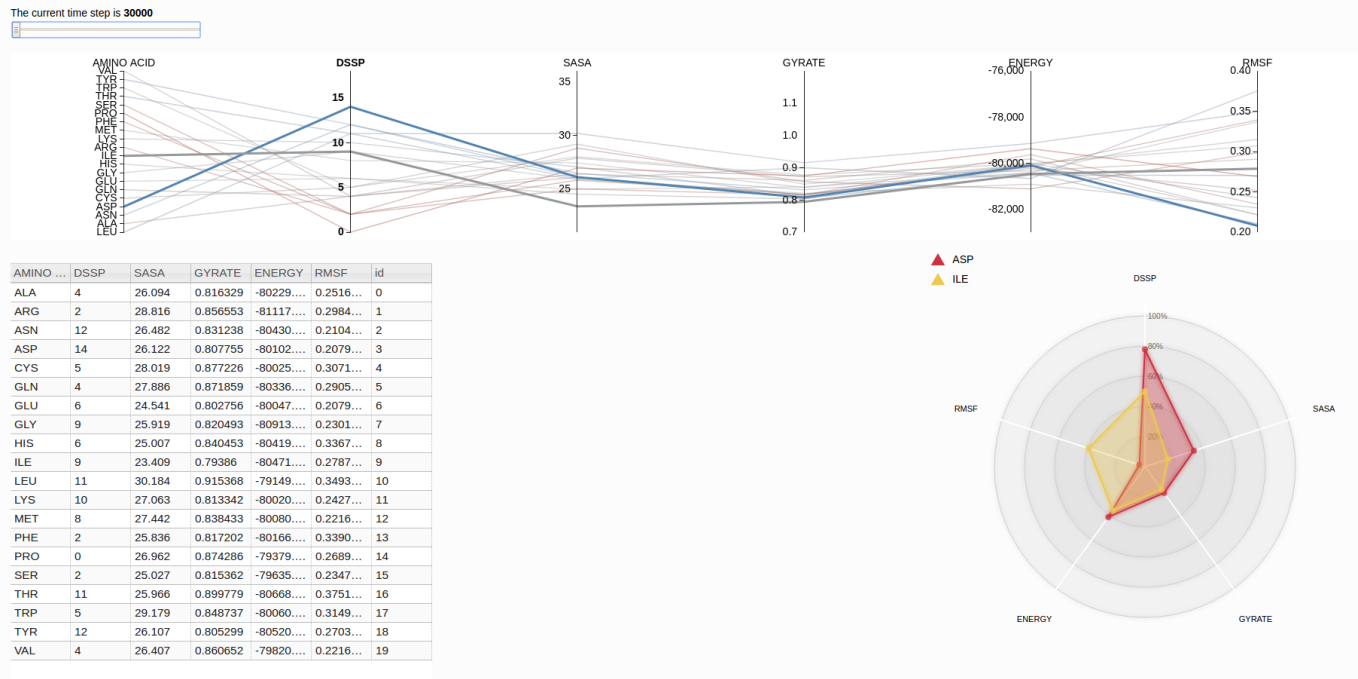
Visualization tool for comparison of a single amino acid change in protein simulation

Bruno Iochins Grisci
Instituto de Informatica
UFRGS
Porto Alegre, Brazil
Email: bigrisci@inf.ufrgs.br



