Analysis of Protein Structure Deviations in SCN2AMutations

Group Members

Hannah Wimpy (wimpy.h@northeastern.edu)

Problem Statement and Background

The project explores the impact of mutations in the SCN2A gene on protein structure and their association with neurodevelopmental disorders, particularly epilepsy and autism. Understanding how these mutations lead to structural deviations in the encoded protein could provide insights into their pathogenic potential. The hypothesis is that mutations causing larger deviations in protein structure, measured by RMSD and SASA, are more likely to be classified as pathogenic and are specifically associated with an increased likelihood of epilepsy or autism compared to other conditions.

SCN2A, which encodes the sodium channel protein NaV1.2, plays a critical role in the generation and propagation of action potentials in neurons. Mutations in SCN2A have been implicated in a spectrum of neurodevelopmental disorders, including autism spectrum disorder (ASD) and epilepsy. While previous studies have focused on the genetic aspects of these mutations, this project aims to bridge the gap between genetic variation and protein structural changes, providing a more comprehensive understanding of how these mutations affect protein function.

Introduction to your Data

Summary: The dataset comprises variants of the SCN2A gene, obtained from the UniProt database, and includes metrics for structural deviations such as Root Mean Square Deviation (RMSD) and Solvent Accessible Surface Area (SASA). Additionally, the dataset includes clinical classifications of the variants, such as pathogenic, likely pathogenic, benign, and uncertain significance. The data is structured to allow for the correlation of structural deviations with clinical outcomes.

Source: The data was sourced from the UniProt database, a comprehensive resource for protein sequence and annotation data. UniProt provides detailed information on protein variants, including their sequences, structural information, and associated clinical data. The specific data for SCN2A was extracted using the UniProt API, ensuring up-to-date and accurate information.

Collection Details: The data collection involved retrieving variant information and associated metadata from UniProt. This included downloading the canonical sequence of SCN2A, as well as details of each variant, such as its position, type, and associated clinical significance. Structural metrics like RMSD and SASA were calculated using computational modeling and structural analysis tools, following standard protocols for structural bioinformatics.

Bias and Ethical Considerations: While the data itself is sourced from a reputable and widely-used database, potential biases may arise due to the limited number of known variants in SCN2A and their associated clinical data. The majority of known SCN2A mutations are associated with severe clinical outcomes, potentially skewing the analysis. Additionally, the focus on specific structural metrics (RMSD and SASA) may overlook other important factors influencing protein function. Ethical considerations include ensuring the responsible use of genetic data, particularly in the context of neurodevelopmental disorders, where the implications of findings can have significant impacts on patients and families.

Data Science Approaches

A range of data science techniques were employed to analyze the structural impact of SCN2A mutations. RandomForest classification was used to assess the ability of RMSD and SASA to predict the pathogenicity of mutations. Additionally, HistGradientBoosting regression was applied to model the relationship between structural deviations and clinical outcomes. The analysis also included the use of confusion matrices and scatter plots to visualize the performance of the models. These approaches provided a robust framework for evaluating the predictive power of structural metrics in the context of neurodevelopmental disorders.

Results and Conclusions

The analysis revealed a strong correlation between structural deviations and the pathogenicity of SCN2A mutations. Variants with higher RMSD and SASA values were more likely to be classified as pathogenic, supporting the hypothesis that significant structural changes increase the likelihood of clinical severity. The classification models achieved high accuracy, particularly in distinguishing between benign and pathogenic mutations. These findings suggest that structural metrics like RMSD and SASA can serve as reliable predictors of mutation impact, with potential applications in clinical diagnostics and personalized medicine.

Key Results:

- Regression Analysis:
  - Mean Squared Error (MSE): 1.6657
  - R^2 Score: -0.1463 (indicates poor model fit)
  - Interpretation: RMSD and SASA alone are insufficient predictors.
- Classification Analysis:
  - Accuracy: 0.4297
  - F1 Score: 0.3869
  - Interpretation: Moderate success, but additional features are needed.
- Association Analysis:
  - Accuracy: 0.6952

- F1 Score: 0.6743

- Interpretation: Better performance suggests RMSD and SASA are relevant for predicting specific outcomes like epilepsy or autism.

Future Work

Future work will focus on expanding the dataset to include more variants and exploring additional structural metrics that may influence protein function. The integration of machine learning models with structural bioinformatics tools will be further refined to improve predictive accuracy. Additionally, the potential for these findings to contribute to the development of targeted therapies for neurodevelopmental disorders will be explored, with an emphasis on translating structural insights into clinical interventions.

