17. Übung

Einführung in die Bioinformatik I, 2. Teil
Sommersemester 2021

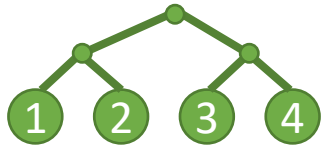
Aufgabe 1 (3 Punkte): Erklären Sie den Unterschied zwischen der Kante (u, v) in einem gerichteten Graph und der Kante $\{u, v\}$ in einem ungerichteten Graph? Was ändert sich, wenn man u und v vertauscht? Ist $u = v$ erlaubt?

	gerichtete Kante	ungerichtete Kante
„Datentyp“		
Tausche u, v		
u=v		

Aufgabe 1 (3 Punkte): Erklären Sie den Unterschied zwischen der Kante (u, v) in einem gerichteten Graph und der Kante $\{u, v\}$ in einem ungerichteten Graph? Was ändert sich, wenn man u und v vertauscht? Ist $u = v$ erlaubt?

	gerichtete Kante	ungerichtete Kante
„Datentyp“	(u, v) ist ein Tupel von Vertices u und v	{u, v} ist eine Menge von Vertices u und v
Tausche u, v	(v, u) bedeutet eine Umkehr der Kantenrichtung	{v, u} hat keinen Effekt
u=v	Erlaubt, entspricht einer Selbstschleife	Nicht erlaubt laut Mengendefinition (aber theoretisch sind auch ungerichtete Selbstschleifen möglich)

Aufgabe 2 (6 Punkte): Gegeben sind die vier Sequenzen $s_1 = \text{GAA}$, $s_2 = \text{GACA}$, $s_3 = \text{ACAC}$ sowie $s_4 = \text{CACG}$ und der Leitbaum $((1,2),(3,4))$. Bestimmen Sie das multiple globale Alignment der vier Sequenzen mit Hilfe des Feng-Doolittle-Verfahrens für die Ähnlichkeitsfunktion $S(A,A) = 2$, $S(a,a) = 1$ für alle $a \neq A$, $S(a,b) = -1$ für $a \neq b$ und $S(a,-) = S(-,a) = -1$ für alle a.



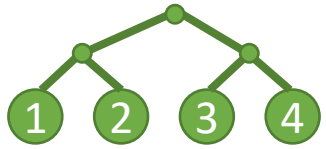
s_1 vs. s_2

	ε	G	A	A
ε				
G				
A				
C				
A				

s_2 vs. s_3

	ε	A	C	A	C
ε					
C					
A					
C					
G					

Aufgabe 2 (6 Punkte): Gegeben sind die vier Sequenzen $s_1 = \text{GAA}$, $s_2 = \text{GACA}$, $s_3 = \text{ACAC}$ sowie $s_4 = \text{CACG}$ und der Leitbaum $((1,2),(3,4))$. Bestimmen Sie das multiple globale Alignment der vier Sequenzen mit Hilfe des Feng-Doolittle-Verfahrens für die Ähnlichkeitsfunktion $S(A,A) = 2$, $S(a,a) = 1$ für alle $a \neq A$, $S(a,b) = -1$ für $a \neq b$ und $S(a,-) = S(-,a) = -1$ für alle a.



s_1 vs. s_2

	ε	G	A	A
ε	0	-1	-2	-3
G	-1	1	0	-1
A	-2	0	3	2
C	-3	-1	2	2
A	-4	-2	1	**4**

```
G A - A
G A C A

G A x A   s'1
G A C A   s'2
```

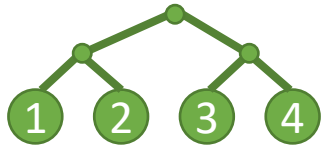
s_2 vs. s_3

	ε	A	C	A	C
ε	0	-1	-2	-3	-4
C	-1	-1	0	-1	-2
A	-2	1	0	2	1
C	-3	0	2	1	3
G	-4	-1	1	1	**2**

```
A C A C -
- C A C G

A C A C x   s'3
x C A C G   s'4
```

Aufgabe 2 (6 Punkte): Gegeben sind die vier Sequenzen $s_1 = \text{GAA}$, $s_2 = \text{GACA}$, $s_3 = \text{ACAC}$ sowie $s_4 = \text{CACG}$ und der Leitbaum $((1,2),(3,4))$. Bestimmen Sie das multiple globale Alignment der vier Sequenzen mit Hilfe des Feng-Doolittle-Verfahrens für die Ähnlichkeitsfunktion $S(A,A) = 2$, $S(a,a) = 1$ für alle $a \neq A$, $S(a,b) = -1$ für $a \neq b$ und $S(a,-) = S(-,a) = -1$ für alle a.