AMINO …	DSSP	SASA	GYRATE	ENERGY	RMSF	id
ALA	4	26.094	0.816329	-80229.…	0.2516…	0
ARG	2	28.816	0.856553	-81117.…	0.2984…	1
ASN	12	26.482	0.831238	-80430.…	0.2104…	2
ASP	14	26.122	0.807755	-80102.…	0.2079…	3
CYS	5	28.019	0.877226	-80025.…	0.3071…	4
GLN	4	27.886	0.871859	-80336.…	0.2905…	5
GLU	6	24.541	0.802756	-80047.…	0.2079…	6
GLY	9	25.919	0.820493	-80913.…	0.2301…	7
HIS	6	25.007	0.840453	-80419.…	0.3367…	8
ILE	9	23.409	0.79386	-80471.…	0.2787…	9
LEU	11	30.184	0.915368	-79149.…	0.3493…	10
LYS	10	27.063	0.813342	-80020.…	0.2427…	11
MET	8	27.442	0.838433	-80080.…	0.2216…	12
PHE	2	25.836	0.817202	-80166.…	0.3390…	13
PRO	0	26.962	0.874286	-79379.…	0.2689…	14
SER	2	25.027	0.815362	-79635.…	0.2347…	15
THR	11	25.966	0.899779	-80668.…	0.3751…	16
TRP	5	29.179	0.848737	-80060.…	0.3149…	17
TYR	12	26.107	0.805299	-80520.…	0.2703…	18
VAL	4	26.407	0.860652	-79820.…	0.2216…	19

Fig. 1. Image of the tool with all elements presented.

***Abstract*—**

***Keywords*-molecular dynamics simulation; bioinformatics; proteins; protein folding; amino acid mutation; temporal visualization; multidimensional visualization; parallel coordinates; radar chart;**

I. Introduction

In a molecular dynamics simulation, many features can be measured, depending on the researchers interests. One of the mainstays of simulation analysis is the inspection of trajectory convergence, an indication of simulation stability [1]. It is, however, well known that convergence is a subjective matter, and that the evaluation of single structural features (such as RMSD) for this purpose is very unreliable [2] [3]. Thus, the simultaneous evaluation of multiple structural analysis via a simple visual tool, as the one presented here, is of great utility to the field of macromolecular simulation.

Proteins are polymers formed by a sequence of around 20 different possible amino acids (Fig. 2) that under physiological conditions fold into a precise shape known as its native state [4]. Each amino acid has an alpha carbon (CA) with bonds to amino (NH2) and carboxyl (COOH) groups and a variable side-chain (R) that determines the particular physicochemical properties of each residue.

A peptide is a molecule composed of two or more amino acids chained by a chemical bond called the *peptide bond*. This bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, releasing a water molecule. The interaction between amino acids in a protein causes the polypeptide chain to fold, usually in a proper configuration, as α-helix, β-sheet, coil or turn. These local folding patterns represent the secondary structure of a protein. The topology or fold is given by the succession of secondary structures connected in a 3D space. The specific characteristics of the peptide bond have significant implications for the 3D fold that can be adopted by polypeptides. The peptide bond (C-N) has a double bond, and

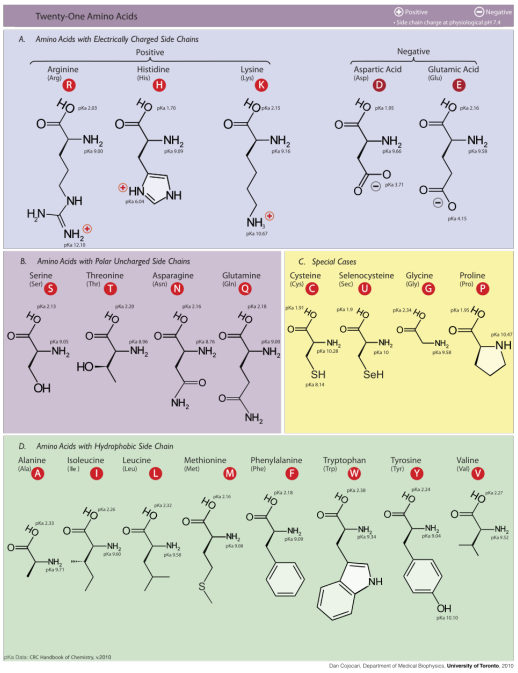


Fig. 2. Table of amino acids physicochemical properties.

it is not allowed rotation of the molecule around this bond. The rotation is only permitted around the bonds $\mathrm{N}-\mathrm{C}\alpha$ and $\mathrm{C}_\alpha-\mathrm{C}$.

