Heuristic-Based Primer Set Minimization for PCR

Thys Kok - T.A.Kok@student.tudelft.nl

Supervisor: Jasper van Bemmelen
Responsible professor: Jasmijn Baaijens

1) Background

Environmental samples are important for studying metagenomic data. In order to be able to analyze them, the useful DNA regions first need to be multiplied in a technique called PCR. Before that can be done, we first need to know which regions are useful. The AmpliDiff algorithm [2] is designed for finding these regions and corresponding DNA pieces needed to initiate the DNA synthesis of these regions, respectively called primers and amplicons.

However, AmpliDiff is held back by its runtime. In the third step of the algorithm, a variant of the Set Cover problem is solved exactly, which takes a lot of time. This project researches heuristics as a possible time-saving alternative.

2) The problem

Primers can only bind in certain sequences depending on the DNA strand; this gives us a similar problem to Set Cover.

3) Iterated Local Search (ILS) framework

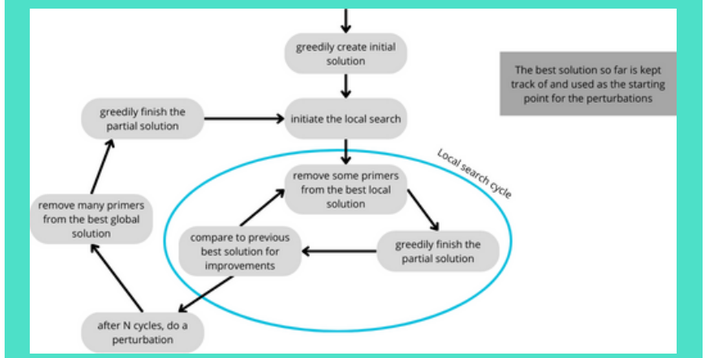
The heuristic chosen is the ILS framework:

- Flexible and easy to extend
- Successful on Set Cover problems [1]
- Based on local search principles

(Initial) Solution creation

To create solutions, all possible primers are sorted on how many sequences they can amplify, then they are greedily picked until every sequence is covered.

In the Figure below, the flow of the algorithm is shown



Hyperparameters:

- The percentage of primers removed during a local search iteration
- The percentage of primers removed during a perturbation (larger shake-up of the solution)

4) Experimental Setup

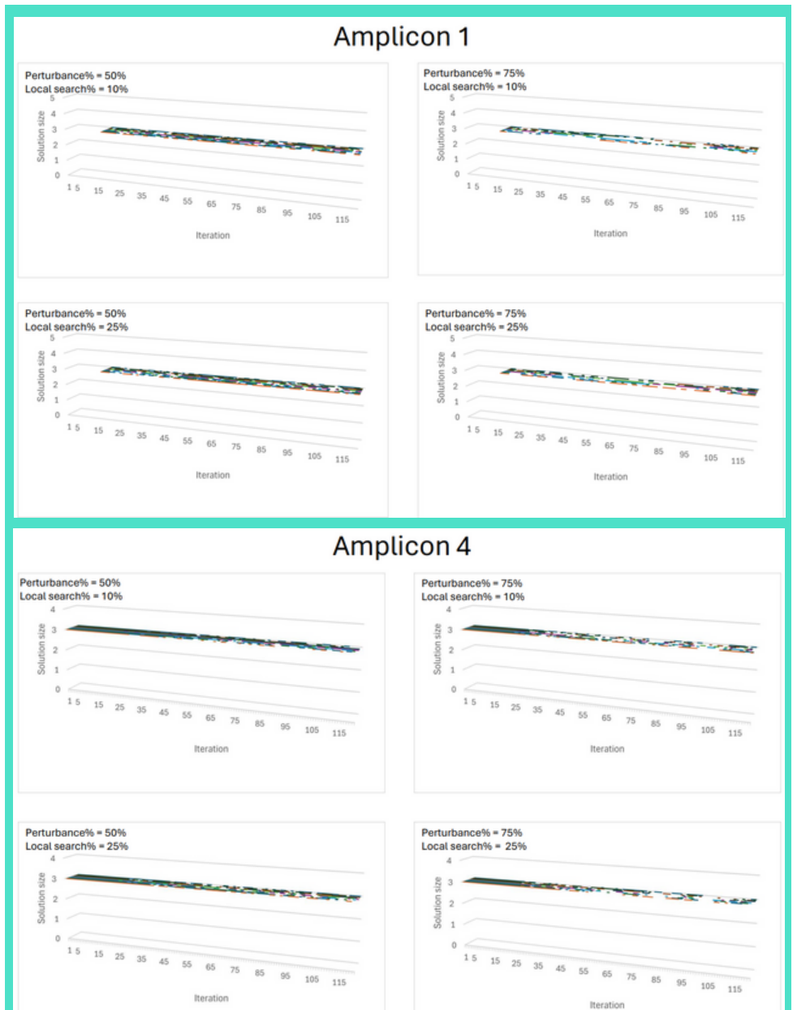
- To test the effectivity of the heuristic, it is compared to the original AmpliDiff.
- Dataset = AmpliDiff example dataset containing 24 sequences
- A larger dataset would be preferred as it better reflects what AmpliDiff normally solves, but due to time constraints this smaller set was used instead
- 10 runs are done to properly evaluate the heuristic, as it is nondeterministic
- For the 2 hyperparameters 10% & 25% and 50% & 75% are tested.

5) Results

- 4 amplicons were needed to distinguish every sequence
- The graphs on the right portray the results for amplicon 1 and amplicon 4
- Each colour is a different run
- X-axis = current iteration
- Y-axis = found solution size for that iteration

For both amplicon 1 and 4 the same solution size as in the original AmpliDiff is found. Depending on the hyperparameters, the frequency with which infeasible solutions are found differs, as the algorithm further explores the solution space.

For amplicon 2 and 3 the optimal solution is instantly found.



6) Conclusion and Future Work

- The heuristics version of AmpliDiff works very well: it finds the same solution as the exact version and does so within 20 cycles for each amplicon.
- However, the dataset that was used, is not representative because of its size. Therefore we recommend future testing on larger datasets. Based on the results, there is reason to do so.

References

[1] Helena Lourenço, Olivier Martin, and Thomas Stützle. Iterated Local Search: Framework and Applications, volume 146, pages 363–397. 09 2010.
[2] Jasper van Bemmelen, Davida S. Smyth, and Jasmijn A. Baaijens. Amplidiff: An optimized amplicon sequencing ap□proach to estimating lineage abundances in viral metagenomes. bioRxiv, 202