Project 3 Algorithms in Bioinformatics

Noa, Stinna, Anna and Anastasia

Algorithms in Bioinformatics
Project 3

Noa, Stinna, Anna and Anastasia

Overview

- Alignments and scores
- Methods
- Time and space considerations
- References

Alignments and scores

Case 1 (sequence 1-3):

Exact alignment:

```
>brca1_bos_taurus
atggatttatctgcggatcatgttgaagaagtacaaaatgtcctcaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtctctacaaagtgtga-cca-ca-tattttgcaaattttg-tatg
ctgaa-ac-ttctcaacca-gaagaaagggccttcacaatgtcc--tttgtgtaagaatga-
>brca1_canis_lupus
atggatttatctgcggatcgtgttgaagaagtacaaaatgttcttaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtttctacaaagtgtga-tca-ca-tattttgcaaattttg-tatgc
tgaa-ac-ttctcaacca-gaggaaggggccttcacagtgtcc--tttgtgtaagaacga-
>brca1_gallus_gallus
gcgaa---atgta-aca-cg-gtagaggtgat-cggggtg-cgtt-atac-gtgcgtggtgacctcggtcggtgt-tgacggtgcctggggttcctcagagtgttttggggtctgaaggatg-gacttgtcagtg-attgccattggagacgt
gcaaaatgtgctttcagccatgcagaa-gaa-ctt-ggagtgtccagtctgtttagatgtgat
```

Score:

790

Alignments and scores

Case 1 (sequence 1-3):

Approximate alignment:

```
>brca1_bos_taurus
atggatttatctgcggatcatgttgaaga-agtacaaaatgtcctcaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtctctacaaagtgtgac-ca-ca-tattttgcaaattttg-tat
gctgaa-ac-ttctcaacca-gaagaaagggccttcacaatgtcc--tttgtgtaagaatga-
>brca1_canis_lupus
atggatttatctgcggatcgtgttgaaga-agtacaaaatgttcttaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtttctacaaagtgtgat-ca-ca-tattttgcaaattttg-tatg
ctgaa-ac-ttctcaacca-gaggaaggggccttcacagtgtcc--tttgtgtaagaacga-
>brca1_gallus_gallus
gcgaa---a--tgt-aa-cacggtagaggtgat-cggggtg-cgtt-atac-gtgcgtggtgacctcggtcggtgt-tgacggtgcctggggttcctcagagtgttttggggtctgaaggatg-gacttgtcagtg-attgccattggagacg
tgcaaaatgtgctttcagccatgcag-aagaa-ctt-ggagtgtccagtctgtttagatgtgat
```

Score:

879

Alignments and scores

Case 2 (sequence 1-4):

Approximate alignment:

```
>brca1_bos_taurus
atggatttatctgcggatcgtgttgaagaag-tac--aa-aat-g-ttcttaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtttctacaaagtgtga--tca-c--a-tattttgcaaat-tt
tg-tatgctgaa-ac-ttctcaacca-gagga-aggggcctt-ca--ca-gtgtcc--tttgtgtaagaacga-
>brca1_canis_lupus
atggatttatctgcggatcatgttgaagaag-tac--aa-aat-g-tcctcaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtctctacaaagtgtga--cca-c--a-tattttgcaaat-t
ttg-tatgctgaa-ac-ttctcaacca-gaaga-aagggcctt-ca--ca-atgtcc--tttgtgtaagaatga-
>brca1_gallus_gallus
gcgaa---atgta-aca-cg-gtagaggtga-t-c--gg-ggt-g--cgtt-atac-gtgcgtggtgacctcggtcggtgt-tgacggtgcctggggttcctcagagtgttttggggtctgaaggatg-gac-ttgtc--agtg-attgccatt-g
gagacgtgcaaaatgtgctttcagccatgcaga-a-gaa-ctt--g--ga-gtgtccagtctgtttagatgtgat
>brca1_homo_sapiens
gtaccttgattt-cgtattctg-agaggctgctgcttagcggtagccccttggt-ttccgt--ggcaacggaaa--agcg-cggga-at-tacaga-taaattaaa-a---ct-gcgactgcgcggcgtgagctcg-ctga-gacttcctggac
gggggacaggctgtg-gg-gtttc--tca-gataactgggcccctgcgctcaggaggcc--ttcac-c---ctc-t-
```

Score:

2 770

Alignments and scores

Case 3 (sequence 1-5):