The analysis of experimental protein structures (X-ray data) reveals that amino acid residues can assume many conformations in proteins [5]. Each amino acid has a set of physiochemical properties which contributes to its intrinsic conformational preference [6]. Amino acids in a secondary structure usually adopt a particular set of backbone torsion angles [5].

The amino acids sequence of a protein is directly related to its structure in three dimensional space, which is defining of the protein biological function. The change of only one amino acid in the chain is capable of great modification of the structure and function of a protein because of the differences in size and physical-chemical properties among amino acids [7].

A. Related work

II. Data characterization

The data for testing this visualization tool comes from the Kappa-conotoxin PVIIA, with amino acids sequence: $CRIPNQKCFQHLDDCCSRKCNRFNKCV$. Its three dimensional structure, obtained with nuclear magnetic resonance (NMR) from the Protein Data Bank (PDB) under the code 1AV3, can be seen in Fig. 3. Kappa-conotoxins are neurotoxical proteins extracted from sea slugs poison. They are inhibitors of the potassium channel and when injected in different organisms some alterations in their effect can be observed, from hyperactivity in fish to death when combined with other variants of conotoxins. The substitution of the asparagine (N) in the fifth position in the amino acids chain of the Kappa-conotoxin PVIIA by an alanine (A) is

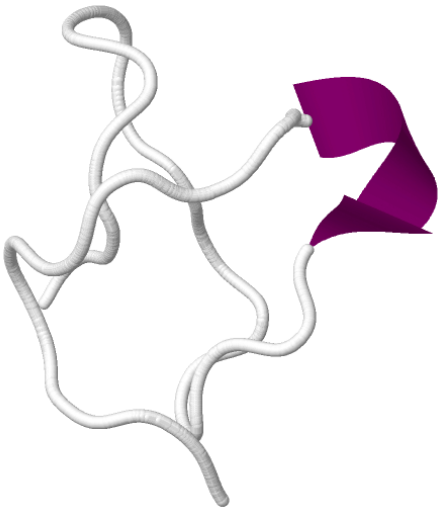


Fig. 3. Three dimensional structure of the Kappa-conotoxin PVIIA.

know to cause a reduction of 100% in the toxicity of this protein [8] [9] [10].

For this study, this asparagine was changed by all possible 20 amino acids, and for each resulting structure a molecular dynamics simulation was performed in order to evaluate its behavior during a predetermined period of time under chosen conditions. The simulations were performed with GROMACS [11], a software for biomolecules simulation, with the simulated time of 50ns. These simulations mimic the behavior of biomolecules in a solvent under controled temperature. From the data created versus time, the following attributes were studied.

A. Data dimensions

DSSP: Analysis of the secondary structure of the protein for each frame of the simulation. The secondary structure is a local structural conformation of a region in the amino acids sequence which follows specific patterns that, once folded, will originate the three dimensional functional structure of the protein [12].

SASA: Solvent-accessible surface area is the surface area of the protein that is accessible to the solvent. By analysing the area exposed to the solvent, i.e., what is around the protein, it is possible to infer how denatured is the protein, so the greater the SASA value, the greater the denaturation [13].

GYRATE: The radius of gyration can be used as a measurement of the compactness of a protein structure. It describes the overall spread of the molecule and is calculated taking the root mean square distance of the atoms from their common centre of gravity. Greater values of radius of gyration means a less compact protein [14].

ENERGY: This attribute measures the total energy of the system. The energy is a stability indicator, usually comparable, in which systems with less energy will be more stable (less entropy) [15].

