CutAdapt Notes

Wednesday, October 09, 2024 6:10 PM

https://cutadapt.readthedocs.io/en/stable/algorithms.html
https://github.com/marcelm/cutadapt

- The algorithm itself is based on free-shift, which shows free or overlap alignments.
- The sequences are allowed to freely shift relative to each other and differences are only penalized in the overlapping region between them.

CutAdapt Procedure:

1. Consider all possible overlaps between sequences and compute an alignment for each
2. Keep only alignments that do not exceed a certain error rate
3. Keep only those alignments that have a maximal number of matches (Updated) [The main problem was the idea of maximizing the number of matches which didn't give desired results]
4. When there is multiple alignments with the same number of matches, it will only keep those with the smallest error rate
5. If multiple candidates are left, choose the alignment that starts at the leftmost position.

The algorithm itself is now seen as a hybrid measuring both distance and score.

- Edit distance is used to fill out the dynamic programming matrix. Which is seen as computing the edit distance for all overlaps between the read and the adapter. We need to use this distance for optimization because we want to let the user provide a certain error rate.
- A second matrix is used with scores filled in simultaneously. The value in a cell is the score of the edit-distance-based alignment, the score is not used as optimization criterion.
- Finally the score is used to decide which overlap between read and adapter is the best.

Score function is: match: +1, mismatch: -1, indel: -2