Aufgabe 5
Globales Alignment

Ulf Leser

Wissensmanagement in der Bioinformatik

Daten

- Wir verwenden neue Daten
- Im Web finden Sie
  - Fünf FASTA Dateien, die jeweils 11 Sequenzen enthalten
    - Jede Datei enthält nur Sequenzen einer Spezies
      - Das sind: Fruchtfliege, Menschen, Fadenwurm, Maus, Schimpanse
    - Die i-ten Sequenzen der Dateien sind Sequenzen desselben Genes in den verschiedenen Spezies
      - Also homologe Sequenzen
  - Eine Datei, die eine Substitutionsmatrix für DNA in einfach lesbarer Form enthält
    - Inklusive Werte für Insertion / Deletion

1. Ähnlichkeit (10 Punkte)

- Schreiben Sie ein Programm, dass
  - Alle Daten einliest
  - Die Sequenzen aller homologen Gene miteinander vergleicht
  - Pro Gen alle paarweisen globalen Ähnlichkeiten ausgibt
    - Sequenz 1

	Mensch	Maus	Schimpanse	Fadenwurm	Fliege
Mensch		?	?	?	?
Maus			?	?	?
Schimpanse				?	?
Fadenwurm					?

    - Sequenz 2
    - ...
  - Für die erste Sequenz alle optimalen Alignments ausgibt

2. Optimale Alignments (6 Punkte)

- In der Regel kann der optimale globale Alignmentscore von mehreren Alignments erreicht werden
- Wir suchen die Zahl optimaler Alignments
  - Eine Möglichkeit, diese zu finden, ist es, in O(n*m) den optimalen Score zu bestimmten und dann alle Pfade per Traceback abzulaufen und zu zählen
  - Da es theoretisch exponentiell viele Pfade geben kann (siehe auch nächste Aufgabe), kann das sehr teuer sein
- Aufgabe: Entwickeln Sie einen Algorithmus, der in O(n*m) die Zahl optimaler Alignments zwischen zwei Zeichenketten A mit |A|=n und B mit |B|=m berechnet
  - Die aktuellen Alignments sind nicht gefragt
  - Hinweis: Dynamische Programmierung hilft!

3. Pfade (4 Punkte)

- Geben Sie eine Formel an, die die Zahl aller Pfade durch eine Matrix der Größe m*n bestimmt
  - Rekursionsformel reicht

Wettbewerb

- Berechnen Sie die Ausgabe möglichst schnell

- Tipp: K-Band könnte helfen
  - Aber: Wenn die Sequenzen sehr unähnlich, braucht man mit dem iterativen K-Band sogar länger
  - Außerdem: Die Sequenzen sind nicht immer exakt gleich lang

Programmaufruf

- Das Programm muss wie folgt aufrufbar sein
  - `java –jar Assignment5.jar file1 file2 file3 file4 file5 subfile`
  - filei: Dateinamen der Speziesdateien mit den Gensequenzen
  - Subfile: Name der Datei mit der Substitutionsmatrix
  - Ausgabe auf STDOUT
  - Für die erste Gengruppe pro Paar alle optimalen Alignments
  - Pro Gen Ausgabe aller paarweisen Scores

Teillösungen

- Für die erste Gengruppe gibt es immer zwischen 1 und 12 optimale Alignments. Beispiele:
  - Mensch – Schimpanse: 1 optimales Alignment
  - Chimpanse – Fliege: 9 optimale Alignments
  - Chimpanse – Fadenwurm: 4 optimale Alignments
- Die Ähnlichkeiten zwischen den Genen der ersten Gruppe sind:

	Mensch	Maus	Schimpanse	Fliege	Fadenwurm
Mensch		16	49	-52	-61
Maus			17	-48	-73
Schimpanse				-51	-55
Fliege					-54

Abgabe

- Bis 10.1.2010 bzw. 12.1.2010, 23.59
- Nur per Mail als TXT Datei

1. Die Ausgabe des Programms aus Aufgabe 1
2. Lösung der Aufgabe 2 als Text
3. Lösung der Aufgabe 3 als Text

- Der Code muss im Quelltext und als ausführbare JAR Datei eingeschickt werden