Aufgabe 2 und 3 im Vergleich

Smith-Waterman-Algorithmus mit verschiedenen Gapkostenmodellen

Verglichen wurden die Implementierungen des Smith-Waterman-Algorithmus mit konstanten und affinen Gapkosten, die eine diverse Reads enthaltende FASTA-Datei mit der Lambdaphagen-Referenzsequenz matchen. Es wurde primär mit kurzen, beliebigen Sequenzen (siehe unten im Quellcode) und einer verkürzten Lambdaphagen-Sequenz (enthalten als lambda_ref_test.fasta) von unter 500 Basen getestet, da die Laufzeit bei den vollen Sequenzen auf einem Rechner mit 8 GB Arbeitsspeicher die 40 Minuten überschritt.

Vergleichstabelle für die gekürzte Lambdaphagen-Referenz mit den vollen Reads:

Größe	**Konstantes Modell (gap: -2)**	**Affines Modell (op: -2/ext: -1)**	**Affines Modell (op: -5/ext: -1)**
Gefundene Alignments	39	40	35
Max. Score	26 (Read 3)	15 (Read 15)	15 (Read 15)
Max. Alignmentlänge	84 (in Read 15)	72 (in Read 5)	50 (in Read 8)
Gemittelte Scores	13,69	11,18	10,86
Max. Matchanteil	94,1% (Read 15)	94,1% (Read 15)	94,1% (Read 15)

Nimmt man nun zwei kurze Testsequenzen (s.u. im Code, auskommentiert), die sich so sehr ähneln, dass in der Alignment-Suche keine "Gaps" in die Sequenzen eingebaut werden, um Deletions / Insertions auszugleichen, so ähneln sich die Scores für die gefundenen Alignments vollständig. Dies macht auch Sinn, da sie sich lediglich in der Behandlung von Gaps unterscheiden. Wird nun eine sehr lange Sequenz mit einer ebenso langen Referenzsequenz auf Alignments überprüft, so werden sehr viele Gaps eingefügt um stellenweise bessere Übereinstimmungen zu erzwingen. Dies resultiert jedoch bei beiden Programmen in sehr unterschiedlichen Ergebnissen.

So fand die Implementierung mit affinem Scoring (bei gap opening = -2, gap extension = -1) 40 Alignments mit einem best score von 15 (94,1 % Match) in Read 15, während das konstant Modell bei 39 Alignments den best score von 26 (73,4%) im Read 3 fand – bei komplett gleichen Eingabesequenzen. Interessanterweise enthält beim konstanten Modell der Read 15 auch das maximale Match (94,1%, 15 Scorepunkte), das jedoch hier NICHT den höchsten Score erhalten hat. Aus den Ergebnissen lässt sich schließen: Je weiter die Kosten für opening und extension auseinandergehen oder überhaupt implementiert werden, umso geringer die gemittelten Scores.

Allgemein lässt sich jedoch beobachten, dass in der Implementierung mit konstanten Gapkosten viel häufiger, dafür aber nur kürzere Gaps eröffnet werden, während die Gaps beim affinen Modell seltener, dafür aber länger sind – je größer der Unterschied zwischen der gap opening penalty und der gap extension penalty, umso mehr wurde dies deutlich.

Außerdem lässt sich feststellen, dass beim konstanten Modell im Schnitt viel längere Alignments (dafür mit geringeren Matchprozenten) gefunden wurden, was sich auf die niedrigere Schwelle zum Einfügen von Gaps bzw. den fehlenden Unterschied von openings und extensions zurückführen lässt.

Dahingehend ist der Faktor der Gapkosten im Allgemeinen auch zu relevant für die Ergebnisse, denn je nachdem, wie diese gewählt werden, kann es auch innerhalb der gleichen Gapkosten-Art zu sehr unterschiedlichen Ergebnissen kommen. Setzt man beispielsweise die Gap Openings zu hoch an, so können biologisch gesehen punktuelle Insertions/Deletions das Alignment eher „zerschießen“. Die Gapkostenansetzung lässt sich also kaum getrennt davon betrachten, wie viel Verwandtschaft man zwischen den beiden Sequenzen grundsätzlich annehmen kann, ähnlich wie bei den Werten von Substitutionsmatrizen.

