,2,3.....11.)
    - **aa:** This attribute signifies the sequence of the amino acid structure and aa values are spread out totaling up to 26 values. (possible categorical values for each amino acid structure's aa are A, B, C, D, ... Z)

- **Numerical:**
    - **k1,k2,k3 .....k11:** All these attributes represent the coordinates of the amino acid protein structure. The values usually fall between 0 and 1(including decimals).
    - **w:** This data attribute denotes the variability factor. In other terms, It illustrates the highest changeability in a particular amino acid structure.
- The Dataset is in the form of a table with each record denoting the amino acid structure with its respective class, coordinates, and variability factor.
- The data is **static** as we are not using real-time data.

Task Abstraction

- Have to create a landing page where users can enter their input protein sequence for plotting the folded version of the amino acid.
- Have to validate the input sent by any user and should convert the data into the format that is feasible to construct the amino acid
- Have to create some interactions like providing the variability score on every sequence of the amino acid.
- Have to provide the color variation depending on the variability score of the particular sequence of the amino acid
- Have to find various sequencing techniques to generate sequences from the input data of the user. These sequences are useful for creating the amino acid protein
- Have to provide the user with access to find the places of the amino acid where it is variable and where it is constant.

- Have to provide the count of most variable sequences of the amino acid and the count of least variable sequences of the amino acid protein

Existing tools/systems for this problem (if any):

- Advantages of d3 over tableau
  - D3 focuses on data, so it is the most appropriate and specialized tool for data visualizations.
  - D3 is open-source. So you can work with the source code and add your own features.
  - It works with web standards so you don't need any other technology or plugin other than a browser to make use of D3.
  - D3 works with web standards like HTML, CSS and SVG, there is no new learning or debugging tool required to work on D3.
  - Since D3 is lightweight, and works directly with web standards, it is extremely fast and works well with large datasets.
- Three.js over Babylon.js
  - Lots of examples, git repositories, and tutorials
  - Easy to start
  - Big community
  - Lots of Stack overflow and forum posts for common issues

Visual encodings of the dataset types you have identified

- **Paper Title**: Visualization of Biomolecular Structures: State of the Art Revisited
  Link: https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.13072
  The paper contains a discussion on molecular structures, and strategies for efficient display regarding image quality and frame rate, covers different aspects of level of detail, and reviews visualizations illustrating the dynamic aspects of molecular simulation data. However, the paper doesn't stress enough major challenges like depicting physical phenomena in more detail, improving the perceptual and cognitive efficiency of visualizations, as well as depicting longer trajectories of larger molecular systems.
- **Paper Title**: Visual analysis of Biomolecular cavities
  Link: https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.12928
  The paper contains a novel classification for cavity detection approaches and structures them into four distinct classes: grid-based, Voronoi-based, surface-based, and probe-based methods. Then match these approaches with corresponding visualization technologies starting with direct 3D visualization, followed by non-spatial visualization techniques that for example abstract the interactions between structures into a relational graph. However, the paper fails to explore the current state of methods for the visual analysis of cavities in dynamic data such as molecular dynamics simulations.
- **Paper Title**: A Visualization system of Dynamic Protein Structure and Amino Acid Network.
  Link:https://www.researchgate.net/publication/319296617_A_Visualization_System_for_Dynamic_Protein_Structure_and_Amino_Acid_Network
  This paper provides insights into design and implements a visual analysis system of dynamic protein structure and amino acid network. Then uses the molecular simulation data of the b2AR protein to do experiments with the mixed layout method, dynamic visual analysis model, and visualization framework with multiple coordinated views. However, there is no discussion on how to analyze the link between protein sequence and structure and there is much emphasis on amino acid topology which deviates from our original project idea.
- **Paper Title**: Large Scale Analyses and Visualization of Adaptive Amino Acid Changes

Link: https://link.springer.com/article/10.1007/s12539-018-0282-7#Sec7
This paper contains a discussion on ADOPS which basically provide the nucleotide and protein sequences in FASTA format (aligned and non-aligned), which can be used for many other types of analyses, including the identification of invariant (likely functionally important) amino acid sites.However, The ADOPS project doesn't contain an option for adding new sequences to a given project, a tool that is certainly useful when the sequences that a given researcher needs are not all contained in the original ADOPS project.

- **Paper Title**: A Survey on 3D Virtual Object Manipulation: From the Desktop to Immersive Virtual Environments
  Link: https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.13390
  Here the authors talk about how different techniques can be used to manipulate 3D virtual objects efficiently. They survey the state-of-the-art techniques used for 3D object manipulation, talking about how they vary from traditional desktop methods, to touch and mid-air interactions. They emphasize on mimicking real life movements, such as translation, rotation and resizing and show different ways that state-of-the-art methods use in order to translate them in virtual environments.
- **Paper Title**: Symmetry in 3D Geometry: Extraction and Applications
  Link: https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.12010
  Geometric symmetries are very common in real world objects as well as their respective virtual representations. In this paper the authors discuss how exploiting these geometric symmetries benefits the understanding of the objects, as well as in terms of creating effective processing techniques for these objects. Based on their survey, they conclude that using algorithms which compress the images and classify them into a set number of distinct categories should be used when the size of the data is very huge. These compact representations once stored can then be used as a reference point to subsequently store deviations from the initial symmetrized shapes.
- **Paper Title**:: A Survey of 3D Interaction Techniques
  Link: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8659.00194
  The paper discusses how the development in terms of 3D graphics rendering systems and hardwares have not been in line with the improvement of user knowledge in terms of how to interact with these virtual environments. They explain how mouse-based techniques as well as 3D input devices help in user interaction with the 3D graphics. In addition to presenting a variety of 3D interaction approaches for object manipulation, viewpoint manipulation, and application control, the authors also provide an overview of how these techniques have developed through time.
- **Paper Title**: Tasks, Techniques, and Tools for Genomic Data Visualization
  Link: https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.13727
  Here the authors provide taxonomies for the information, the display, and the activities related to visualizing genetic data. They also provide a thorough analysis of available genomic visualization tools in relation to the suggested taxonomies. They explain how there is a need for privacy protections and safe data access and storage when it comes to genome visualizations. They also explain how researchers outside of this subject have a steep learning curve due to the complexity involved in the genomics datasets.
- **Paper Title**:cis-Regulatory element prediction in mammalian genomes:
  Link https://ieeexplore.ieee.org/document/1540599

  In this paper, they tried to understand the regulation of genes by identifying cis-regulatory elements and modules. They assembled the input sequence sets comprised of the upstream

regions of a target gene, its orthologues and co-expressed genes on the premise that such genes will share promoters by evolution or share regulatory control mechanism.

- **Paper Title**: Detection of Genome Sequence Outliers Across Pan-Genomes
  Link : https://ieeexplore.ieee.org/document/8883624

  Pan-genomes denote sets of all unique gene families found in multiple related genomes in a given taxon - for instance related strains of a bacterial species, thus representing the entire gene pool of the taxon.They demonstrated that characterizing a pan-genome of a given taxon using sequences generated from different genome projects can misguide subsequent genome comparison studies when a set of incorrect strains is selected as input.
- **Paper Title**:Clous processing of 1000 genomes sequencing data using Amazon Web Service:
  Link : https://ieeexplore.ieee.org/document/6736809.
  In this paper, genetic variant pipeline SNPTools is deployed in the cloud utilizing the Amazon Web Service (AWS). With the cloud SNPTools pipeline, they performed the SNP calling and genotype imputation on the 1000 Genomes Project Phase 3 data and assessed the quality of SNPs. The analysis shows that cloud computing will be indispensable to the Next Generation Sequencing data processing.
- **Paper Title**:Flowering plants pollination robotic system for greenhouse by means of nano copter(drone aircraft) :
  Link : https://ieeexplore.ieee.org/document/7751907
  The external environment that is capable to provide pollination of plants it is not enough for intensive agricultural production. In this regard there is a need of searching a new more efficient pollination instrument to increase productivity of crops. The article argues that the simulated pollination of agricultural plants by means of nano copter can provide collecting and delivering pollen in the mode of automatic control.