AMINO ...	DSSP	SASA	GYRATE	ENERGY	RMSF	id
ALA	4	26.094	0.816329	-80229....	0.2516...	0
ARG	2	28.816	0.856553	-81117....	0.2984...	1
ASN	12	26.482	0.831238	-80430....	0.2104...	2
ASP	14	26.122	0.807755	-80102....	0.2079...	3
CYS	5	28.019	0.877226	-80025....	0.3071...	4
GLN	4	27.886	0.871859	-80336....	0.2905...	5
GLU	6	24.541	0.802756	-80047....	0.2079...	6
GLY	9	25.919	0.820493	-80913....	0.2301...	7
HIS	6	25.007	0.840453	-80419....	0.3367...	8
ILE	9	23.409	0.79386	-80471....	0.2787...	9
LEU	11	30.184	0.915368	-79149....	0.3493...	10
LYS	10	27.063	0.813342	-80020....	0.2427...	11
MET	8	27.442	0.838433	-80080....	0.2216...	12
PHE	2	25.836	0.817202	-80166....	0.3390...	13
PRO	0	26.962	0.874286	-79379....	0.2689...	14
SER	2	25.027	0.815362	-79635....	0.2347...	15
THR	11	25.966	0.899779	-80668....	0.3751...	16
TRP	5	29.179	0.848737	-80060....	0.3149...	17
TYR	12	26.107	0.805299	-80520....	0.2703...	18
VAL	4	26.407	0.860652	-79820....	0.2216...	19

Fig. 4. Overview of the used grid.

RMSF: The root mean square fluctuation of atomic positions is the measure of the average distance between the atoms of the simulated protein and a well-defined average position [16].

All the dimensions are continuous numerical values. As can be seen, the data of a single simulation is multidimensional and changes over the simulated time.

III. Technical Background

A. Grid

B. Parallel Coordinates

C. Radar Chart

IV. Technique Overview

The proposed visualization technique takes advantage of the use of three different visualization tools working in synchrony, and the capability of changing the data over time. The parallel coordinates and radar chart data is normalized in the range of minimum and maximum values for each dimension along all time steps.

A. Grid

The grid is an interactive table where the raw values of the data are displayed as in Fig. 4. Its function is to provide a place where the data can be observed with more detail for scientific purposes.

Item highlighting: When hoving the mouse over a line in the grid, that line is highlited. This operation also reflects on the other charts.

Item selection: When right-clicking on a line of the grid with the mouse that line is selected and the item is fixed on the charts. The selected line is becomes darker on the parallel coordinates chart.

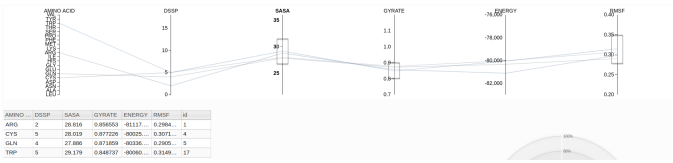


Fig. 5. Example of use of the brush.

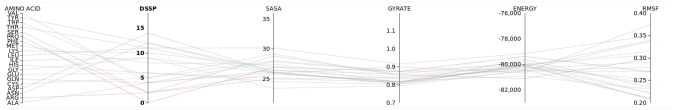


Fig. 6. Colors of the parallel coordinates for the DSSP dimension.

B. Parallel Coordinates

The parallel coordinates chart is used to show a global view of the data, displaying all the items in a given time at once and allowing the user to view the general ”shape” of the simulations and spread of the values between items.

Axis brushing: The brush is the filtering operation of the parallel coordinates and can be seen in Fig. 5. Draging the mouse along an axis select the items in that range. This change is reflected on the grid.

Axis sorting: When double clicking an axis the order of its values is inverted. This change is reflected on the radar chart.

Axis reordering: When right-clicking an axis and draging it to left or right the axes are reordered. This change is reflected on the radar chart.

Recoloring by axis selection: The color of the lines of the parallel coordinates is defined by the range of the selected axis. Clicking an axis change the dimension used for coloring. Figures 6 and 7 show the colors for two different dimensions. This function is useful when analysing one dimension while wanting to keep in mind the behavior of another.