s'$_1$ vs. s'$_3$

	ε	G	A	x	A
ε					
A					
C					
A					
C					
x					

s'$_1$ vs. s'$_4$

	ε	G	A	x	A
ε					
x					
C					
A					
C					
G					

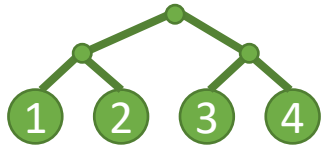
s'$_2$ vs. s'$_3$

	ε	G	A	C	A
ε					
A					
C					
A					
C					
x					

s'$_2$ vs. s'$_4$

	ε	G	A	C	A
ε					
x					
C					
A					
C					
G					

Aufgabe 2 (6 Punkte): Gegeben sind die vier Sequenzen $s_1 = \text{GAA}$, $s_2 = \text{GACA}$, $s_3 = \text{ACAC}$ sowie $s_4 = \text{CACG}$ und der Leitbaum $((1,2),(3,4))$. Bestimmen Sie das multiple globale Alignment der vier Sequenzen mit Hilfe des Feng-Doolittle-Verfahrens für die Ähnlichkeitsfunktion $S(A,A) = 2$, $S(a,a) = 1$ für alle $a \neq A$, $S(a,b) = -1$ für $a \neq b$ und $S(a,-) = S(-,a) = -1$ für alle a.



s'$_1$ vs. s'$_3$

	ε	G	A	x	A
ε	0	-1	-2	-3	-4
A	-1	-1	1	1	0
C	-2	-2	0	1	0
A	-3	-3	0	0	3
C	-4	-4	-1	0	2
x	-5	-4	-1	0	**2**

s'$_1$ vs. s'$_4$

	ε	G	A	x	A
ε	0	-1	-2	-3	-4
x	-1	0	-1	-1	-2
C	-2	-1	-1	-1	-2
A	-3	-2	1	1	1
C	-4	-3	0	1	0
G	-5	-3	-1	0	**0**

s'$_2$ vs. s'$_3$

	ε	G	A	C	A
ε	0	-1	-2	-3	-4
A	-1	-1	1	0	-1
C	-2	-2	0	2	1
A	-3	-3	0	1	4
C	-4	-4	-1	1	3
x	-5	-4	-1	1	**3**

s'$_2$ vs. s'$_4$

	ε	G	A	C	A
ε	0	-1	-2	-3	-4
x	-1	0	-1	-2	-3
C	-2	-1	-1	0	-1
A	-3	-2	1	0	2
C	-4	-3	0	2	1
G	-5	-3	-1	1	**1**

G A C A - - s''$_2$
- A C A C X s''$_3$

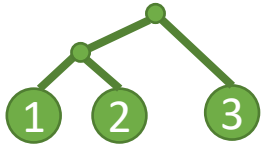
G A - A - -
G A C A - -
- A C A C -
- - C A C G

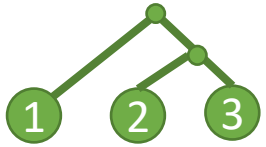
Aus dem besten Alignment (hier s'$_2$ vs. s'$_3$) wird nun das Gesamtalignment berechnet: Neue Gaps aus s"$_2$ in s'$_1$ und s'$_2$ einfügen, sowie neue Gaps aus s"$_3$ in s'$_3$ und s'$_4$ einfügen, dann alle x wieder in Gaps umwandeln.

Aufgabe 3 (6 Punkte): Überlegen Sie sich ein Beispiel für ein progressives globales Alignment mit drei Sequenzen, bei dem sich mit dem Feng-Doolittle-Verfahren je nach Leitbaum unterschiedliche multiple Alignments ergeben.

Aufgabe 3 (6 Punkte): Überlegen Sie sich ein Beispiel für ein progressives globales Alignment mit drei Sequenzen, bei dem sich mit dem Feng-Doolittle-Verfahren je nach Leitbaum unterschiedliche multiple Alignments ergeben.

Versucht erst zu überlegen, was grundsätzlich passieren muss, damit ein MSA aus drei Sequenzen für zwei verschiedene Leitbäume unterschiedlich sein muss. Überlegt euch dann wie die drei Sequenzen aussehen müssen.



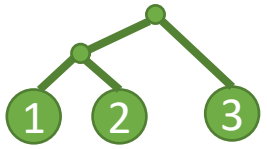


Noch zwei Tipps:
Es klappt schon bei einer Sequenzlänge von 3, mit insgesamt 2 verschiedenen Buchstaben.