Interaction techniques/systems for the tasks you have identified

https://www.researchgate.net/figure/Visualization-of-A-the-plant-pollinator-interaction-network-from-DeBarros-2010-B_fig2_305795508

In this plant-pollinator interaction network, we don't have any interactions. We need some interactions like showing the variability score and changing the color of the particular sequence of the amino acid. So this visualization is not sufficient for our project.

Functional Specifications

1. **Scenario 1: Sequence viewer - Venkata Laxmi Mounika Batchu**

   Aslihan, a botany student, wants to know about the Natural version of the amino acid sequence. So, he opens our website and input the required protein name. Then, a sequence viewer appears. He can clearly see any part of the sequence by hovering on the desired location of the sequence. Also can view the length of the amino acid sequence with partitions. Aslihan can easily find the count of each alphabet in the sequence as different colors are given to the different alphabets of the sequence.

2. **Scenario 2: Comparison of multiple Amino Acids - Junaid Mohammad Shaik**

   Trehan, A botany PhD student, wants to do some research on amino acids and wants to view multiple amino acid structures at a time. Then he can open our website and provide the names of amino acids he wants to explore. Then all the amino acids folded versions pop up on the screen. So that he can dig deeper into every structure. He can examine the confidence levels of each amino acid in the graph we are showing. Some tool tips are available for analyzing all the amino acids in a single glance. Trehan can analyze the parts of the amino acid by giving the starting and ending index of the part.

1. **Scenario 3: Protein Structure Viewer - Harsh Mishra**

   Dejan, a Biological Research Professor, wants to show his students how Proteins are built as chains of amino acids, which fold into unique 3-D shapes. In order to get a 3-D visualization of these structures, all Dejan will have to do is to input the natural protein sequence in our application. Once he hits run, our application will then render a 3-D structure respective to the input sequence, which can be interactively analyzed by him and his students. The application will offer its users to zoom in and out in order to learn about the protein structure in greater detail, while also giving the users the information about the sub-structures via the tooltip.

Non-Goals

- Read the protein sequence from the input file
- Give the users the ability to save the visualizations

Non-functional Requirements

- Providing the feasibility that users can enter the amino acid input.
- Validating the user input and modifying according to the project.
- Good color scheme for the website.

Goals and Targets

Low Target:

The low level tasks are as follows:

- A visualization of the sequence viewer.
- Being able to visualize the 3-D version of Alpha fold protein structure
- Coloring according to the start and end index provided by the user.
- Visual representation of temperature factor.
- Being able to rotate the 3-D protein structure
- Being able to zoom in and out of the protein structure

Desirable Target:

- Being able to visualize the 3-D version of Alpha fold protein structure depending on the input given by the user
- Being able to provide some interactions using the sequence viewer.

High Target

- Being able to show multiple amino acids in a single UI.
- Being able to show comparisons of the multiple amino acids given by the user.
- Being able to learn about the sub-structures of the main structure viz., using tooltips

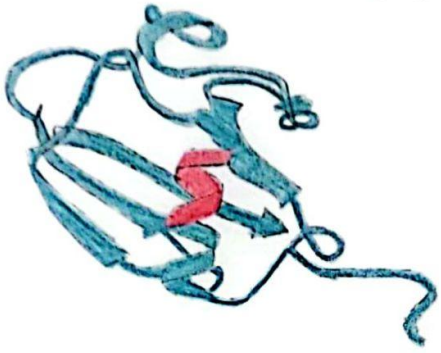
Extras

- The users being able to use a 2D plane to cut some parts of the protein structure in order to analyze them further

Prototypes -

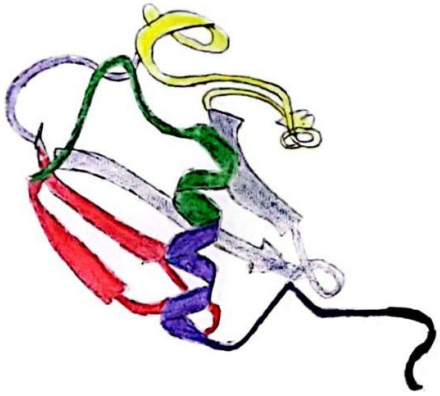
1. ARBITRARY USER SELECTION

SEQUENCE: NHIRFVQSWIPPIYCTBERYGITCYR



⇒ User is given a choice to select a part of sequence to color it with the default color (Red)

2. DIRECTIONALITY COLOR INDEX

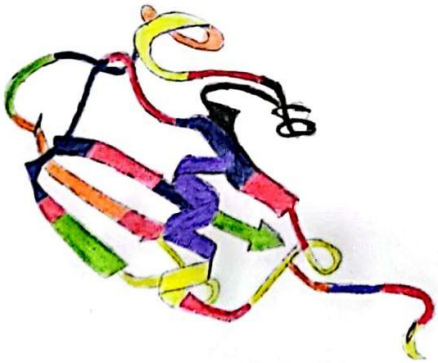


⇒ Colors are blended from start index to end index of amino acid sequence.

start index

end index

3. CLASS BASED COLOR



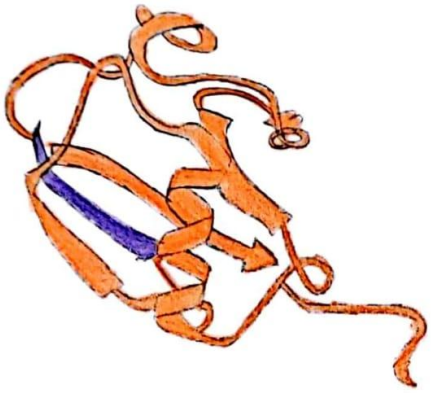
⇒ Each amino acid is colored based on the class that it belongs to.

Class: 1 2 3 4 5 6 7

4. PI (ISOELECTRIC POINT)

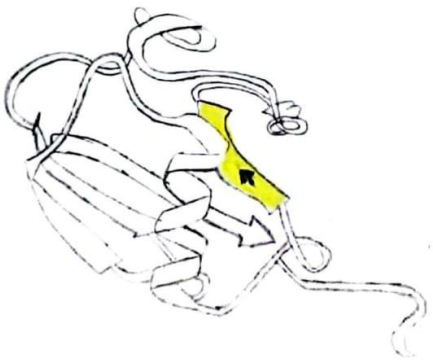
⇒ Each amino acid is colored based on pI values

5. DYNAMIC COLOR SELECTION



⇒ User can select a window of amino acids & can choose the color of his choice from the given colors.

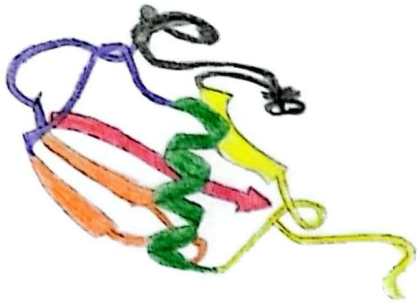
6. TRANSPARENT AMINO ACID



⇒ User can hover on a particular amino acid and colored amino acid sequence is highlighted with a colour while the rest still remains transparent.