Eine typisches Alignment für das konstante Gapkostenmodell: hohe Länge, hoher Score, mehrere kurze Gaps und Matchprozente um die 70%.

```
Lambda-Referenz:        Länge: 495
    ['G', 'G', 'G', 'C', 'G', 'G', 'C', 'G', 'A', 'C', 'C', 'T', 'C', 'G', 'C', 'G', 'G', 'G', 'T', 'T', 'T', 'T', 'C', 'G', 'C',

Read 3                  Länge: 1219
    ['A', 'A', 'A', 'A', 'C', 'C', 'G', 'G', 'G', 'T', 'C', 'C', 'T', 'T', 'T', 'T', 'G', 'C', 'G', 'G', 'G', 'C', 'T', 'C', 'T',

    Maximalwert der Sequenzmatrix: 26

    ##  Maximum 26 found at:
            i: 167  j: 1087

    ##  Maximum 26 found at:
            i: 169  j: 1089

    Referenz:  AAAGAAAGGAAACGACAGGTGCTGAAAG--CGAGGCTTTTTGGCCTCT-GTCGTTTCCTTTCTC
    Symmetrie: | ||||||||||||||||  ||  ||||  ||  || |||  || | | |||| ||| ||||||
    Read-Seq:  AGAGAAAGGAAACGACAGAGGCCAAAAGCTCGCTGC-TTTCAGCATGTCGTCGATTCTTTTCTC

    Sequence Length:    64
    Übereinstimmung:    73.4 Prozent
    Scoring:            26      (Match: 1 ; Mismatch: -1 ; Gap: -2)
    Matches:            47
    MisMatches:         17
    Insertions:         3
    Deletions:          1
-------------------------

    Referenz:  AAAGAAAGGAAACGACAGGTGCTGAAAG--CGAGGCTTTTTGGCCTCT-GTCGTTTCCTTTCTCTG
    Symmetrie: | ||||||||||||||||  ||  ||||  ||  || |||  || | | |||| ||| |||||| |
    Read-Seq:  AGAGAAAGGAAACGACAGAGGCCAAAAGCTCGCTGC-TTTCAGCATGTCGTCGATTCTTTTCTCAG

    Sequence Length:    66
    Übereinstimmung:    72.7 Prozent
    Scoring:            26      (Match: 1 ; Mismatch: -1 ; Gap: -2)
    Matches:            48
    MisMatches:         18
    Insertions:         3
    Deletions:          1
-------------------------
```

Ein typisches Alignment für das affine Modell auf dem gleichen Read: geringere Länge, geringerer Score, längere Gaps, dafür aber weniger, resultierend in geringerer Gesamtübereinstimmung:

```
Lambda-Referenz:        Länge: 495
    ['G', 'G', 'G', 'C', 'G', 'G', 'C', 'G', 'A', 'C', 'C', 'T', 'C', 'G', 'C', 'G', 'G', 'G', 'T', 'T', 'T', 'T', 'C', 'G'

Read 3                  Länge: 1219
    ['A', 'A', 'A', 'A', 'C', 'C', 'G', 'G', 'G', 'T', 'C', 'C', 'T', 'T', 'T', 'T', 'G', 'C', 'G', 'G', 'G', 'C', 'T', 'C'

    Maximalwert der Score-Matrix: 20

    ##  Maximum 20 found at:
            i: 167  j: 1087

    ##  Maximum 20 found at:
            i: 169  j: 1089

    Referenz:  CGAGGCTTTTTGGCCTCTGTCGTTTCCTTTCTC
    Symmetrie: ||  |||||  |      |||| ||| ||||||
    Read-Seq:  CGCTGCTTTCAGCATGTCGTCGATTCTTTTCTC

    Sequence Length:    33
    Übereinstimmung:    63.6 Prozent
    Scoring:            9       (Match: 1 ; Mismatch: -1 ; Gap-Op: -5 ; Gap-Ex: -1)
    Matches:            21
    MisMatches:         12
    Insertions:         0
    Deletions:          0
-------------------------

    Referenz:  CGAGGCTTTTTGGCCTCTGTCGTTTCCTTTCTCTG
    Symmetrie: ||  |||||  |      |||| ||| |||||| |
    Read-Seq:  CGCTGCTTTCAGCATGTCGTCGATTCTTTTCTCAG

    Sequence Length:    35
    Übereinstimmung:    62.9 Prozent
    Scoring:            9       (Match: 1 ; Mismatch: -1 ; Gap-Op: -5 ; Gap-Ex: -1)
    Matches:            22
    MisMatches:         13
    Insertions:         0
    Deletions:          0
-------------------------
```