Übungsblatt 4  
Multiple Sequence Alignment  
Deadline: 1.6.2022 um 20:00 MEZ  
Bioinformatik für Biochemiestudierende  
Dr. Florian Klimm  
Sommersemester 2022  
Aufgabe 4.1 Progressives Alignment (50 %)  
Sie haben die folgenden vier DNA Sequenzen:  
• AATCG  
• AACG  
• CG  
• AA  
und die Scoring Methode  
• Match = +1  
• Mismatch = −1  
• Gap = −2

Teilaufgabe 4.1.1 Alignment mit Progressiver Methode  
Berechnen Sie das progressive Alignment. Berechnen Sie hierzu erst das paarwei-  
se Alignment und eine p-Distanzmatrix. Danach nutzen Sie UPGMA um einen  
phylogenetischen Guide tree zu erstellen. Entlang des Guide tree fügen Sie dann  
schrittweise die Sequenzen hinzu.

**1. Paarweises Alignment**

Es gibt 4 Sequenzen, daher müssen Alignments berechnet werden.

Sequenz 1: AATCG

Sequenz 2: AA-CG

Sequenz 1: AATCG

Sequenz 3: ---CG

Sequenz 1: AATCG

Sequenz 4: AA---

Sequenz 2: AACG

Sequenz 3: --CG

Sequenz 2: AACG

Sequenz 4: AA—-

Sequenz 3: CG-- oder: Sequenz 3: --CG

Sequenz 4: --AA Sequenz 4: AA--

**2. Berechnung der p-Distanz**

Mit D = Anzahl der Positionen, die sich zwischen den beiden Sequenzen unterscheiden

L = Länge der Sequenz

Sequenz 1&2:

Sequenz 1&3:

Sequenz 1&4:

Sequenz 2&3

Sequenz 2&4

Sequenz 3&4

**Distanzmatrix:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |
| **A** | **0** | **0.2** | **0.6** | **0.6** |
| **B** | **0.2** | **0** | **0.5** | **0.5** |
| **C** | **0.6** | **0.5** | **0** | **1** |
| **D** | **0.6** | **0.5** | **1** | **0** |

**3. Guide-Tree (UPGMA)**

X

A

B

A: AATCG

B: AACG

|  |  |  |  |
| --- | --- | --- | --- |
|  | **C** | **D** | **X** |
| **C** | **0** | **1** | **0.55** |
| **D** | **1** | **0** | **0.55** |

**2 Möglichkeiten: C oder D**

X

A

B

C

Y

A: AATCG

B: AACG

C: CG

|  |  |  |
| --- | --- | --- |
|  | D | Y |
| D | 0 | 0.7 |

A

B

C

D

X

Y

Z

A: AATCG

B: AACG

C: CG

D: AA

**4. Multiple Sequence Alignment**

Wichtig: Bestehende Lücken können nicht verändert werden!!

• Match = +1  
• Mismatch = −1  
• Gap = −2

A: AATCG

B: AA-CG

C: ---CG

D: ---AA

Teilaufgabe 4.1.2 Konsens Alignmnet  
Berechnen Sie das Konsens Alignment des erzeugeten Multiple Sequence Align-  
ments.

**Konsens-Alignment**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| A | 2 | 2 | 0 | 1 | 1 |
| T | 0 | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 0 | 3 | 0 |
| G | 0 | 0 | 0 | 0 | 3 |
| - | 2 | 2 | 3 | 0 | 0 |

Teilaufgabe 4.1.3 Bewertung des Multiple Sequence Alignments  
Berechnen Sie den Sum-of-Pairs-Score des erzeugeten Multiple Sequence Align-  
ments.

• Match = +1  
• Mismatch = −1  
• Gap = −2

A: AATCG

B: AA-CG

C: ---CG

D: ---AA

**Sum-of-Pairs-Score:**

Aufgabe 4.2 Multiple Sequence Alignment mit CLUSTAL OMEGA (30 %)  
Downloaden Sie die Primärsequenz des Hemoglobin subunit beta als FASTA Files  
https://www.uniprot.org/ für mindestens sechs verschiedene Organismen.  
Nutzen Sie nun CLUSTAL OMEGA um ein Multiple Sequence Alignment zu erzeugen.  
Untersuchen Sie auch den Phylogenetischen Baum der aus diesem Alignment  
erzeugt wurde.  
Nutzen Sie nun **MVIEW** um das erzeugte Alignment darzustellen und das Konsens  
Alignment zu untersuchen.

Table

Description automatically generated

**Phylogenetischer Stammbaum**

Text

Description automatically generated with low confidence

**A picture containing table

Description automatically generatedMVIEW**

Die Hämoglobin subunit beta Sequenz folgender 6 Arten wurde verglichen:

NOTFU = Nothobranchius furzeri (Türkiser Prachtgrundkärpfling)

PANTS = Panthera tigris sumatrae (Sumatra-Tiger)

PANON = Panthera onca (Jaguar)

SHEEP = Ovis aries (Hausschaf)

HUMAN = Homo sapiens

GORGO = Gorilla gorilla gorilla (Gorilla)

Aus dem von Clustal Omega erstellten phylogenetischen Stammbaum geht, wie schon erwartet, hervor, dass der Mensch mit dem Gorilla clustert und der Jaguar mit dem Sumatra-Tiger. Während das Hausschaf auf der Basis der Hämoglin-Sequenz mit Mensch und Gorilla in einem Cluster liegt, befindet sich der Türkise Prachtgrundkärpfling in keinem der beiden Cluster. Die Konsensus-Sequenzen, welche unter dem Alignment dargestellt sind, wurden nach unterschiedlichen Thresholds berechnet – 100, 90, 80 und 70%. Die Konsensus-Sequenz trägt an Positionen, an denen bei 100, 90, 80 oder 70% der Sequenzen die gleiche Aminosäure ist, eben diese Aminosäure. An den Positionen, an denen das nicht der Fall ist, wird tritt an Stelle einer Aminosäure ein Symbol, dass die an dieser Stelle vorwiegendste physikochemische Eigenschaft der Aminosäuren beschreibt (o=alkoholisch, a=armonatisch, l=aliphatisch, h=hydrophob, p=polar, c=geladen, - = negativ, + = positiv, s=klein, u=sehr klein, t=“turnlike“). Punkte zeigen an, dass in nur einer der Sequenzen die Aminosäure abweicht. In diesem Beispiel hier weicht an mehreren Positionen nur Nothobranchius furzeri von allen anderen Arten ab, was sich auch im phylogenetischen Stammbaum widerspielgelt.

Aufgabe 4.3 Konsensus Alignment mit R (20 %)  
Erstellen Sie ein R Programm das einen Textfile mit einem Multiple Sequence Align-  
ment auswertet und das Konsensus Alignment ausgibt. Wenden Sie Ihr Programm  
auf das Alignment in der Datei human\_ACTG\_isoforms.txt an.  
Hinweis: Der Befehl  
sort(table(x),decreasing= TRUE)[1]  
ermittelt das am häufigsten vorkommende Element im Vektor x und könnte nützlich  
für diese Aufgabe sein.

* Consensus string konstruieren
* Letters an jeder Stelle anschauen
* Sortiermechanismus (siehe oben) 🡪 für jede Stelle bekommt man den am häufigsten auftretenden Wert