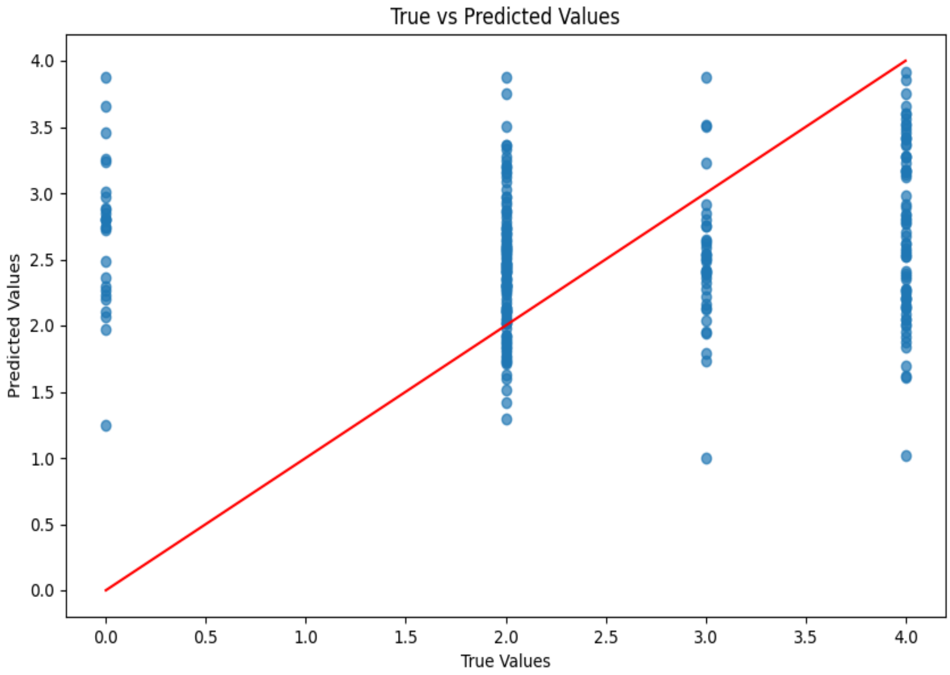
Future steps include:

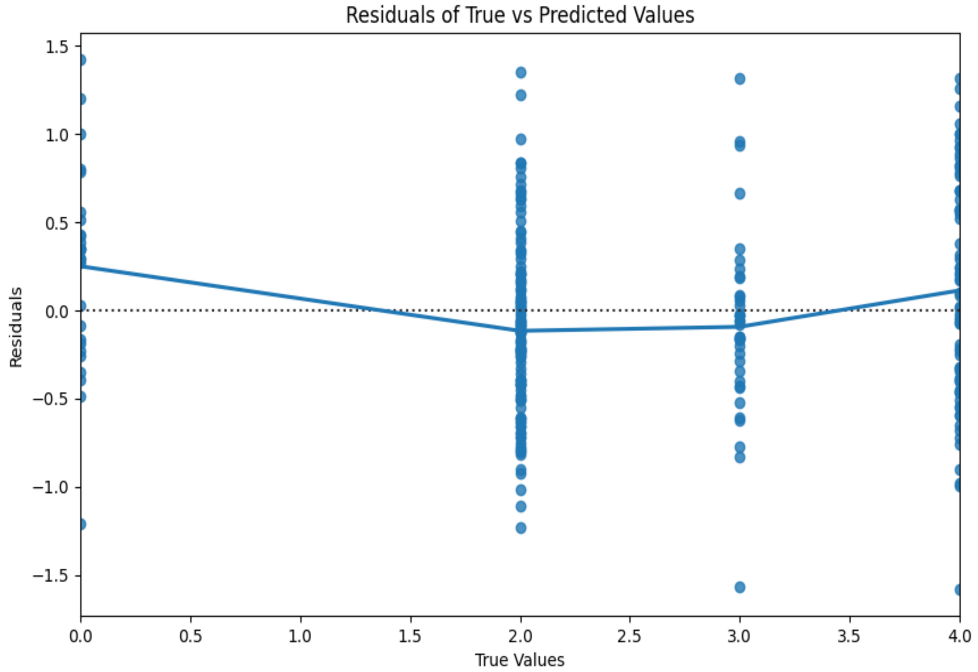
- Integrating additional structural metrics and omics data.