C. Radar Chart

The objective of the radar chart, showed in Fig. 8, is to visualize and compare specif items of the data, i.e., single simulations. The hypothesis is that overlaping the shapes of the lines from the parallel coordinates in a radial way makes it easier com see the similarities and differences between a small number of items, while the global view is better represented by the parallel coordinates chart since for many shapes the radar chart becomes convoluted.

Item highlighting: When hoving the mouse over a shape in the radar chart or its label, the shape is filled in order to be higlighted.

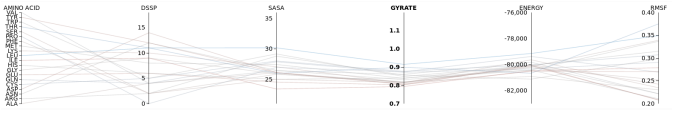


Fig. 7. Colors of the parallel coordinates for the GYRATE dimension.

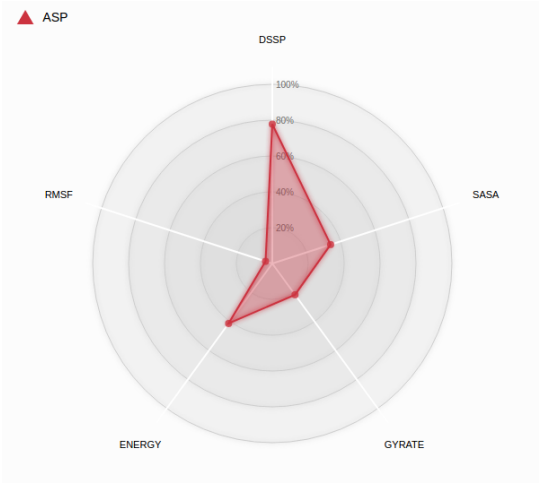


Fig. 8. Example of one simulation being visualized in the radar chart.

Fig. 9. View of the elements of the slider, that enables the change of time steps.

D. Slider

The slider is the element that allows the user to change the data being visualized over time (Fig 9). It can be clicked or dragged, then selecting a specific time step of the simulation, and the charts and grid will be updated. It is also possible to visualize the changes in time with animations. For this there are two buttons, "Play" and "Rec", that update the data automatically untill the end of the time steps. Interacting with any other element of the visualization pauses the animation. A text box informs the current time step.

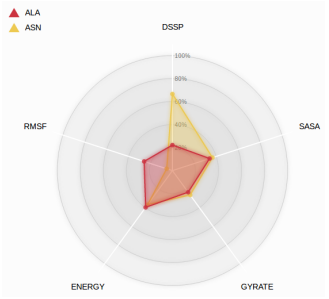
V. Experiments and Discussion

All elements of this visualization tool were implemented using JavaScript, HTML and CSS, with the visualization library D3.js. The parallel coordinates chart uses the visual toolkit for multidimensional detectives Parallel Coordinates (0.7.0) (http://syntagmatic.github.io/parallel-coordinates).

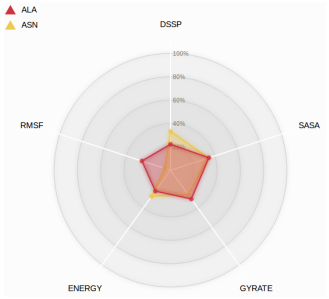
The radar chart was build from the implementations of Nadieh Bremer (http://visualcinnamon.com/2015/10/different-look-d3-radar-chart.html) and Micah Stubbs (http://bl.ocks.org/micahstubbs/a772306d6fd49874ec92).

As previously mentioned, the Kappa-conotoxin PVIIA was chosen as example for testing the visualization tool. We know from the start that the simulation containing the asparagine (N) in the fifth position of its amino acid chain represents the protein in its natural state. Also, the change of this asparagine by an alanine is reported to reduce 100% of the proteins toxicity. When comparing the simulations with asparagine and alanine, thus, it would be expected to see some similarity between then, since the mutation don't destroy the protein, but also enough differences to justify the loss of function.

