Grundlagen der Bioinformatik

Exercises – Assignment 2

Ulf Leser

Assignment 2 - Overview

- Read a pair of DNA sequences in FASTA format
- Computer their local alignment(s)
- Align some real sequences

Task 1

- In moodle, you find two files, each containing two sequences
  - example1.fasta (short sequences)
  - example2.fasta (long sequences)

- Write a program that reads a pair of sequences
- .. and computes their local alignment score
  - with unit cost model (M: 1; I/R/D: -1)
- … and outputs "all" optimal local alignments
- Submission: program called "**localalign**" with one parameter: Name of FASTA file

Expected input / output

- Input file
```
>seq1
AATCG
>seq2
TGCCG
```
- Output (nothing else)
  - 2
  - AATCG
    ***CG
- Input file
```
>seq1
AATCG
>seq2
AACG
```
- Output (nothing else)
  - 3
  - AATCG
    AA_CG
- Input file
```
>seq1
AATTTCG
>seq2
AACG
```
- Output (nothing else)
  - 2
  - AATTTCG
    *****CG
  - AATTTCG
    AA*****

Longer example

- Input file

```
>seq1
AATCGGGATCGATAGCTACGATAGTGTACAGATACTCGCATAGCTA
>seq2
GTCTGATACTACG
```

- Output (nothing else)
  - 9
  - `AATGGGGATC_GATAGCTACTGATAGTGTACAGATACTCGCATAGCTA`
    `******G_TCTGATA_CTAC_G**************************`
  - `AATGGGGATC_GATAGCTACTGATAGTGTACAGATACTCGCATAGCTA`
    `********TCTGATA_CTAC_G**************************`
  - `AATGGGGATC_GATAGCTACTGATAGTGTACAGATACTCGCATAGCTA`
    `******G_TCTGATA_CTAC*****************************`
  - …

What to Output?

- We output alignment cores
- Definition: An alignment core is a trace-back from a global maximum in the table to a cell with value 0
- But: there can be many paths from a maximum to a zero, and the maximum an occur multiple times in the matrix
- Compute one trace-back for each core
- Before and after the core (from optimal score to 0) – pad with "*"
- Thus: Only one output per occurrence of the maximum in the table

Competition

- Compute the local alignment score as fast as possible
  - Use whatever tricks you find
  - No need to compute or output the alignments
  - Implementation may be different from solution to Task 1
  - Submission (voluntarily): program called "**localalignment_competition**" with one parameter
- We will measure wall clock time (unix time) for a few new pairs of sequences
  - Alphabet is {A,C,G,T}

Things you may consider for speeding-up alignment

- This is difficult – essentially only engineering tricks

- Stop early when no better-than-current solution is possible anymore
- Use compact, branch-free code
- Use byte matrix instead of integer matrix
- Parallelize (not so obvious)
- Four Russians trick (complex)
- …

Task 2: Align some real sequences

- KRAS is a member of the RAS protein family
- KRAS is a transcription factor central for many signaling pathways controlling cell proliferation and differentiation
- KRAS is an important and long known oncogene
- Mutation status determines applicability of certain drugs
  - "Approximately 30-50% of colorectal tumors are known to have a mutated (abnormal) *KRAS* gene […] Patients with mutated KRAS CRC are unlikely to benefit from anti-EGFR therapy; it remains unclear whether patients with KRAS wild-type CRC will definitely respond, although these individuals may be able to derive some benefit from anti-EGFR therapy. Patients with metastatic CRC who are being considered for anti-EGFR antibody therapy should be tested for the presence of a KRAS mutation prior to therapy"
    - https://emedicine.medscape.com/article/1690010-overview

Example of a Targeted Therapy

- Chemotherapy essentially evict all dividing cells
  - Affecting also immune system, hair growth, and blood production
- Targeted therapies try to directly block/activate oncogenes or tumor suppressor genes
  - Often depending on mutational status
  - Very successful; driving force in last decade
  - Breast cancer. About 20% to 25% of breast cancers have too much of a protein called human epidermal growth factor receptor 2 (HER2). If the cancer is "HER2 positive", there are many targeted therapy options.
  - Chronic myeloid leukemia. Almost all cases of CML are driven by the formation of a gene called *BCR-ABL*. This was the very first mutation and cancer treated with targeted therapy.
  - Colorectal cancer. Colorectal cancer often makes too much of a protein called epidermal growth factor receptor (EGFR). Drugs that block EGFR may help stop or slow cancer growth. These cancers have no mutation in the *KRAS* gene.
  - Lung cancer. Drugs that block EGFR may also stop or slow lung cancer growth. There are also drugs for lung cancer with mutations in the *ALK* and *ROS* genes.
  - Lymphoma … melanoma … cervical cancer … brain …

source: https://www.cancer.net/

Example: Antibody-based Drugs

Cancer drugs A to Z list

There are many cancer drugs and cancer drug combinations. They have individual side effects. The list includes chemotherapy, hormone therapies, targeted cancer drugs and bisphosphonates. The drugs are listed in alphabetical order by pharmacy (generic) name and brand name.

A to Z list of cancer drugs including combination treatments

ab

A

- Alemtuzumab (Campath, MabCampath)
- Atezolizumab (Tecentriq)
- Avelumab (Bavencio)

B

- Bevacizumab (Avastin)
- Blinatumomab (Blincyto)
- Brentuximab (Adcetris)

C

- Cetuximab (Erbitux)

D

- Denosumab (Prolia, Xgeva)
- Durvalumab (Imfinzi)

F

- Fludarabine, cyclophosphamide and rituximab (FCR)

G

- Gemtuzumab ozogamicin, daunorubicin and cytarabine

I

- Inotuzumab ozogamicin (Besponsa)
- Ipilimumab (Yervoy)
- Ipilimumab and nivolumab

N

- Nivolumab (Opdivo)

O

- Obinutuzumab (Gazyvaro)

P

- Panitumumab (Vectibix)
- Pembrolizumab (Keytruda)
- Pertuzumab (Perjeta)
- Polatuzumab vedotin, bendamustine and rituximab (Pola-BR)

R

- Rituximab (Mabthera, Rixathon, Truxima)

T

- Trastuzumab and pertuzumab
- Trastuzumab emtansine (Kadcyla)

Source: https://www.cancerresearchuk.org/

Task 2: Align real sequences

- Download the human KRAS wild type gene from Genbank
  - Genbank: www.ncbi.nlm.nih.gov/nuccore
  - NM_004985.5
- Download the murine KRAS gene
  - NM_021284.7
- Compute the optimal local alignment of both sequences and output one alignment for each core
- Use the „EMBOSS Water" web server to compute a local alignment
  - Use default settings
- Compare yours and the EMBOSS alignment, describe differences, and try to find an explanation for these

Submission (as PDF)

- URL to the human entry and length of DNA sequence
- URL to the murine entry and length of DNA sequence
- Your alignments (one for each core)
- EMBOSS Alignment (screenshot)
- Thoughts on differences

General requirements

- Remember to name all programs as requested
- All programs must run without further installations on GRUENAU2
  - *ssh username@gruenau2.informatik.hu-berlin.de*
- For all programs, source code must be submitted as well
  - Document your code
  - For Java/C etc.: Submit the source code and the compiled binary
- All responses must be submitted as PDF, where the task /assignment of every answer is clearly recognizable
- Zip everything into one file per task and upload via Moodle
  - **AssignmentX_groupY_taskZ.zip**
- Deadline for submissions: 31.5.2022, 0 o'clock

Questions?