생물정보학 및 실습 2 실습 보고서 5

2021-20471 박명규
https://github.com/PMKYU98/BioInfo2

Q1 (ba5g)

두 문자열에서 edit distance 구하기

Global alignment 알고리즘 문제 풀었던 것에서 gap 및 mismatch penalty를 양수 1로 변경하고 각 cell에서 최대의 score를 내도록 한 뒤 backtracking 하였다.

문제에서는 distance만 구하면 되므로 backtrack하여 align한 것은 표현하지 않고 점수 값만 제출하였다.

Q2 (ba5h)

Fitting alignment problem

두 문자열 v와 w를 받았을 때, 더 긴 문자열 v의 substring v'와 w의 global alignment를 진행하여 가장 최대의 점수를 찾는 substring v'를 찾는 것이 문제이다.

따라서 global alignment를 진행하되, 문자열 v의 어느 지점에서 시작하여도 w의 끝에만 도달하면 문제가 없으므로 score matrix를 initialization할 때 첫번째 row에는 penalty를 적용하지 않고 0으로 초기화하였다.

그 다음 첫번째 row에 도달할때까지 backtracking하여 표기하였다.

Q3 (ba5i)

Overlap alignment problem

두 문자열 v와 w를 받아서 suffix(v)와 prefix(w) 간에 alignment를 진행하는 문제이다.

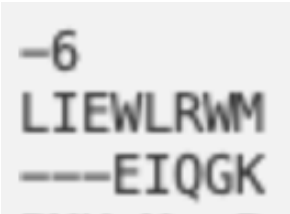
Local alignment를 진행하되 v에서는 suffix를 사용하고 w에서는 prefix를 사용하고 prefix 기준으로 start gap을 피해야하므로 score matrix backtracking에서 그 시작점을 택하지 않도록 해당 부분에 더 큰 페널티를 적용하였다.

Score matrix 계산이 끝난 후에 마지막 column에서 최대 점수를 찾아서 그 부분부터 첫번째 row에 도달할 때까지 backtracking하였다.

Q4 (ba5j)

Global alignment problem with affine gap

Global alignment에서 gap opening에 더 큰 penalty를 주는 문제이나 test를 통과하지 못하였다.



주어진 example case는 통과하였으나, 수정을 진행하며 여러번의 test에 도전하였으나 모두 실패하였다. 그 중 하나의 일부를 추출하여 다시 실행하면 왼쪽과 같다.

직관적으로 보았을 때 아래 sequence의 EI가 아래에 붙어 I에 대한 match 점수를 받고 나머지 gap을 진행하는 것이 더 높은 점수를 받을 것으로 보이는데 그렇지 않은 결과가 나왔다.

위 알고리즘은 각 칸마다 Arrow (진행 방향)을 기록하며 gap을 넣을때 그 직전 칸의 arrow를 참조하여 gap opening인지 extend인지 정하고 점수를 산정하는 방식인데, 실패한 것이다. 다른 case에서는 gap을 열 때 gap을 여는 것보다 일단 mismatch를 줘버리는 게 더 높은 점수를 받아서 올바른 위치에 gap을 열지 못하는 문제도 있었다.

손으로 해 보니 알고리즘 자체에 뭔가 보완점이 있어야 하는데 다른 방식을 생각해내지 못하고 실패하였다.

Q5 (ba5k)

Middle edge in alignment graph

이전에 했던 global alignment 문제에서 middle column과 middle node의 위치를 구해서 어떻게 진행되는지 확인하는 코드를 작성하였다.

Q6 (ba5l)

Alignment using linear space

시간이 부족하여 해결하지 못하였다.