CS Scanned with CamScanner

7. AVERAGE MODEL CONFIDENCE



- >110
- >90 & <110
- >70 & <90
- >50 & <70
- 25> & <50
- <25

Scanned with CamScanner

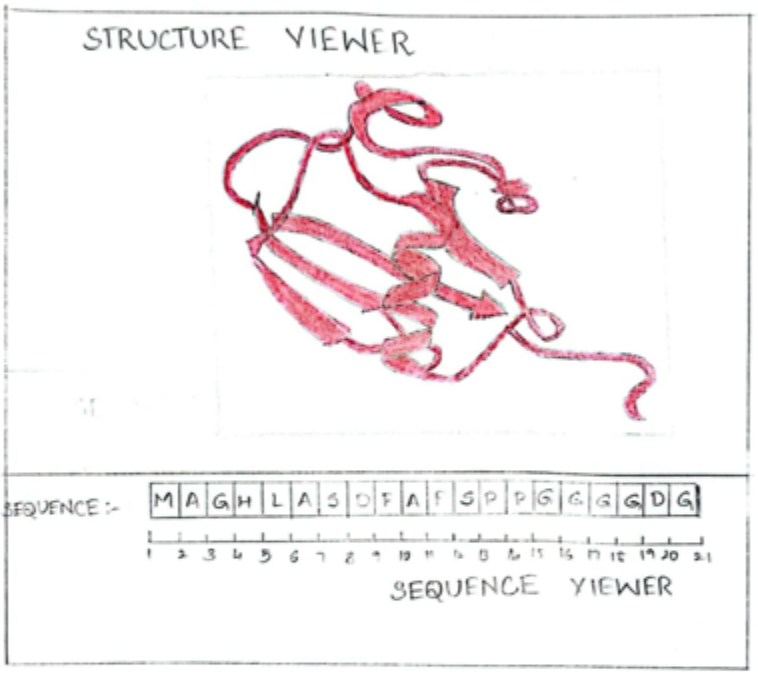
PROTOTYPE-I (VENKATA LAXMI MOUNIKA BATCHU)

SEQUENCE VIEWER :

⇒ Displays amino acid sequence given by the user.

STRUCTURE VIEWER

⇒ Displays entire amino acid fold structure.



Scanned with CamScanner

Discussion

pros:

providing the access for the user to give the input sequence

cons:

No any other feature like sequence, color, & PI selection

Focus

Input for Sequence

SEQUENCE :

PROTOTYPE-II (VENKATA LAXMI MOUNIKA BATCH(U)

⇒ User can provide the part which he wants to select from amino acid sequence.

⇒ the textfield validates the sequence upon clicking the submit button

⇒ If the sequence entered is valid then the structure viewer highlights the selected part in the 3D structure.

STRUCTURE VIEWER

SEQUENCE :

M	A	G	A	H	L	A	S	D	F	A	F	S	P	P	G	G	G
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

SEQUENCE PART : HLASDF

SUBMIT

Discussion point

Pros

* user can select any part of the sequence which he wants to view

* cons

zoom in feature is not there for detailed viewing

Focus

Highlighting

M	A	G	A	H	L	A	S	D	F	A	F	S	P	P	G	G	G

Highlighting the selected part

Scanned with CamScanner

Button

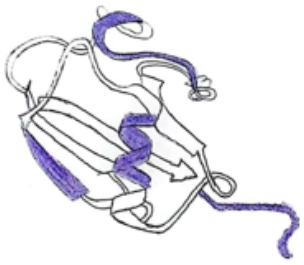
SUBMIT

PROTOTYPE-III (VENKATA LAXMI MOUNIKA BA...

⇒ Sequence viewer lets the user to select PI range (i.e:, low, medium and high).

⇒ Based on the user input, the structure viewer displays the 3D structure with highlighted amino acids with the given PI range

STRUCTURE VIEWER



Scanned with CamScanner

SEQUENCE VIEWER.

PI RANGE :-

LOW	MEDIUM	HIGH
○	◉	○

SUBMIT

Discussion points

Pros:

Isoelectric point of an amino acid is the point at which amino acid has not electric charge. Displaying radio buttons for this feature is a pro

Cons:

just provided low, medium, high instead of providing the starting and ending range of everything

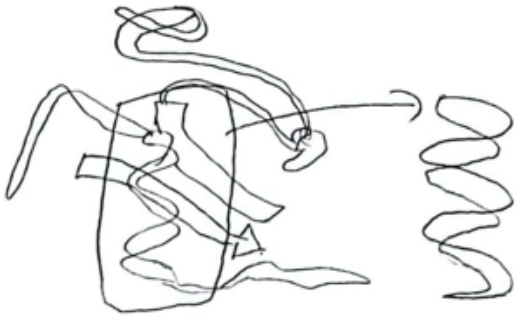
Focus

Radio buttons

Low Medium High

Button

SUBMIT



Name: Junaid Mohammad Shaik

PROTOTYPE - I

PROTOTYPE-1 (JUNAID MOHAMMAD SHAIK)

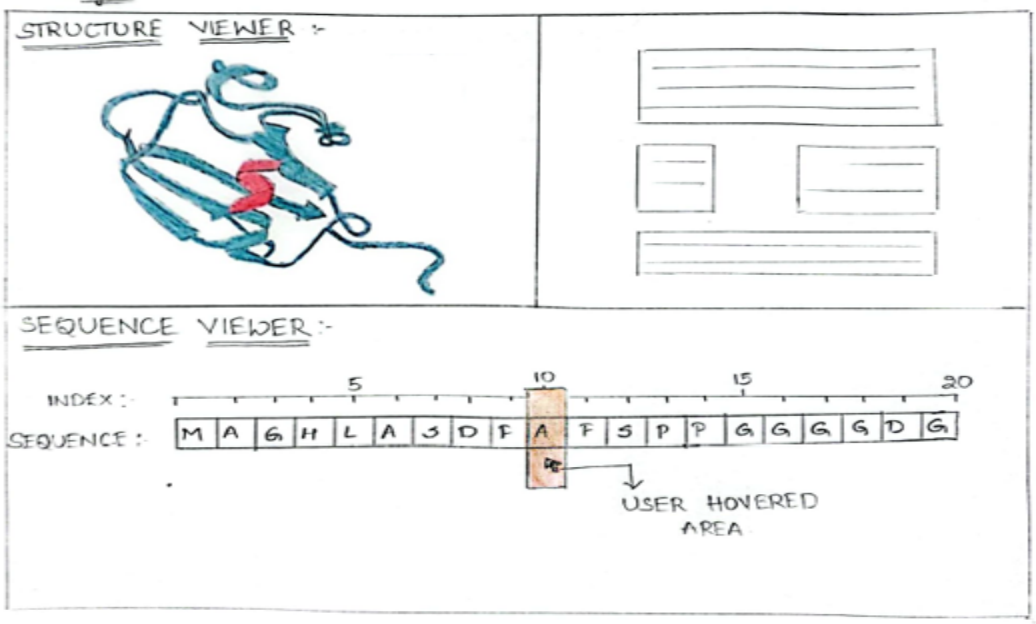
Operation:-

⇒ Sequence viewer displays the index of each amino acid

⇒ User can hover over a particular amino acid and it gets highlighted

⇒ Upon doing so, the selected amino acid gets highlighted in the structure viewer that displays 3D structure.

Layout:-



Discussion:-

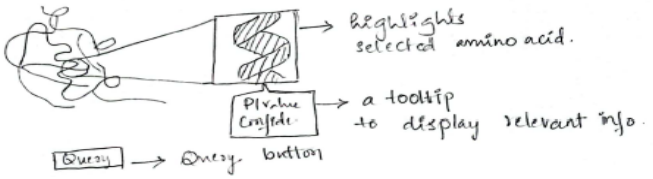
⇒ Pros:

- Displaying based on users choice of amino acid.
- Clear differentiation with respect to other amino acids

⇒ Cons.

- Only able to choose one amino acid at a time.
- No information regarding PI & Confidence.