Approximate alignment:

```
>brca1_bos_taurus
atggatttatctgcggatcatgttgaaga-ag-tac--aa-aat-g-tcctcaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtctctacaaagtgtga-c-ca-c--a-tattttgcaaat-tttg-tatgctga
a-ac-ttctcaacca-gaagaaagggccttcacaatgtcc--tttg-tgtaagaatga-
>brca1_canis_lupus
atggatttatctgcggatcgtgttgaaga-ag-tac--aa-aat-g-ttcttaatgctatgca-gaaaatcttag--agtgtccaat-atgtctggagttgatcaaagag-cct-gtttctacaaagtgtga-t-ca-c--a-tattttgcaaat-tttg-tatgctgaa-
ac-ttctcaacca-gaggaaggggccttcacagtgtcc--tttg-tgtaagaacga-
>brca1_gallus_gallus
gcgaa---a--tgt-aa-cacggtagaggtga-t-c--gg-ggt-g--cgtt-atac-gtgcgtggtgacctcggtcggtgt-tgacggtgcctggggttcctcagagtgttttggggtctgaaggatg-ga-cttgtc--agtg-attgccatt-ggagacgtgcaa
aatgtgctttcagccatgcag-aagaa-ctt-ggagtgtccagtctg-tttagatgtgat
>brca1_homo_sapiens
gtaccttgattt-cgtattctg-agaggc-tgctgcttagcggtagccccttggt-ttccgt--ggcaacggaaa--agcg-cggga-at-tacaga-taaattaaa-a---ct-gcgactgcgcggcgtgagctcg-ctga-gacttcctggacgggggacagg
ctgtg-gg-gtttc--tca-gataactgggcccctgcgct-cag--gaggccttcaccctct-
>brca1_macaca_mulatta
atggatttatctgctgttcgcgttgaaga-ag-tac--aa-aat-g-tcattaatgctatgca-gaaaatcttag--agtgtccaat-ctgtctggagttgatcaaggaa-cct-gtctccacaaagtgtga-c-ca-c--a-tattttgcagat-tttg-catgctga
a-ac-ttctcaacca-gaagaaagggccttcacagtgtcc--tttg-tgtaagaatga-
```

Score:

4 103

Alignments and scores

Case 4 (sequence 1-6):

Approximate alignment:

```
>brca1_bos_taurus
a-t-ggatttatctgcggatcgtgttgaagaag-tac--aa-aat-g-ttcttaatg-c-ta-tgca-gaaaatcttag--a-gtgtc-caa-t-atgtctggagttgatcaaagag-cct-gtttctacaaagtgtga--tca-c--a-tattttgcaaa
t-tttg-tatgct-gaa-ac-ttctcaacca-gagga-aggggcct-t-ca--ca-gtgtcc--tttgtgtaagaacga-
>brca1_canis_lupus
a-t-ggatttatctgcggatcatgttgaagaag-tac--aa-aat-g-tcctcaatg-c-ta-tgca-gaaaatcttag--a-gtgtc-caa-t-atgtctggagttgatcaaagag-cct-gtctctacaaagtgtga--cca-c--a-tattttgcaa
at-tttg-tatgct-gaa-ac-ttctcaacca-gaaga-aagggcct-t-ca--ca-atgtcc--tttgtgtaagaatga-
>brca1_gallus_gallus
g-c-gaa---atgta-aca-cg-gtagaggtga-t-c--gg-ggt-g--cgtt-ata-c--g-tgcgtggtgacctcggtcg-gtgt--tga-cggtgcctggggttcctcagagtgttttggggtctgaaggatg-gac-ttgtc--agtg-attgcca
tt-ggagacgtgca-aaatgtgctttcagccatgcaga-a-gaa-ct-t--g--ga-gtgtccagtctgtttagatgtgat
>brca1_homo_sapiens
g-t-accttgattt-cgtattctg-agaggctgctgcttagcggtagccccttggt--t-tc-cgt--ggcaacggaaa--a-gcg-c-ggg-a-at-tacaga-taaattaaa-a---ct-gcgactgcgcggcgtgagctcg-ctga-gacttcctg
gacgggggacaggct-gtg-gg-gtttc--tca-gataactgggcccc-tgcgctcaggaggcc--ttcac-c---ctc-t-
>brca1_macaca_mulatta
a-t-ggatttatctgctgttcgcgttgaagaag-tac--aa-aat-g-tcattaatg-c-ta-tgca-gaaaatcttag--a-gtgtc-caa-t-ctgtctggagttgatcaaggaa-cct-gtctccacaaagtgtga--cca-c--a-tattttgcag
at-tttg-catgct-gaa-ac-ttctcaacca-gaaga-aagggcct-t-ca--ca-gtgtcc--tttgtgtaagaatga-
>brca1_mus_musculus
gttccgaaaggctagcgctaggcgcc-aagcgg-c-c-----ggt-t-tccttggcgacggagagcgcgggaattttag--atagattgtaatt-gcggct-gcg-cggccgctgcc-cgt-gcagccagaggatccag---ca-c--c-tctctt
ggggct-tctc-cgtcctcggc-gc-tt-ggaagta--cgga-tcttttttct-cg--ga-gaaaag--ttcac-t-ggaactg-
```