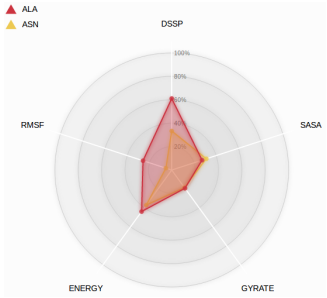
This comparison is showed in Fig. 10 for three different moments in the simulations. The greatest discrepancy observed are the DSSP and RMSF attributes, what could indicate the 3D structure of the proteins have significant differences.



(a) Time step 30000



(b) Time step 40000



(c) Time step 50000

Fig. 10. Comparison between the simulations with asparagine (yellow) and alanine (red) for at three different moments.

In another test, we can compare the simulation with asparagine to the ones with glycine and proline, as in Fig. 11. The amino acids glycine and proline, due to their characteristics, usually interfere greatly in the structure of a protein when placed instead of another.

Finally, it would be of interest to identify possible mutation candidates that would interfere the least with the original protein structure and function. Using visual inspection (Fig. 12), the amino acids aspartic acid and methionine, based only on the simulation attributes, seem to be potential options. Of course, any further conclusion is depended of biological analysis, with the visualization tool serving only as a first step for insights.

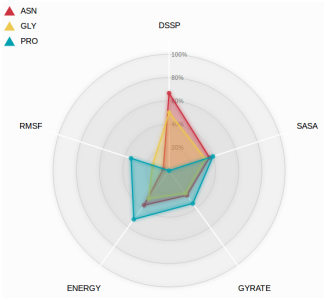
VI. Conclusion

A demo can be tested at: http://inf.ufrgs.br/~bigrisci/parallel-coordinates

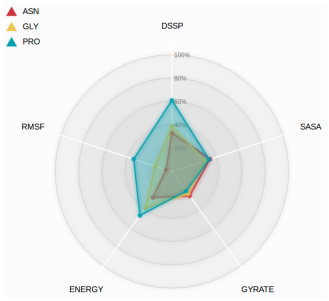
A. Future work

This tool, although functional, is open for further improvements. It would be desired to expand the synchrony between different charts and the addition of new charts that could provide new insights about the data.

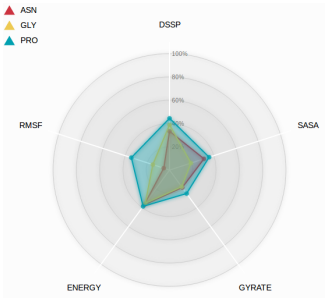
A future extension of this project is to make it an available webtool allowing researchers to visualize their molecular dynamics simulation data. The user should be able to upload the



(a) Time step 30000

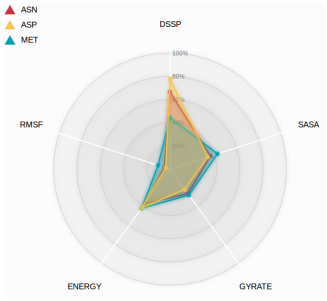


(b) Time step 40000

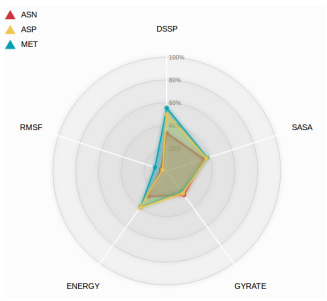


(c) Time step 50000

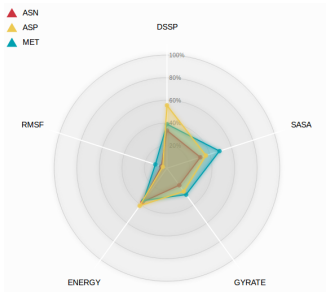
Fig. 11. Comparison between the simulations with asparagine (red), glycine (yellow) and proline (blue) for at three different moments.