Focus:-



PROTOTYPE-II (JUNAID MOHAMMAD SHAIK)

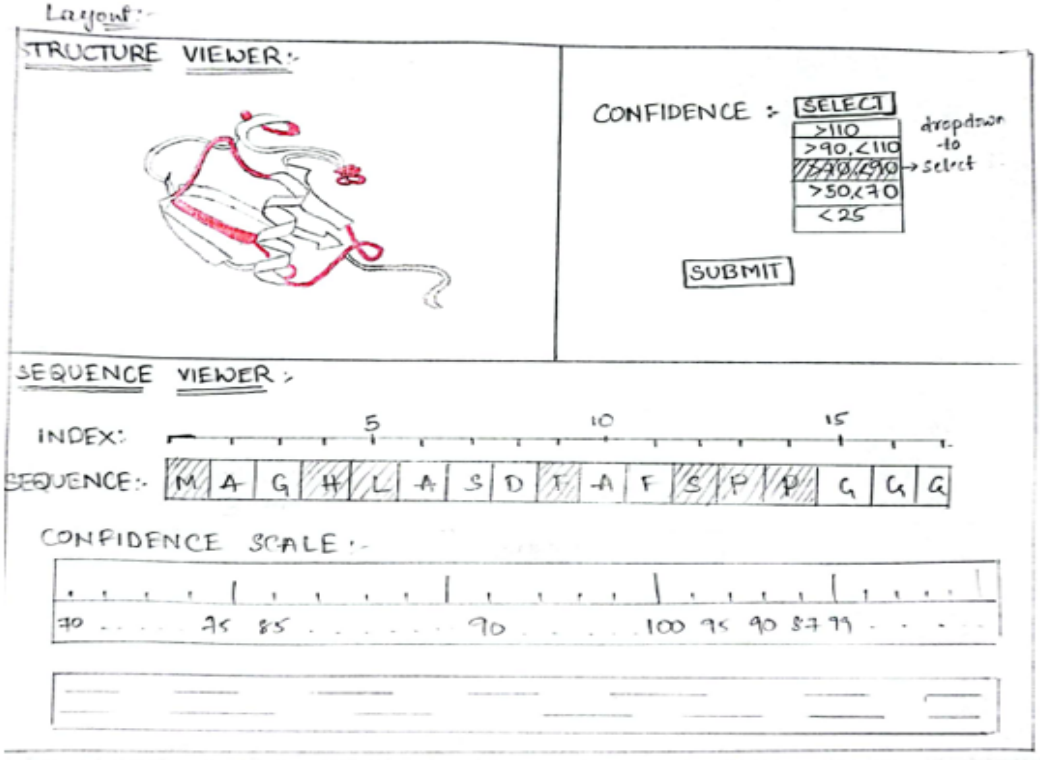
operation:-

⇒ Sequence viewer lets the user to view amino acid sequence which are indexed accordingly

⇒ User can select the confidence range from the dropdown

⇒ Structure viewer displays 3D structure that highlights the amino acid that falls within the given confidence range

Layout:-



Discussion:-

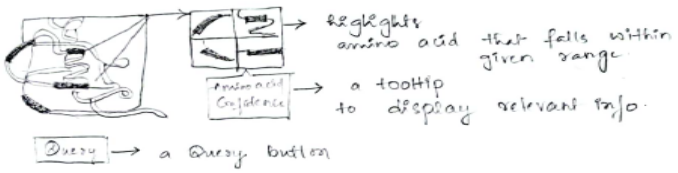
⇒ Pros:-

- User interaction is done based on confidence values thereby giving users the flexibility to choose amino acids based on confidence scores.
- Can select multiple amino acids at a given time.

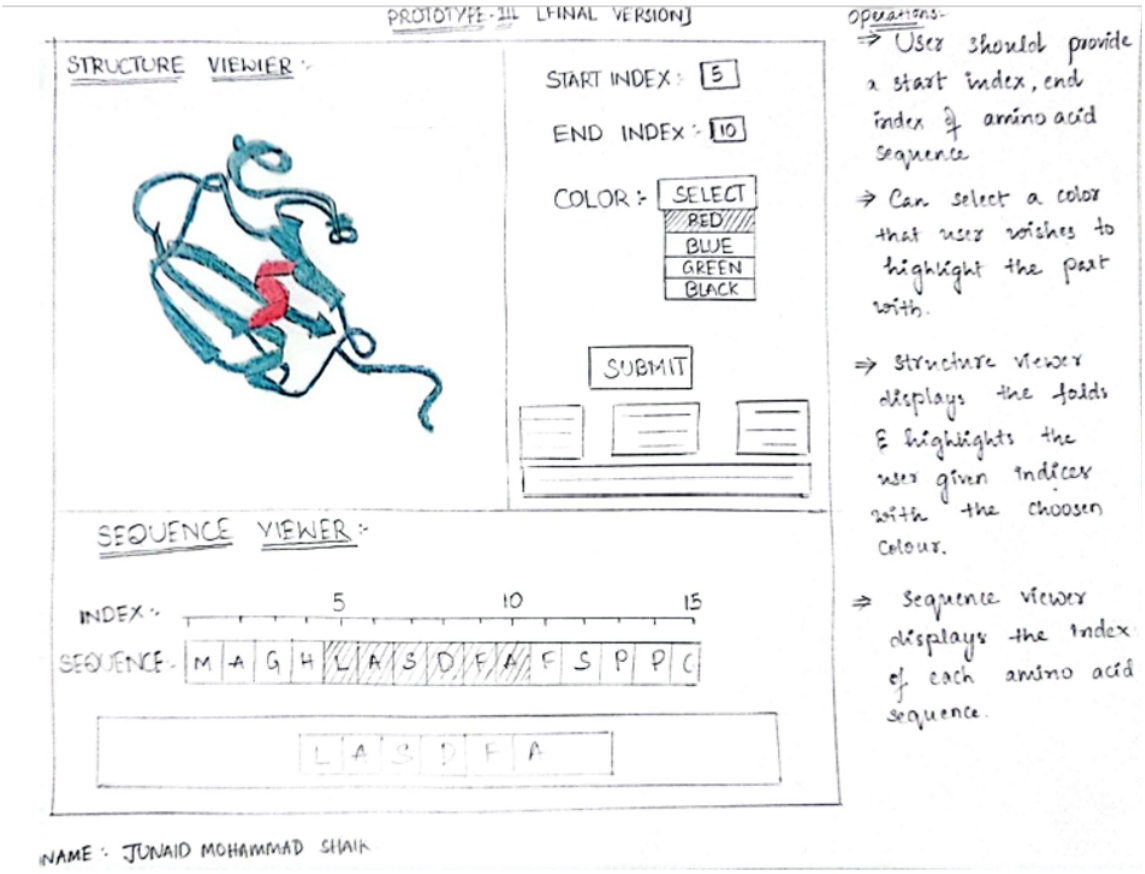
⇒ Cons:-

- If no amino acids falls under the confidence range given by user, no highlighting is done.
- There is no guarantee that every amino acid contains a confidence score.

focus:-



PROTOTYPE-3 (FINAL VERSION)



Discussion

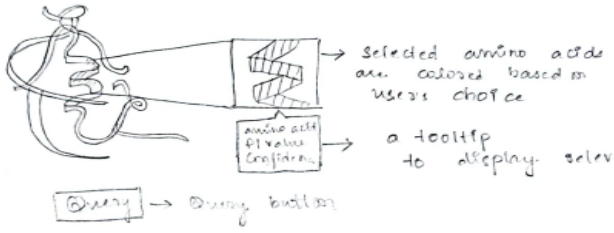
⇒ Pros:

- User selection based on index gives user the flexibility to only choose amino acid that are within the structure.
- validates the index. so, only valid indexed amino acids are selected.
- Every amino acid is indexed accordingly. helps the user to identify required amino acid based on index.
- Can also provide color choice to highlight in 3D structure

⇒ Cons:

- No major cons with respect to previous prototype. Therefore, this is the final version (approved by client)

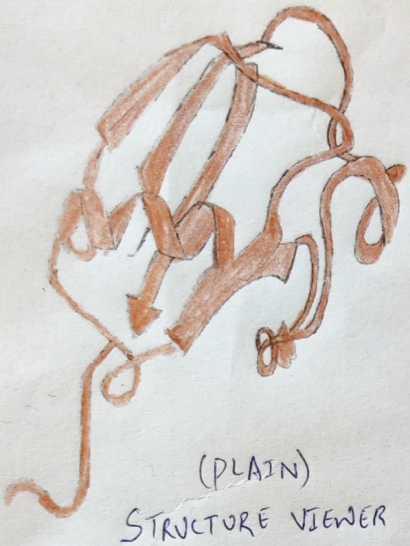
Focus:



Name: Harsh Mishra

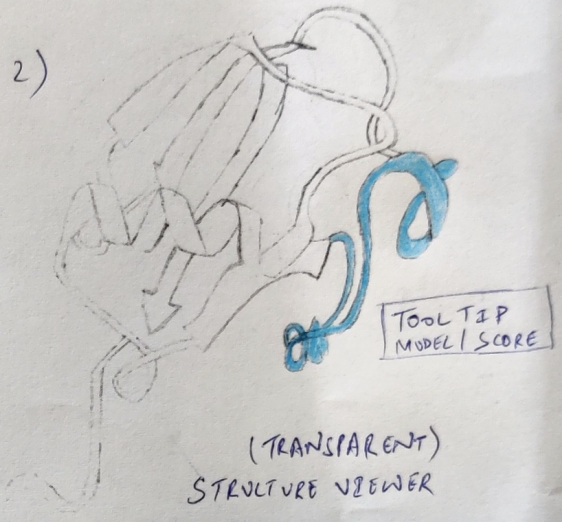
Ideas : Sheet : 1 , AUTHOR: HARSA MISHRA

1)



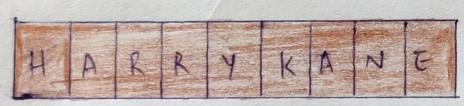
(PLAIN)
STRUCTURE VIEWER

2)

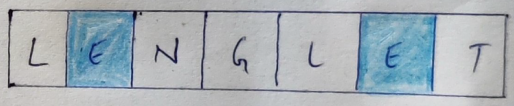


(TRANSPARENT)
STRUCTURE VIEWER

3)



4)



5-10) Give the users the ability to select amino acid sequences based on Classes (2), based on Directionality Color Index (2) and based on Input Confidence Scores (2).

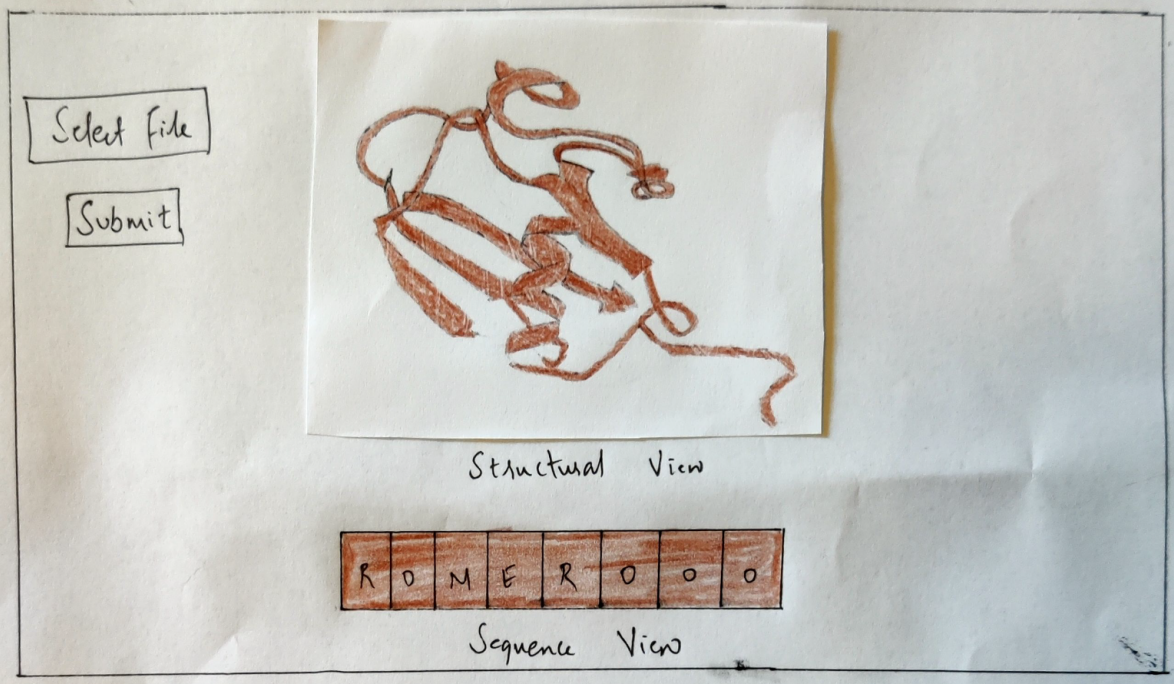
FILTERS : Every 2 pairs of options enable the users to see the amino acid sequences in two different ways, Structurally and Sequentially. Various ways can be used to enable the user to analyze the amino acid sequences in a better manner. Giving them the option to zoom in-out and get further info. about a sequence through tool tips enables them further.

TITLE : VISUALIZATION OF AMINO ACID PROTEIN SEQUENCES

AUTHOR : HARSH MISHRA

SHEET : 2

LAYOUT



Operations :

→ Input the PDB file containing the 3D locations of all the atoms of the Amino Acid Sequences and Submit.

→ Generate the structural view (plain) and the sequence view (plain) of the provided amino acid sequence

Discussions

→ Users can understand about the overall structure of the amino acid sequence.

→ The structural view and the sequence view together helps understand the sequence better.

→ Users can zoom in-out and rotate the structural view.