Score:

7 946

Alignments and scores

Case 5:

Alignment:

In FASTA file (see references)

Clustal Omega: N is always a match

Score:

276 756

This and all previous alignments can be found in this drive:
https://drive.google.com/drive/folders/1ICUdJpiAc6UPY2JVn3ohWIA3zl_Q-jvL?usp=share_link

Methods

Exact algorithm

- Same as for 2 sequences, just with 7 possible last columns instead of 3.
- Indexing with 3 coordinates in 3 dimensions

Approximate algorithm

- Find middle string by performing pairwise alignment of all sequences -> find smallest row sum for the matrix= middle string.
- Align all other seqs to middle string.
- Backtrack for all these, getting alignments.
- Merge alignments so that the multiple alignment is consistent with the pairwise alignments.
- Loop through all the columns in the alignment and add their cost to the total score.

Time and space

Exact algorithm

- Time
  - Filling out 3D-matrix: $O(n^k)$, here **$O(n^3)$**
  - Backtracking: $O(n^k)$, here **$O(n^3)$**
    - The loop makes at *most* n^k iterations, since we iterate over the lengths (at most n) of the k sequences.
    - Each iteration performs lookups (constant time) in the precomputed score matrix.
- Space
  - **$O(n^3)$** aka $O(n^k)$ as the 3D table we fill out accounts for the biggest space consumption in the algorithm.

Approximate algorithm

- Time
  - Find center string by aligning all strings to each other: $O(k^2 \cdot n^2)$
  - Building alignment: $k \cdot n^2$ (fill out) + $k \cdot n^2$ (backtrack) + $k \cdot n^2$ (merge)
    - For each string, we align with the reference string, backtrack pairwise through the alignment with with the reference string.
    - We “merge” with the reference string.
    - The merge takes time proportional to the length of the strings.
  - All in all: $k^2 \cdot n^2 + k \cdot n^2 + k \cdot n^2 + k \cdot n^2 =$ **$O(n^2)$**
- Space
  - $O(n^k)$ here **$O(n^3)$**

References

LINK HERE

How do our algorithms work? | Exact

Exact global alignment for 3 sequences

- Same as for 2 sequences, just with 7 possible last columns instead of 3.
- Indexing with 3 coordinates in 3 dimensions

Time and space complexity

- Time
  - Filling out 3D-matrix: O(n^k), here O(n^3)
  - Backtracking: O(n^k) ??

    How far we backtrack depends on…
    - The length of the sequences
    - How many cases we have to consider each time depends on *k*.
- Space
  - O(n^3) aka O(n^k) as the 3D table we fill out accounts for the biggest space consumption in the algorithm.

How do our algorithms work? | SP- algorithm

SP-approximation algorithm for *k* sequences

- Find middle string by performing pairwise alignment of all sequences -> find smallest row sum for the matrix= middle string.
- Align all other seqs to middle string.
- Backtrack for all these, getting alignments.
- Merge alignments so that the multiple alignment is consistent with the pairwise alignments.
- Loop through all the columns in the alignment and add their cost to the total score.