Aufgabe 3 (6 Punkte): Überlegen Sie sich ein Beispiel für ein progressives globales Alignment mit drei Sequenzen, bei dem sich mit dem Feng-Doolittle-Verfahren je nach Leitbaum unterschiedliche multiple Alignments ergeben.

<u>Grundidee:</u> Wir brauchen ein Beispiel, in welchem schon im ersten Alignmentschritt, je nach Baum, in Sequenz 2 an unterschiedlichen Stellen Gaps eingefügt werden, denn:
„Once a gap, always a gap."

s_1 = ACA s_2 = AAC s_3 = CAA Ähnlichkeitsfunktion wie bei Aufgabe 2.

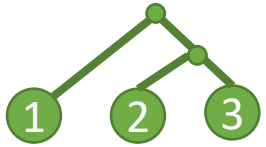


s_1 vs. s_2

	ε	A	C	A
ε	0	-1	-2	-3
A	-1	2	1	0
A	-2	1	1	3
C	-3	0	2	2

A C A - s_1
A – A C s_2

Da im jeweils ersten Alignmentschritt des FD-Verfahrens in s_2 schon unterschiedliche Gappositionen enthalten sind, können die fertigen MSAs nicht mehr identisch sein.



s_2 vs. s_3

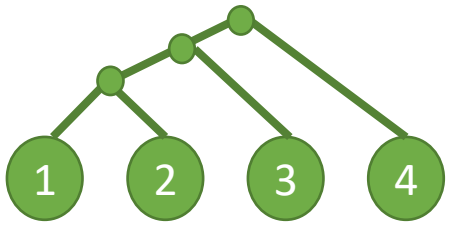
	ε	A	A	C
ε	0	-1	-2	-3
C	-1	-1	-2	-1
A	-2	1	1	0
A	-3	0	3	2

- A A C s_2
C A A - s_3

Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

<u>Zwecks Speicherbedarf:</u>
Erstmal müssen wir überlegen was genau gespeichert wird. Dann überlegen wir wie groß der Speicherbedarf dafür maximal ist und wie oft wir das speichern müssen.

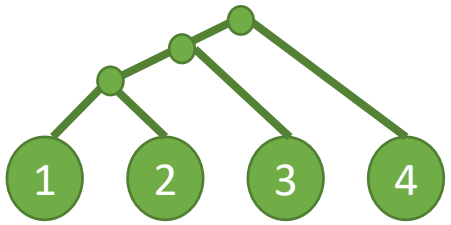


Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

<u>Zwecks Speicherbedarf:</u>
Erstmal müssen wir überlegen was genau gespeichert wird. Dann überlegen wir wie groß der Speicherbedarf dafür maximal ist und wie oft wir das speichern müssen.

Wir speichern Sequenzalignments. Angenommen wir haben k=3 Sequenzen der Länge n=4, dann wäre das Alignment mit dem größten Speicherbedarf für diese Sequenzen das folgende:

```
A A A A - - - - - - - -
- - - - B B B B - - - -
- - - - - - - - C C C C
```



Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

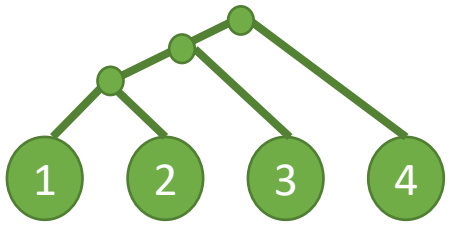
<u>Zwecks Speicherbedarf:</u>
Erstmal müssen wir überlegen was genau gespeichert wird. Dann überlegen wir wie groß der Speicherbedarf dafür maximal ist und wie oft wir das speichern müssen.

Wir speichern Sequenzalignments. Angenommen wir haben k=3 Sequenzen der Länge n=4, dann wäre das Alignment mit dem größten Speicherbedarf für diese Sequenzen das folgende:

```
A A A A - - - - - - - - -
- - - - B B B B - - - - -
- - - - - - - - - C C C C
```

Pro Zeile haben wir einen Speicherbedarf von $k * n$
und wir haben insgesamt genau k Zeilen
→ Speicherbedarf von $k^2 * n$

Wie viele Alignments müssen wir speichern?



Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

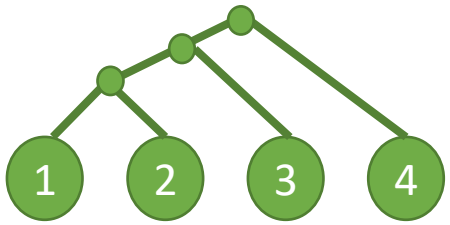
<u>Zwecks Speicherbedarf:</u>
Erstmal müssen wir überlegen was genau gespeichert wird. Dann überlegen wir wie groß der Speicherbedarf dafür maximal ist und wie oft wir das speichern müssen.

Wir speichern Sequenzalignments. Angenommen wir haben k=3 Sequenzen der Länge n=4, dann wäre das Alignment mit dem größten Speicherbedarf für diese Sequenzen das folgende:

```
A A A A - - - - - - - -
- - - - B B B B - - - -
- - - - - - - - C C C C
```

Pro Zeile haben wir einen Speicherbedarf von $k * n$
und wir haben insgesamt genau k Zeilen
→ Speicherbedarf von $k^2 * n$

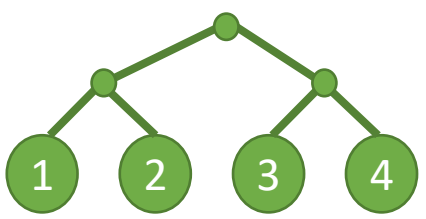
Wie viele Alignments müssen wir speichern?
Für jeden inneren Knoten unseres Leitbaums ein MSA

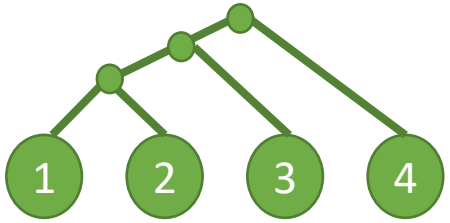


→ Es gibt immer genau k-1 innere Knoten in einem Leitbaum
→ <u>Gesamtspeicherbedarf</u> liegt in $O(k^3 * n)$

Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

<u>Zwecks Laufzeit:</u>
Erstmal müssen wir überlegen was genau der Basisalgorithmus ist, welcher angewendet wird.
Dann überlegen wir welche Laufzeit dieser hat und dann wie oft wir diesen anwenden müssen.



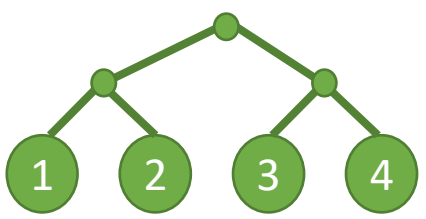


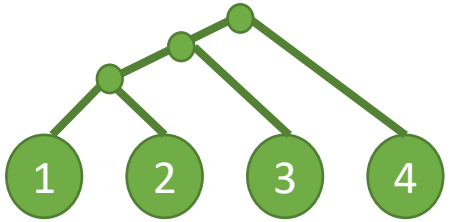
Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

<u>Zwecks Laufzeit:</u>
Erstmal müssen wir überlegen was genau der Basisalgorithmus ist, welcher angewendet wird.
Dann überlegen wir welche Laufzeit dieser hat und dann wie oft wir diesen anwenden müssen.

Feng-Doolittle führt eigentlich immer nur paarweise Alignments durch.
→ Laufzeit für paarweises Alignment ist $O(n^2)$

Wie viele paarweise Alignments müssen wir durchführen?





Aufgabe 4 (5 Punkte): Analysieren Sie Laufzeit und Speicherbedarf des Feng-Doolittle-Verfahrens bei Eingabe von k Sequenzen der Länge n und gegebenem Leitbaum.
Annahme: Die Sequenzen dürfen sich während des Progressieven Alignemnents nur um einen konstanten Faktor verlängern (Das bedeutet, selbst mit eingefügten Gaps ist die Länge einer Sequenz immer $O(n)$).

<u>Zwecks Laufzeit:</u>
Erstmal müssen wir überlegen was genau der Basisalgorithmus ist, welcher angewendet wird.
Dann überlegen wir welche Laufzeit dieser hat und dann wie oft wir diesen anwenden müssen.

Feng-Doolittle führt eigentlich immer nur paarweise Alignments durch.
→ Laufzeit für paarweises Alignment ist $O(n^2)$

Wie viele paarweise Alignments müssen wir durchführen?

Betrachten wir den obersten inneren Knoten; dort müssen wir maximal $\left(\frac{k}{2}\right)^2$ paarweise Vergleiche durchführen.

Da wir immer noch $k-1$ innere Knoten haben
→ <u>Gesamtlaufzeit: $O(k^3 * n^2)$</u>