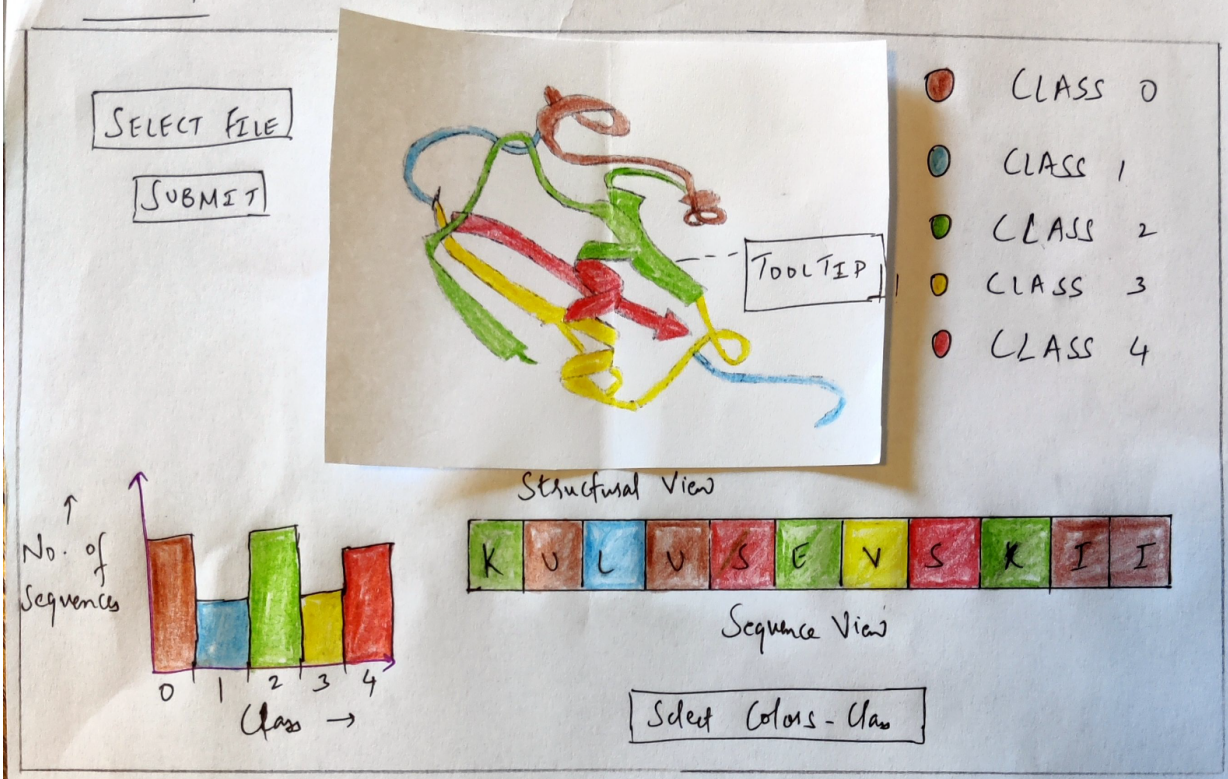
→ Getting to know about different sequences in the same protein is not possible.

TITLE - VISUALIZATION OF AMINO ACID PROTEIN SEQUENCES

AUTHOR - HARSH MISHRA

SHEET - 3

LAYOUT



Operations:

→ Input the PDB file with 3D locations of the amino acid sequence and submit.

→ Generate the structural view, sequential view and the histogram for no. of sequence per class, with default colors for classes.

→ Give the user the option to change colors based on class type of the sequence.

Discussions:

→ Users can zoom in - zoom out and rotate the sequence's structure.

→ The tooltip with additional info. as per user's hover will add extra info.

→ Distinction based on class type of the sequences will aid user's ability to know the composition of the overall protein sequence.

TITLE - VISUALIZATION of AMINO ACID PROTEIN SEQUENCES

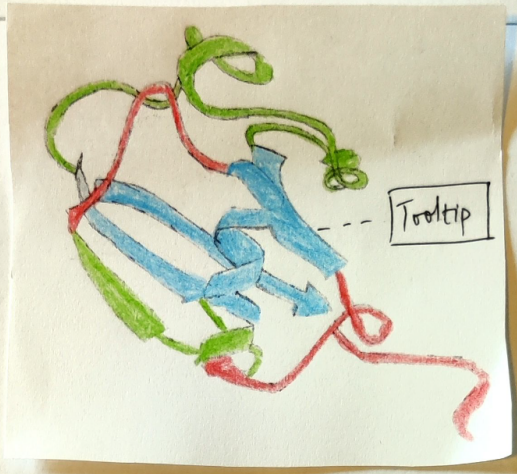
AUTHOR - HARSH MISHRA

SHEET - 4

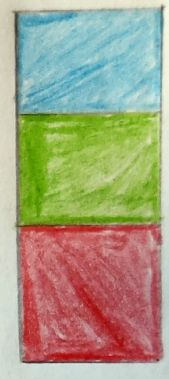
LAYOUT

SELECT FILE

SUBMIT



Sequential Color Scheme

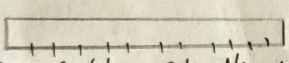


Confidence Score

Structural View

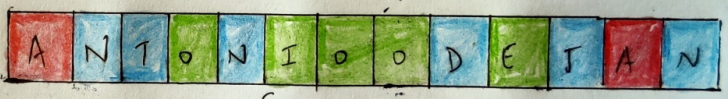
Tooltip

SELECT RANGE



0 Confidence Interval/Score 100
in %

Default: 100



Sequential View

Select Colors - Confidence Score

Operations :-

→ Input the PDB file with 3D locations of the amino acid sequence & submit.

→ Generate the structural view, sequential view with colors based on the confidence scores received from the PDB file or any other Json.

→ Users can select the range of confidence scores they want to be visualized. Others not in range will be colorless.

Discussions:

→ Users can zoom-in-out and rotate the structural view of the sequence and can hover over both the structural view and the sequential view to get extra info. from tooltip.

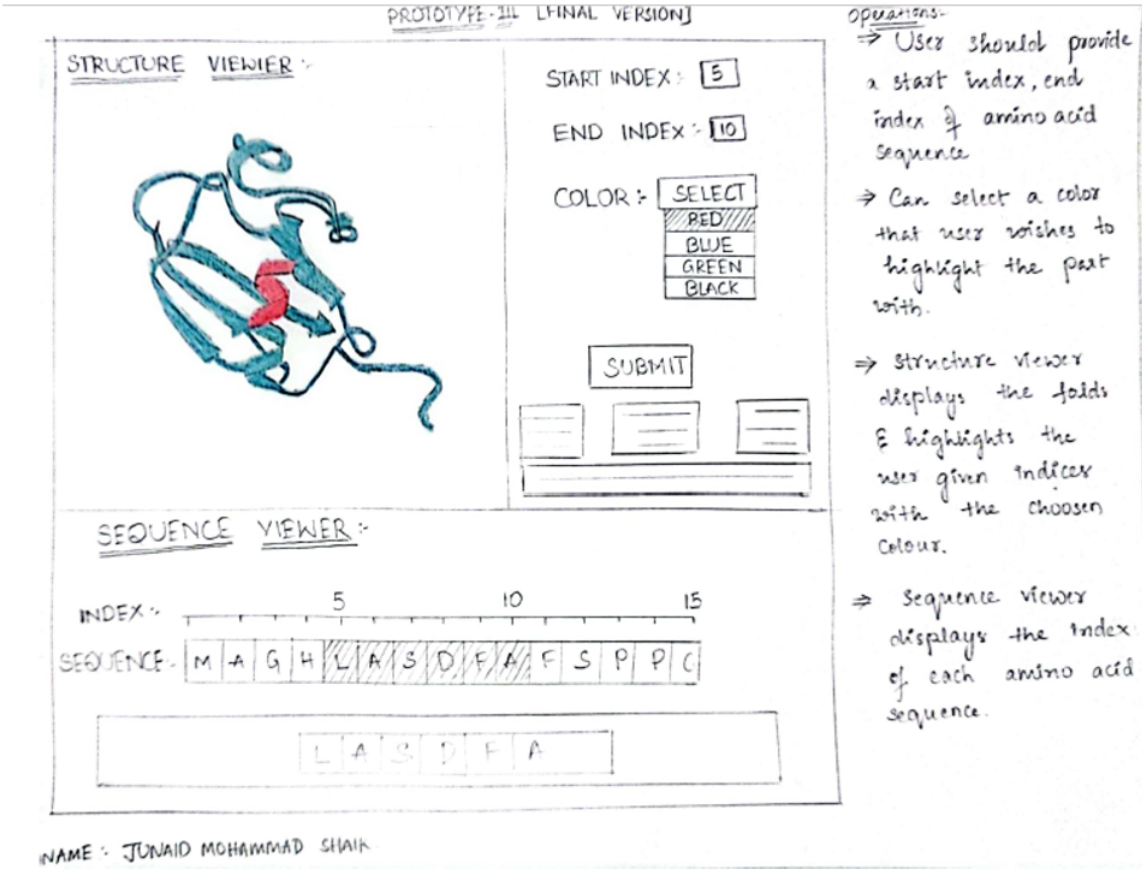
→ Users can select the range of confidence scores and can then narrow down the coloring of the visualization based on that selection.

