A Mismatch Relaxation to the Primer Selection Process of an Amplicon Sequencing Algorithm

Dean Polimac - d.polimac@student.tudelft.nl

Responsible Professor: Jasmijn Baaijens
Supervisor: Jasper van Bemmelen

INTRODUCTION

This research focuses on the impact of mismatches on the selected amplicons in the AmpliDiff algorithm (cite). The AmpliDiff algorithm finds highly discriminatory parts of genomes, along with primers which can be used to amplify those regions.

RESEARCH QUESTION

The main research questions are:

- Does allowing mismatches impact the amplicons selected?
- If so, how is it reflected on amplicons which were selected in both solution sets?

METHODOLOGY

- First a primer-to primer comparison is done
- Each primer gets assigned a set of similar primers, if similarity ≥ **e**
- Similarity between primers is computed based on two different metrics
  - Hamming Distance [1]
  - Levenshtein Distance [2]
- A distance of region enclosed constraint is added which ensures that the same primer pair cannot occur twice, as shown in Figure 1.

Figure 1: Difference in regions enclosed by the same primer pair.

EXPERIMENT SETUP

- Two datasets comprised of 75, and 150 lineages of the SARS-CoV-2 are used
- The experiment is ran on both datasets using the original algorithm & the two inexact matching versions
- Number of mismatches allowed is set to 2 (e = 2)
- The distance constraint is set to 100bp, 400bp and 1000bp [3]
- Ran on DelftBlue cluster with 200GB RAM and 12 CPU cores

RESULTS

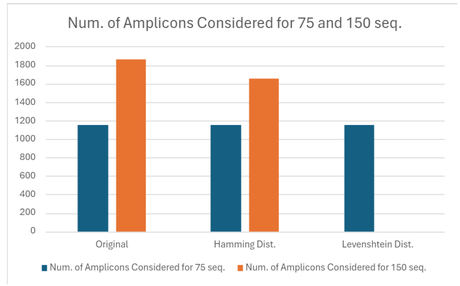

Figure 2: Number of Amplicons considered during the model optimization for distance constraint of 100bp.

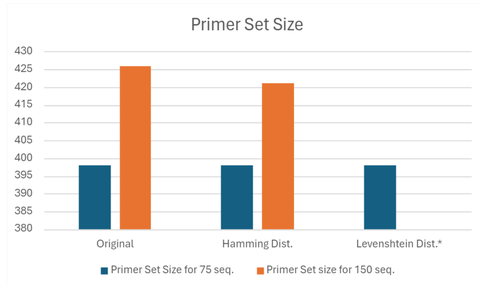

Figure 3: Size of the primer set for distance constraint 100bp.

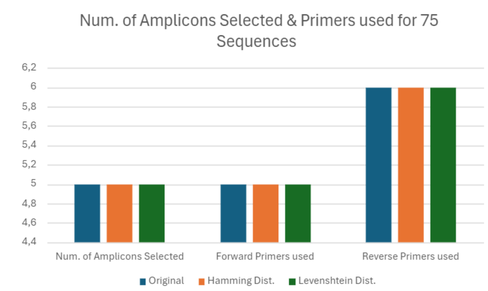

Figure 4: Number of Amplicons selected with their respective primer pair for 75 sequences for distance constraint of 100bp.

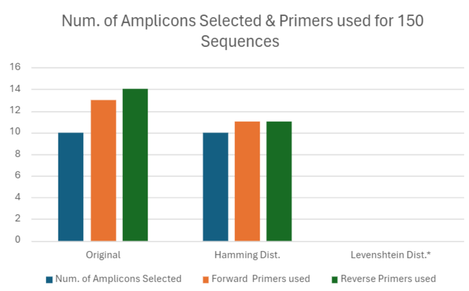

Figure 5: Number of Amplicons selected with their respective primer pair for 150 sequences for distance constraint of 100bp.

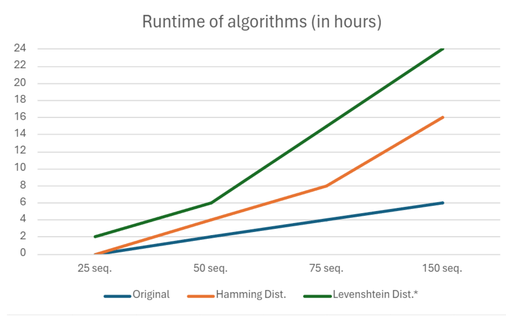

Figure 6: Runtime of the algorithms in hours based on the number of sequences for distance constraint of 100bp.

- Figure 2 suggest:
  - Using Hamming Distance on 150 seq., implies that fewer amplicons need to be considered to find the optimal solution.
  - For 75 sequences, all three variations give the same solution
- Figure 3 shows that:
  - size of the primer set is the same for 75 seq.
  - despite allowing for mismatches, the Hamming distance primer set is smaller for 150 seq.

- As seen in Figures 2, 3 & 5, Levenshtein distance does not produce any results for 150 seq. within 24h of runtime
- Figure 4 shows that when considering 75 sequences, the same amplicons with the same primers are selected
- Figure 5 suggests that when using Hamming distance fewer primers are needed.

CONCLUSION

- Hamming Distance:
  - Allowing for primer mismatches does impact the amplicons in the solution set.
  - It takes fewer amplicons to find the optimal solution
  - The amplicons selected completely differ from the ones in the original algorithms solution
  - It takes significantly more time to compute
- Levenshtein Distance:
  - The similarity introduces overheads such that finding a feasible solution in less than 24h with given computational power
  - Does not contribute to a better solution set for smaller sequences
- Setting the constraint to a value higher than 100bp makes it so that the model is overly stringent, hence, no solution can be found.

FUTURE IMPROVEMENTS

- Using a weighted similarity score
- Relaxing the distance region enclosed constraint
- Optimizing the computation of Levenshtein distance by utilizing a smaller comparison matrix

REFERENCES

[1] Richard W Hamming. Error detecting and error correcting codes. The Bell system technical journal, 29(2):147–160, 1950.

[2] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, volume 10, pages 707–710. Soviet Union, 1966.

[3] Scott W Tighe, Andrew F Hayden, Marcy L Kuentzel, Korin M Eckstrom, Jonathan Foox, Daniel L Vellone, Kristiaan H Finstad, Pheobe K Laaguiby, Jessica J Hoffman, and Sridar V Chittur. Molecular characterization of increased amplicon lengths in sars-cov-2 reverse transcription loop-mediated isothermal amplification assays. Journal of biomolecular techniques: JBT, 32(3):199, 2021.

[4] Jasper van Bemmelen, Davida S Smyth, and Jasmijn A Baaijens. Amplidiff: An optimized amplicon sequencing approach to estimating lineage abundances in viral metagenomes. bioRxiv, pages 2023–07, 2023.