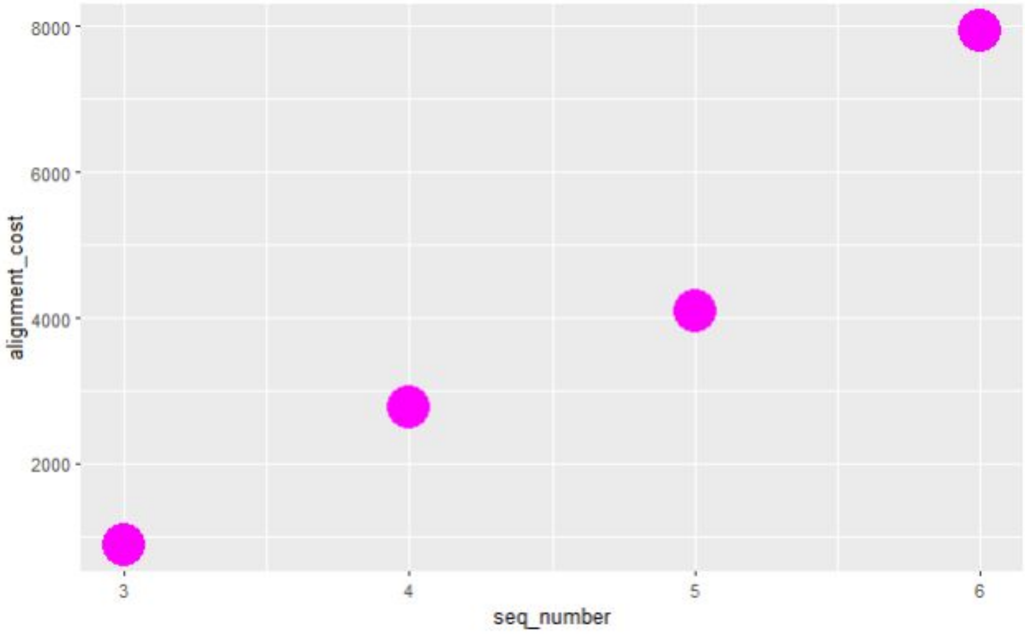
Time and space complexity

- Time
  - Find center string by aligning all strings to each other: $O(k^2 \cdot n^2)$
  - Building alignment: $k \cdot n^2$ (fill out) + $k \cdot n^2$ (backtrack) + $k \cdot n^2$ (merge)
    - For each string, we align with the reference string, backtrack through the alignment with with the reference string.
    - We “merge” with the reference string.
    - The merge takes time proportional to the length of the strings.)
  - All in all: $k^2 \cdot n^2 + k \cdot n^2 + k \cdot n^2 + k \cdot n^2 = O(n^n)$
- Space: $O(n^k)$

Cost of alignment of groups of test BRCA-sequences (SP)

Cost of alignments

seq 1-3	879
seq 1-4	2770
seq 1-5	4103
seq 1-6	7946



Alignments of groups of test BRCA-sequences (SP)

```
>brca1_bos_taurus
a-t-ggatttatctgcggatcgtgttgaagaag-tac--aa-aat-g-ttcttaatg-c-ta-tgca-gaaaatcttag--a-gtgtc-caa-t-atgtctggagttgatcaaagag-cct-gtttctacaaagtgtga--tca-c--
a-tattttgcaaat-tttg-tatgct-gaa-ac-ttctcaacca-gagga-aggggcct-t-ca--ca-gtgtcc--tttgtgtaagaacga-

>brca1_canis_lupus
a-t-ggatttatctgcggatcatgttgaagaag-tac--aa-aat-g-tcctcaatg-c-ta-tgca-gaaaatcttag--a-gtgtc-caa-t-atgtctggagttgatcaaagag-cct-gtctctacaaagtgtga--cca-c-
-a-tattttgcaaat-tttg-tatgct-gaa-ac-ttctcaacca-gaaga-aagggcct-t-ca--ca-atgtcc--tttgtgtaagaatga-

>brca1_gallus_gallus
g-c-gaa---atgta-aca-cg-gtagaggtga-t-c--gg-ggt-g--cgtt-ata-c--g-tgcgtggtgacctcggtcg-gtgt--tga-cggtgcctggggttcctcagagtgttttggggtctgaaggatg-gac-ttgtc--
agtg-attgccatt-ggagacgtgca-aaatgtgctttcagccatgcaga-a-gaa-ct-t--g--ga-gtgtccagtctgtttagatgtgat

>brca1_homo_sapiens
g-t-accttgattt-cgtattctg-agaggctgctgcttagcggtagccccttggt--t-tc-cgt--ggcaacggaaa--a-gcg-c-ggg-a-at-tacaga-taaattaaa-a---ct-gcgactgcgcggcgtgagctcg-c
tga-gacttcctggacgggggacaggct-gtg-gg-gtttc--tca-gataactgggcccc-tgcgctcaggaggcc--ttcac-c---ctc-t-

>brca1_macaca_mulatta
a-t-ggatttatctgctgttcgcgttgaagaag-tac--aa-aat-g-tcattaatg-c-ta-tgca-gaaaatcttag--a-gtgtc-caa-t-ctgtctggagttgatcaaggaa-cct-gtctccacaaagtgtga--cca-c--
a-tattttgcagat-tttg-catgct-gaa-ac-ttctcaacca-gaaga-aagggcct-t-ca--ca-gtgtcc--tttgtgtaagaatga-

>brca1_mus_musculus
gttccgaaaggctagcgctaggcgcc-aagcgg-c-c-----ggt-t-tccttggcgacggagagcgcgggaattttag--atagattgtaatt-gcggct-gcg-cggccgctgcc-cgt-gcagccagaggatccag---
ca-c--c-tctcttggggct-tctc-cgtcctcggc-gc-tt-ggaagta--cgga-tcttttttct-cg--ga-gaaaag--ttcac-t-ggaactg
```

Cost of alignment of all 8 full BRCA-sequences (SP)

- COST

- ADD REPRESENTATION OF THE ALIGMENTS!!

- What did we do to deal with N’s in the rat sequence?

Where to find our beautiful alignments?

- SOME KIND OF LINK (THEY HAVE TO BE IN FASTA-FORMAT) !