CSE 5370: Bioinformatics Homework 1 Write-up

Prit Desai (UTA ID: 1002170533)

March 7, 2025

1 Genome-Wide Association Study (GWAS)

1.1 Fisher's Exact Test and Results

For each SNP, I performed Fisher's Exact Test using the contingency table:

$$\begin{bmatrix} Case_C & Case_T \\ Control_C & Control_T \end{bmatrix}$$

I used the two-sided alternative hypothesis to determine statistical significance. The results were stored in `results.csv` with the following columns:

- **SNP**: The SNP identifier.
- **P-value**: The p-value from Fisher's Exact Test.
- **Significant**: Boolean flag for SNPs where p ¡ 5×10^{-8}.
- **Bonferroni-Significant**: Boolean flag for SNPs where p ¡ ($5 \times 10^{-8}/1000$).

1.2 Manhattan Plot

A Manhattan plot was generated with `-log10(P-value)` plotted against SNP index. The red dashed line represents the 5×10^{-8} threshold, and the green dashed line represents the Bonferroni-corrected threshold. The plot was saved as `manhattan_plot.png`.

2 Sequence Alignment

2.1 Custom Substitution Matrix

A substitution matrix was created for the 26 letters of the English alphabet and space:

- Match: +2
- Semi-match (characters present in "pritdesai"): +1
- Mismatch: -1

This matrix was stored in `1002170533_S.txt`.

2.2 Local Alignment: Smith-Waterman Algorithm

The Smith-Waterman algorithm was used for local alignment of "pritdesai" and the pangram "the quick brown fox jumps over the lazy dog." The final alignment matrix was stored in `1002170533_D.txt`.

2.3 Global Alignment: Needleman-Wunsch Algorithm

The Needleman-Wunsch algorithm was implemented to compute the global alignment between "pritdesai" and the pangram. The resulting matrix was saved in `1002170533_NW.txt`.

3 Conclusion

All required statistical tests and alignments were implemented successfully, with results formatted correctly for grading.

4 Difficulty Adjustment

- **How long did this assignment take you to complete?**
  This assignment took approximately 8 hours to complete.
- **If the assignment took you longer than 10 hours, which parts were overly difficult?**
  The assignment was completed within 8 hours, but the most challenging part was implementing the Fisher's Exact Test correctly and ensuring the proper formatting of the Manhattan plot.