Übungen zur Vorlesung Sequenzanalyse II

Universität Bielefeld, SoSe 2007
Dr. Sven Rahmann · Dipl.-Inf. Peter Husemann · Dipl.-Biol. Constantin Bannert

http://gi.cebitec.uni-bielefeld.de/teaching/2007summer/sequenzanalyse/

Blatt 10 vom 29.06.2007
Abgabe am 06.07.2007 vor der Vorlesung um 8:30 in H14

Aufgabe 1 Es folgt eine Übersicht der behandelten Themen der Vorlesung. Arbeite für die Tutorien konkrete Fragen zu einzelnen Punkten aus, und reiche sie vorab bei den Tutoren per e-mail ein (`phuseman`, `skaestne` [at] `cebitec`; `stwardzi` [at] `techfak`).

Themen der Vorlesung Sequenzanalyse II

- Suffixbaum-Konstruktion (WOTD)
- effiziente Darstellung von Suffixbäumen im Speicher
- Suffixarray-Konstruktion (naiv, Manber-Myers)
- RNA-Sekundärstrukturvorhersage: Nussinov-Algorithmus
- (Protein-Sekundärstruktur, PASSTA)
- Vorwärts-Rückwärts-Trick bei paarweisem Alignment
- paarweises Alignment mit linearem Platzbedarf (auch: lokal, affine Gapkosten)
- längennormalisiertes lokales paarweises Alignment
- parametrisches Alignment
- multiples Alignment
  - Grundlagen und Definitionen
  - Scoring-Funktionen (SP, WSP, Tree)
  - exakter DP Algorithmus
  - Reduktion des Suchraums laut Carillo-Lipman
  - Center-Star Heuristik (2-Approximation)
  - Divide-and-Conquer Heuristik; Wahl der cutpoints
  - Tree Alignment, Algorithmus von Sankoff, Algorithmus von Fitch
  - Praxis: progressive und segment-basierte Alignment Methoden
- Genomvergleich: Chaining-Algorithmus

Algorithmische Design-Prinzipien in der Sequenzanalyse

- vollständige Aufzählung des Lösungsraumes (geht theoretisch immer; wenig praktikabel bei exponentiell großen Lösungsräumen)
- Dynamic Programming (push vs. pull), nutzt überlappende Teilprobleme und Optimalität von Teillösungen, z.B. Needleman-Wunsch Alignment, Nussinov-Algorithmus, Fitch-Algorithmus zum Labeln innerer Knoten eines Baums

- Greedy-Strategie (z.B. Partitionsberechnung von links nach rechts bei der Maximal-Matches-Distanz)
- Approximationsalgorithmen, z.B. Center-Star-Methode ist eine 2-Approximation

Aufgabe 2 Proteinfamilien lassen sich häufig durch eine Abfolge von relativ gut erhaltenen Sequenzmotiven charakterisieren (strukturbildend oder aktiven Zentren); dazwischen findet man häufig weniger konservierte Sequenzen (loops, etc.).

Wie kann man vorgehen, um in einer gegebenen Menge von Sequenzen, die genau eine Proteinfamilie bildet, die charakteristischen Sequenzmotive zu finden?

Diskutiere sowohl Methoden, die auf (multiplem) Alignment basieren, als auch Alignment-freie Methoden.