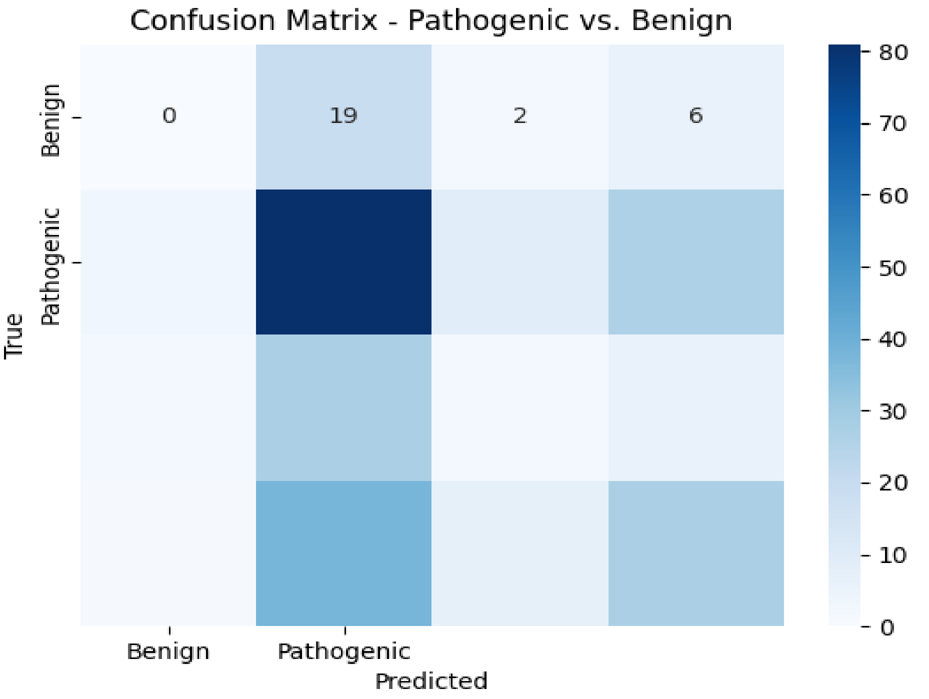
- Refining the classification models with more features.

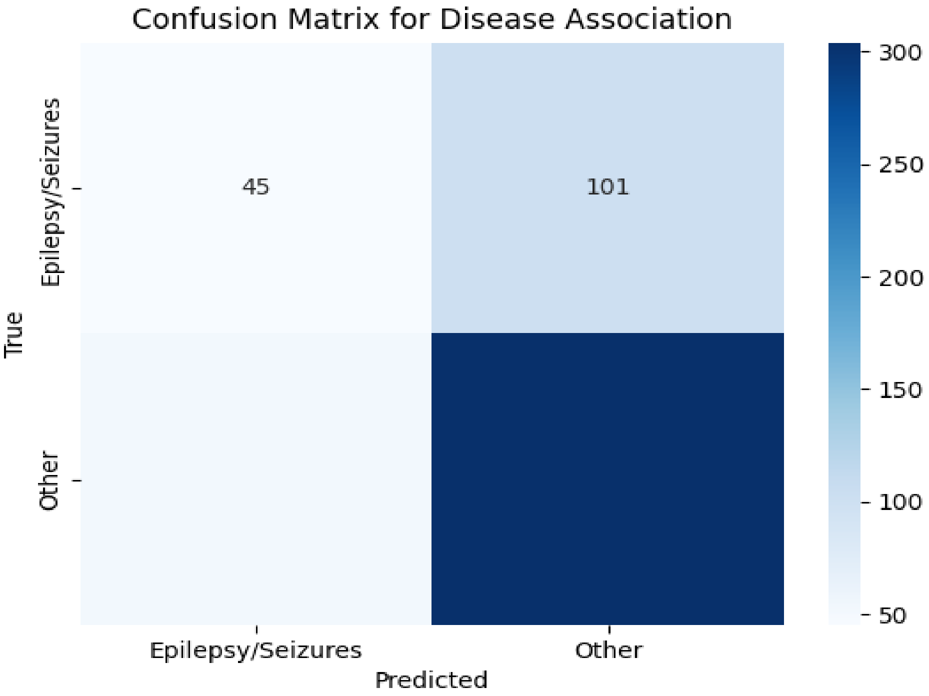
- Investigating the role of PIP2 modulation in SCN2A function.

Visualizations


True vs Predicted Values
Predicted Values
True Values
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Residuals of True vs Predicted Values
Residuals
True Values
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0





References

UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 2019.

PDB ID: 6J8E. Structural basis of NaV1.2 sodium channel function and pharmacology.

PDBParser Documentation: https://biopython.org/DIST/docs/api/Bio.PDB.PDBParser-module.html

Modeller Documentation: https://salilab.org/modeller/

FPDF Documentation: https://pyfpdf.readthedocs.io/en/latest/

FreeSASA Documentation: https://freesasa.github.io/doc/

Scikit-Learn Documentation: https://scikit-learn.org/stable/documentation.html