Nifty: DNA Sequence Matching

Brian R. King, Edward Talmage

Bucknell University, Lewisburg, PA, USA

Bucknell UNIVERSITY

Learning Goals

Upper-level Algorithm Design course

- See abstract algorithms run
- Support application-minded students
- Increase comfort with Dynamic Programming
- Connect algorithms to real world

Assignment

Problem

Given

1. a database of DNA sequences and
2. a query DNA sequence

which database sequence is most similar to the query?

Assignment

Guidelines

Write a program that gives the user a choice among at least three of

- Longest Common Subsequence,
- Longest Common Substring,
- Edit Distance,
- Needleman-Wunsch,
- an algorithm you design

to determine the best match.

Sample Output

```
Give the filename of your query sequence.
q.txt
Give the filename of your sequence database.
s.txt
What algorithm would you like to run?
0: Exit
1: Longest Common Subsequence
2: Longest Common Substring
3: Needleman-Wunsch
1
Sequence: NC_000011.10:c2161209-2159779 Homo sapiens chromosome 11, GRCh38.p13 Insulin, score: 854
Sequence: NT_176377.1:12394966-12397002 Guinea pig insulin gene, score: 921
Sequence: V00179.1 Dog gene encoding insulin, score: 772
Sequence: V01243.1 Rat gene for insulin 2, score: 951
Sequence: AY092023.1 Gorilla gorilla insulin gene, partial cds, score: 657
Sequence: NC_052536.1 Gallus gallus isolate bGalGal1 chromosome 5, Insulin, score: 544
Sequence: NC_045512.2:21563-25384 SARS-Cov-2 - surface spike protein, score: 1060
Sequence: NC_002018.1:21-1385 Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 6, neuraminid
ase, score: 759
Sequence: NC_007366.1:30-1730 Influenza A virus (A/New York/392/2004(H3N2)) segment 4, hemagglutin
in, score: 830


------------------------------------------------
Best match: NC_045512.2:21563-25384 SARS-Cov-2 - surface spike protein with a score of 1060
```

Implementation Details

- Teams
- Coding
- Demo
- Submission

Implementation Details

- Teams
  - Flexible size, 3-4 ideal
  - I assign, mix skill levels
- Coding
- Demo
- Submission

Implementation Details

- Teams
- Coding
  - File handling
  - Control loop
  - Comparison algorithms
  - Bonus Interface features
- Demo
- Submission

Implementation Details

- Teams
- Coding
- Demo
  - Students run code
  - New tests
  - Discussion, real-world connections
    - What happens on a blank query? Why?
    - How good is your match? Can/should you output a strength, as well as the score?
    - What should you do if there are two really close matches?
    - What are strengths and weaknesses of each algorithm?
- Submission

Implementation Details

- Teams
- Coding
- Demo
- Submission
  - After demo, allows bugfixes

Workload

- Low required programming load
- Simple report
- “Crash-Tolerant”
- Easy Grading

Tips

- The code is not the point!
- Demo discussion is the highest value
- Help with algorithmic understanding (only)
- Helps to know just a little about DNA

Reception

Student Quote 1

"On another note, I really enjoyed both of the projects! They were on interesting topics and felt very doable with what we learned in class."

Student Quote 2

"i really enjoyed the projects in this course and finally started to see the connections to real-world examples."

Thank you!