(a) Time step 30000



(b) Time step 40000



(c) Time step 50000

Fig. 12. Comparison between the simulations with asparagine (red), aspartic acid (yellow) and methionine (blue) for at three different moments.

output file generated from GROMACS and select the desired attributes. We hope this could help to better understandment of the simulations and comparisons between them.

Another possible use for this tool, outside the field of molecular dynamics simulation, would be the visualization of convergence in population based optimization algorithms for multidimensional problems. This data seems to fit well within the techniques used in this project.

Acknowledgment

I would like to thank the structural Bioinformatics and Computational Biology Lab (INF-UFRGS) for the great help generating and analysing the data for this project and testing the visualization tool. Special thanks to Dr. Marcio Dorn, Dr. Rodrigo Ligabue-Braun and Leonardo Alves.

References

[1] A. Grossfield, S. E. Feller, and M. C. Pitman, "Convergence of molecular dynamics simulations of membrane proteins," *Proteins: structure, function, and bioinformatics*, vol. 67, no. 1, pp. 31–40, 2007.

[2] B. Knapp, S. Frantal, M. Cibena, W. Schreiner, and P. Bauer, "Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible?" *Journal of Computational Biology*, vol. 18, no. 8, pp. 997–1005, 2011.

[3] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger *et al.*, "Biomolecular modeling: goals, problems, perspectives," *Angewandte Chemie International Edition*, vol. 45, no. 25, pp. 4064–4092, 2006.

[4] C. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 96, pp. 223–230, 1973.

[5] B. Borguesan, M. Barbachan e Silva, B. Grisci, M. Inostroza-Ponta, and M. Dorn, "Apl: an angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction," *Comput. Biol. Chem.*, vol. 59, no. A, pp. 142–157, 2015.

[6] V. Mathura and D. Kolippakkam, "Apdbase: Amino acid physicochemical properties database," *Bioinformation*, vol. 1, no. 1, pp. 2–4, 2005.

[7] D. L. Nelson, A. L. Lehninger, and M. M. Cox, *Lehninger principles of biochemistry*. Macmillan, 2008.

[8] R. B. Jacobsen, E. D. Koch, B. Lange-Malecki, M. Stocker, J. Verhey, R. M. Van Wagoner, A. Vyazovkina, B. M. Olivera, and H. Terlau, "Single amino acid substitutions in κ-conotoxin pviia disrupt interaction with the shaker k+ channel," *Journal of Biological Chemistry*, vol. 275, no. 32, pp. 24 639–24 644, 2000.

[9] R. Mir, S. Karim, M. Amjad Kamal, C. M Wilson, and Z. Mirza, "Conotoxins: structure, therapeutic potential and pharmacological applications," *Current pharmaceutical design*, vol. 22, no. 5, pp. 582–589, 2016.

[10] D. T. Akey, X. Zhu, M. Dyer, A. Li, A. Sorensen, S. Blackshaw, T. Fukuda-Kamitani, S. P. Daiger, C. M. Craft, T. Kamitani *et al.*, "The inherited blindness associated protein aipl1 interacts with the cell cycle regulator protein nub1," *Human molecular genetics*, vol. 11, no. 22, pp. 2723–2733, 2002.

[11] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of chemical theory and computation*, vol. 4, no. 3, pp. 435–447, 2008.

[12] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[13] T. J. Richmond, "Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect," *Journal of molecular biology*, vol. 178, no. 1, pp. 63–89, 1984.

[14] M. Y. Lobanov, N. Bogatyreva, and O. Galzitskaya, "Radius of gyration as an indicator of protein structure compactness," *Molecular Biology*, vol. 42, no. 4, pp. 623–628, 2008.

[15] W. F. van Gunsteren, X. Daura, and A. E. Mark, "Gromos force field," *Encyclopedia of computational chemistry*, 1998.

[16] V. N. Maiorov and G. M. Crippen, "Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins," *Journal of molecular biology*, vol. 235, no. 2, pp. 625–634, 1994.