Final Project Report
CS612 – Algorithms in Bioinformatics

Title: Rational Design of Non-Harmful Hemoglobin Mutations for Enhanced Hydroxyurea Binding in Sickle Cell Anemia

Abstract

This project explores a new idea in the treatment of Sickle Cell Anemia: modifying the hemoglobin protein with carefully chosen mutations that do not cause disease, but instead help hydroxyurea (a known drug) bind better. I focused on designing mutations that are safe and located far from the E6V mutation responsible for sickle cell disease. Using a combination of molecular docking (AutoDock Vina) and machine learning (Logistic Regression and SVM), I evaluated 50 such mutations for their ability to create new drug binding sites. The final model predicts whether hydroxyurea can bind near the mutation site based on features like binding affinity, structural distance, and changes in charge or hydrophobicity.

1. Introduction

Sickle Cell Anemia is caused by a single point mutation in the HBB gene (E6V), which changes the shape and function of hemoglobin, leading to severe clinical symptoms. Hydroxyurea is one of the drugs commonly used to manage the disease, but it doesn't bind directly at the site of mutation and its effectiveness varies. The central idea of this project is: what if we could create new, artificial binding sites for hydroxyurea by introducing additional mutations that do NOT interfere with hemoglobin's function?

Instead of targeting the problematic site (E6V), I tested other sites across the hemoglobin protein, introducing mutations that were selected for safety and surface accessibility. These were evaluated using docking simulations, followed by machine learning to predict whether each mutation allowed the drug to bind at the mutation site.

2. Methodology

2.1 Protein Preparation and Docking

- Selected 50 mutation candidates based on literature, structural accessibility, and chemical diversity.
- Each mutant protein was modeled and converted into .pdbqt format using AutoDockTools.
- Hydroxyurea (ligand) was also prepared using the same pipeline.

- Docking simulations were run using AutoDock Vina for each protein-ligand pair.
- Output data included:
  - Binding Affinity (kcal/mol)
  - RMSD (Å)
  - Minimum Distance from ligand to mutation residue

2.2 Dataset Construction
A custom dataset was created manually from docking results.
 Features included:
- Mutation name, location, original and substituted amino acids
- Change in hydrophobicity and charge
- Type of amino acid shift (polar to hydrophobic, etc.)
- Docking output values (affinity, RMSD, distance)
- Label: Whether the drug bound at the mutation site (Yes/No)

2.3 Machine Learning Pipeline
- Model Types: Logistic Regression and Support Vector Machine
- The dataset was balanced using resampling
- 70/30 Train-Test split was used with cross-validation
- Evaluation metrics: Accuracy, ROC-AUC, Precision, Recall, Confusion Matrix
- Feature importance was assessed using permutation methods

3. Results

Out of 50 mutations:
- 18 mutations enabled hydroxyurea to bind at or near the mutation site
- Top examples: HBB_G25D, HBB_D79G, HBB_P30A, HBB_A136P, HBB_L105P
- These mutations had favorable binding affinities (around -3.8 to -4.0 kcal/mol) and very short distances between the ligand and mutation

ML Model Performance
- Logistic Regression achieved ROC-AUC ~0.86
- SVM reached ROC-AUC ~0.84
- Precision and recall were both above 80%
- Misclassifications were analyzed using Venn diagrams to compare models
- Important features:
  - MinDistance(Å)
  - Hydropathy_Change
  - Is_Charged_Change

4. Harmfulness and Biological Safety

I manually reviewed the structural location of all 18 “Yes” mutations:
- Only 1 mutation (HBB_L105P) could be potentially harmful due to its location in a structured alpha-helix region
- All other mutations were located on solvent-exposed loops or outer surface regions, suggesting minimal interference with normal hemoglobin function

This validates the idea that drug-accessible mutation sites can be designed safely.

5. Conclusion

This project demonstrates a new approach to improving drug interaction using protein engineering. By creating artificial, non-harmful mutations in hemoglobin, I were able to simulate hydroxyurea binding at new positions. The combination of structure-based docking and machine learning allowed us to:
- Discover mutation-induced binding hotspots
- Build a classifier to predict favorable mutations
- Highlight a safe and effective design space for future therapeutic experiments

If validated experimentally, this strategy could open new directions for treating sickle cell anemia or enhancing the action of existing drugs.

6. References

1. Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking.
2. UniProt Database – HBB protein sequence
3. Scikit-learn ML Library (v1.2.2)
4. Biopython – PDBParser
5. PyMOL molecular viewer

Appendix

- All docking files and ligand files are in the proteins_cleaned/ and ligand/ folders.
- The machine learning scripts are provided in the ML Coding/ folder.
- The dataset used (CS612 Dataset.csv) was fully created from the docking output and manual annotations.
- Images and docking simulation frames are included for selected mutations in the submission folder.