→ They can also change the color scheme for the confidence scores.

Software placeholder:

Alpha Release:

Final Prototype



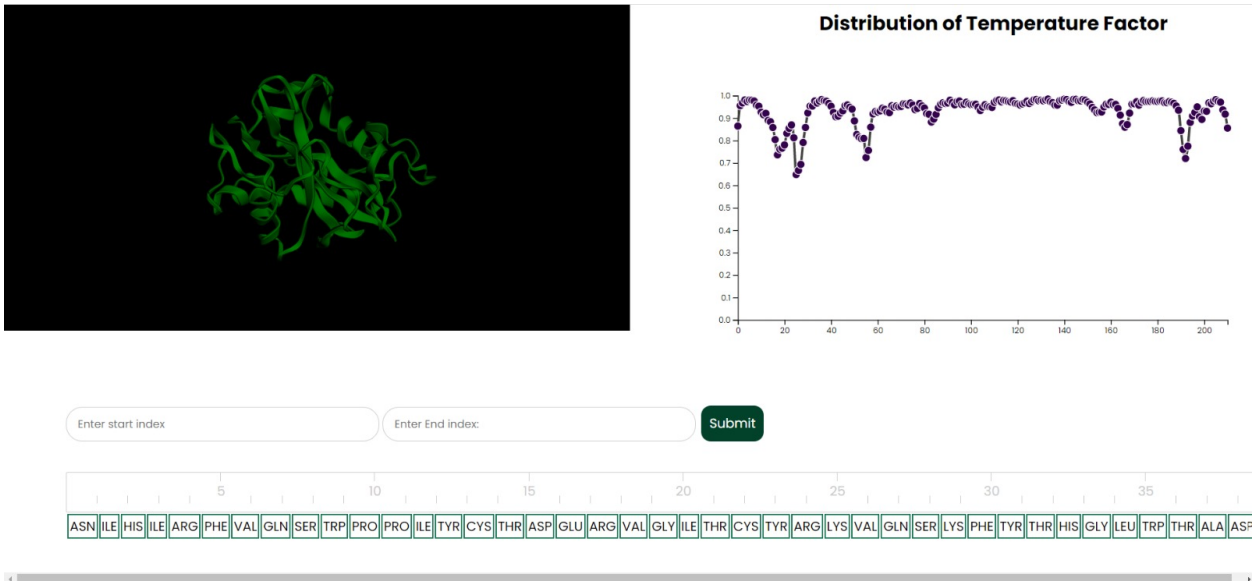
Current state of the interface -:

We have completed all our low level targets that we proposed at the beginning of the project. For alpha release we do not implement the backend side of our system, rather we assume that the protein structure to be visualized is fixed. Given the Protein Data Bank (PDB) file, we extract the X,Y, and Z coordinates of the amino acid sequences to plot the 3D structural view that is visible on the top left side in the screenshot attached below. The user has the ability to zoom in and out of the protein structure as well as the ability to rotate it. The different amino acid sequences are also shown in a sequential format, where the abbreviations for all the sequences are used to represent them. The user can choose the starting index and the end index of the sequence they want to be highlighted. Once the selection of the indices is submitted, the amino acid sequences within that index range get highlighted in white color. We also extract the temperature factor for each amino acid sequence in the given protein structure, and show the distribution of the Temperature factor on the right side of the interface, as seen below.

Overall, we represent the data from a PDB file using the 3 different visualizations, the Structural view, the Sequence view and the Distribution of Temperature Factor.

We were not able to make the sequence viewer interactable for our alpha release, as adding tooltips to the amino acid sequences would need us having more detailed data about each amino acid sequence, which was not present in the PDB file that we are using for our reference. Our alpha version visualizes amino acids from one pdb file, so future versions will include an option for the user to upload PDB file or unique PDB code, which makes our visualization dynamic.

Current screenshots -:



Tasks used for testing -:

1) Select start index and end index from the sequence viewer, to be represented in the structure view.
2) Zoom in-out on the 3D structure

3) Brushing and tooltip for each amino acid sequence

Interview questions -:
1) Was our interface self explanatory?
2) Was the chosen sequence distinctly visible post highlighting?
3) Do the colors used in the interface have the necessary contrast?
4)Did you find difficulty in figuring out the distinction between the different types of visualizations and how they differ?

Alpha Release Feedback -:

Tester 1:

	T1	**T2**	**T3**
Success/Failure	Yes	Yes	Yes
Time taken	10 seconds	15 seconds	12 seconds
Errors/Feedback	Different encodings can be used to display the structure		

Tester 2:

	T1	**T2**	**T3**
Success/Failure	Yes	Yes	Yes
Time taken	10 seconds	15 seconds	12 seconds
Errors/Feedback	Sequence viewer can be interactable		

Tester 3:

	T1	**T2**	**T3**
Success/Failure	Yes	Yes	Yes
Time taken	10 seconds	15 seconds	12 seconds
Errors/Feedback	Option to reset the visualization encoding		

Tester 4:

	T1	**T2**	**T3**
Success/Failure	Yes	Yes	Yes

Time taken	10 seconds	15 seconds	12 seconds
Errors/Feedback	N.A.		

Beta Release:

Current state of the interface -:

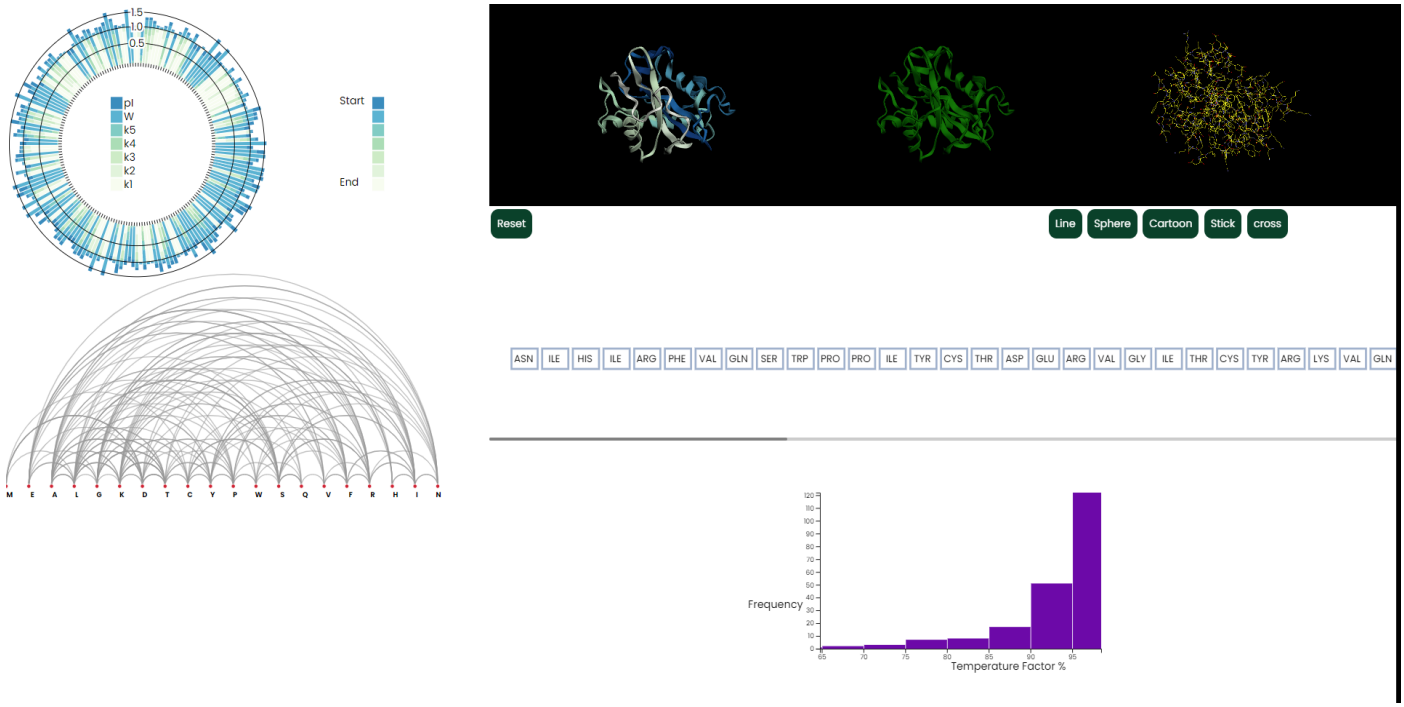
1) Enhancements from Alpha release:

- For the beta release we further enhanced the visualizations that we had for our alpha release. The sequence viewer which was initially static has been transformed into being interactive and dynamic. It can now be directly used to provide the starting and the ending index by the users to color any specific sequence of the amino acid sequence. The selected sequence gets highlighted in white color.
- The temperature factor graph has been modified to show the distribution of the temperature factors via a histogram rather than representing it via a line chart. This way we can show the actual distribution for the input file.

2) Newly added features:

- Radial graph - It represents the various characteristics of each amino acid sequence based on the details present in the input file.
- Arc Diagram - It represents the various links that each type of amino acid sequence has with the other types, the classes of the sequences are considered in our case. The user can click on each class type and know about the other classes that are directly connected to the class that was clicked on.
- 3D structure viewing options - The user can change the 3D representation of the molecule structure by choosing different options to display the connections. The user can choose between the following options: Line, Sphere, Cartoon, Stick, Cross.

Current screenshots -:



Link: https://harshm16.github.io/vds_project_beta/

Tasks used for testing -:

1) What is the temperature factor range with the highest frequency?
2) What are the different ways the user can display the 3D structure of the amino acid sequence?
3) What is the information that you can get for the “ARG” molecule?
4) How would you set the color encodings of the middle 3D structure to default (green)?

Interview questions -:
1) Was our interface self explanatory?
2) Were the buttons responsive?
3) Was the chosen sequence distinctly visible post highlighting?
4) Do the colors used in the interface have the necessary contrast?
5) Was the information displayed via tooltips clearly visible?

Beta Release Feedback -:

Tester 1:

	T1	**T2**	**T3**	**T4**
Success/Failure	Yes	Yes	Yes	Yes
Time taken	20 seconds	12 seconds	15 seconds	10 seconds
Errors/Feedback	1) Title/Headings could help understand the viz. better.			

Tester 2:

	T1	T2	T3	T4
Success/Failure	Yes	Yes	Yes	Yes
Time taken	10 seconds	15 seconds	13 seconds	18 seconds
Errors/Feedback	1) More information about each graph can be provided, for eg. Titles.			

Tester 3:

	T1	T2	T3	T4
Success/Failure	Yes	Yes	Yes	Yes
Time taken	20 seconds	10 seconds	12 seconds	13 seconds
Errors/Feedback	1) More information about each graph can be provided, for eg. Titles.			

Tester 4:

	T1	T2	T3	T4
Success/Failure	Yes	Yes	Yes	Yes
Time taken	10 seconds	12 seconds	15 seconds	20 seconds
Errors/Feedback	1) Double clicking. 2) The XY coordinate for the Tooltip needs to be fixed.			

Tester 5:

	T1	T2	T3	T4
Success/Failure	Yes	Yes	Yes	Yes
Time taken	8 seconds	10 seconds	10 seconds	14 seconds
Errors/Feedback	1) Tooltip, titles 2) Color codings for the radial chart should be in-sync 3) The start-end legend should be shown alongside the viz. that it belongs to.			

Common Feedbacks:

1) Ask the user to specifically click once for selecting the start and the end index.

2) Highlight the start and the end index of the chosen subsequence with the same color as used to encode the 3D structure.
3) Title and additional information should be provided to make the system look self explanatory.
4) The tootips should be placed better

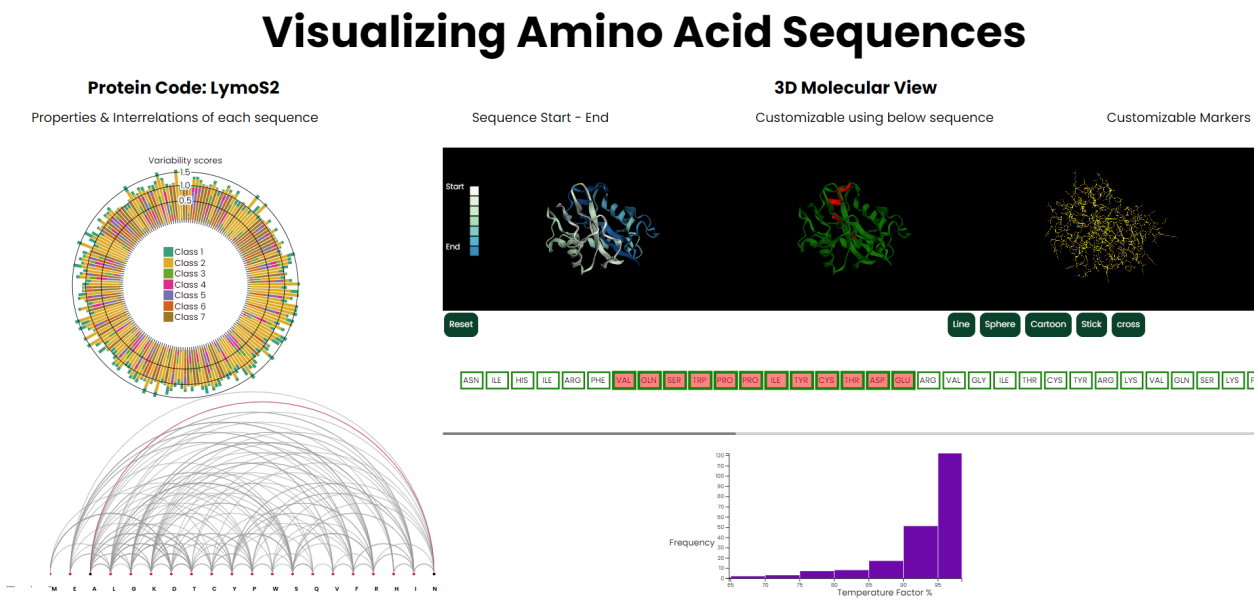
Public Release -

Current state of the interface -:

Enhancements from Beta release:

- For the Final Release, we further enhanced the visualizations that we had for our beta release. As per tester feedback, we enhanced the view of the sequence viewer when the user selected the start and end of the molecule to be highlighted. Now the start and the end index of the chosen subsequence gets highlighted with the same color as used to encode the 3D structure.
- Moved the legend of the sequence start-End more closer to the amino acid.
- Handled the tooltip of the sequence and Temperature factor graph. Previously, it was coming a little bit far away from the graphs.
- Changed the legend of Radial graph to be more descriptive about the colors of each bar. Also, provide an index of the particular bar to the tool tip of the Radial graph.
- Added titles for each sub visualization of the visualization, to make the system self explanatory.

Current screenshots -:



Feedback from Client-:

Link to the source code -: https://github.com/harshm16/VDS_project

Link to the public version -: https://harshm16.github.io/demo/

Link to the video presentation-:
https://drive.google.com/file/d/1UcHy30sbe1vupWMVIG1jUovd9jJAqv3Q/view?usp=share_link

Link to the presentation -: VDS final presentation