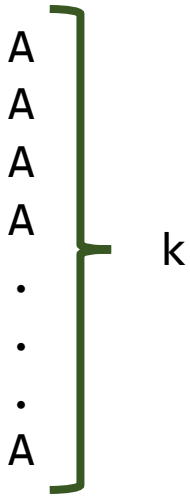
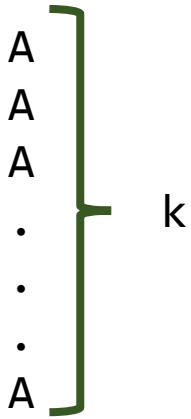
15. Übung

Einführung in die Bioinformatik I, 2. Teil
Sommersemester 2021

Aufgabe 1 (3 Punkte): Gegeben die Ähnlichkeitsfunktion S mit $S(a,a) = 1, S(a,b) = -1$ für $a \neq b$ und $S(a,-) = S(-,b) = -1$ Berechnen Sie den Sum-of-Pairs-Score für eine Spalte mit k Einträgen, in der sich nur 'A's befinden und für eine Spalte mit k Einträgen, in der sich $k-1$ 'A's und ein 'T' befinden. Vergleichen Sie die beiden Werte asymptotisch.



k



k-1

Aufgabe 1 (3 Punkte): Gegeben die Ähnlichkeitsfunktion S mit $S(a,a) = 1, S(a,b) = -1$ für $a \neq b$ und $S(a,-) = S(-,b) = -1$ Berechnen Sie den Sum-of-Pairs-Score für eine Spalte mit k Einträgen, in der sich nur 'A's befinden und für eine Spalte mit k Einträgen, in der sich $k-1$ 'A's und ein 'T' befinden. Vergleichen Sie die beiden Werte asymptotisch.

A
A
A
A
.
.
.
A

k

$$(k-1) * S(A,A) + (k-2) * S(A,A) + (k-3) * S(A,A) + \ldots + S(A,A)$$
$$= \frac{k*(k-1)}{2} = 1/2 * (k^2 - k)$$

oder alternativ:

$$\binom{k}{2} * S(A,A) = \frac{k!}{2! * (k-2)!}$$
$$= 1/2 * \frac{k!}{(k-2)!}$$
$$= 1/2 * \frac{k*(k-1)}{1} = 1/2 * (k^2 - k)$$

A
A
A
.
.
.
A
T

k-1

$$\frac{(k-1)*(k-2)}{2} * S(A,A) + (k-1) * S(A,T)$$
$$= \frac{k^2 - 3k + 2}{2} - (k-1) = 1/2\,k^2 - 2{,}5k + 2$$

Aufgabe 2 (4 Punkte): Bestimmen Sie die besten Consensus-Sequenzen des multiplen Alignments

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

1. für obige Ähnlichkeitsfunktion S mit $S(-,-) = 0$.

Aufgabe 2 (4 Punkte): Bestimmen Sie die besten Consensus-Sequenzen des multiplen Alignments

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

1. für obige Ähnlichkeitsfunktion S mit $S(-,-) = 0$.

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

Mögliche Buchstaben für Position 1 der Consensus-Sequenz: A, G

Sum-of-Pairs Score für A in Spalte 1: $S(A,G) + S(A,A) + S(A,A) + S(A,A) + S(A,A) = -1 + 1 + 1 + 1 + 1 = 3$
Sum-of-Pairs Score für G in Spalte 1: $S(G,G) + S(G,A) + S(G,A) + S(G,A) + S(G,A) = 1 - 1 - 1 - 1 - 1 = -3$
→ A gewinnt.
Analog wird das für alle Spalten des MSA gemacht.
Bei gleichen guten Scores, wie z. B. A und C in Spalte 9 existieren mehrere beste Consensus-Sequenzen.

Lösungen hier: ATGCCC-GCT und ATGCCC-GAT

Aufgabe 2 (4 Punkte): Bestimmen Sie die besten Consensus-Sequenzen des multiplen Alignments

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

2. für die Ähnlichkeitsfunktion S' mit
   - $S'(A, A) = 4$, $S'(C, C) = S'(G, G) = S'(T, T) = 2$,
   - $S'(C, G) = S'(G, C) = 1$, $S'(a, b) = -1$ für alle anderen $a \neq b$, und
   - $S'(a, -) = S'(-, b) = -1$, $S'(-, -) = 0$.

Aufgabe 2 (4 Punkte): Bestimmen Sie die besten Consensus-Sequenzen des multiplen Alignments

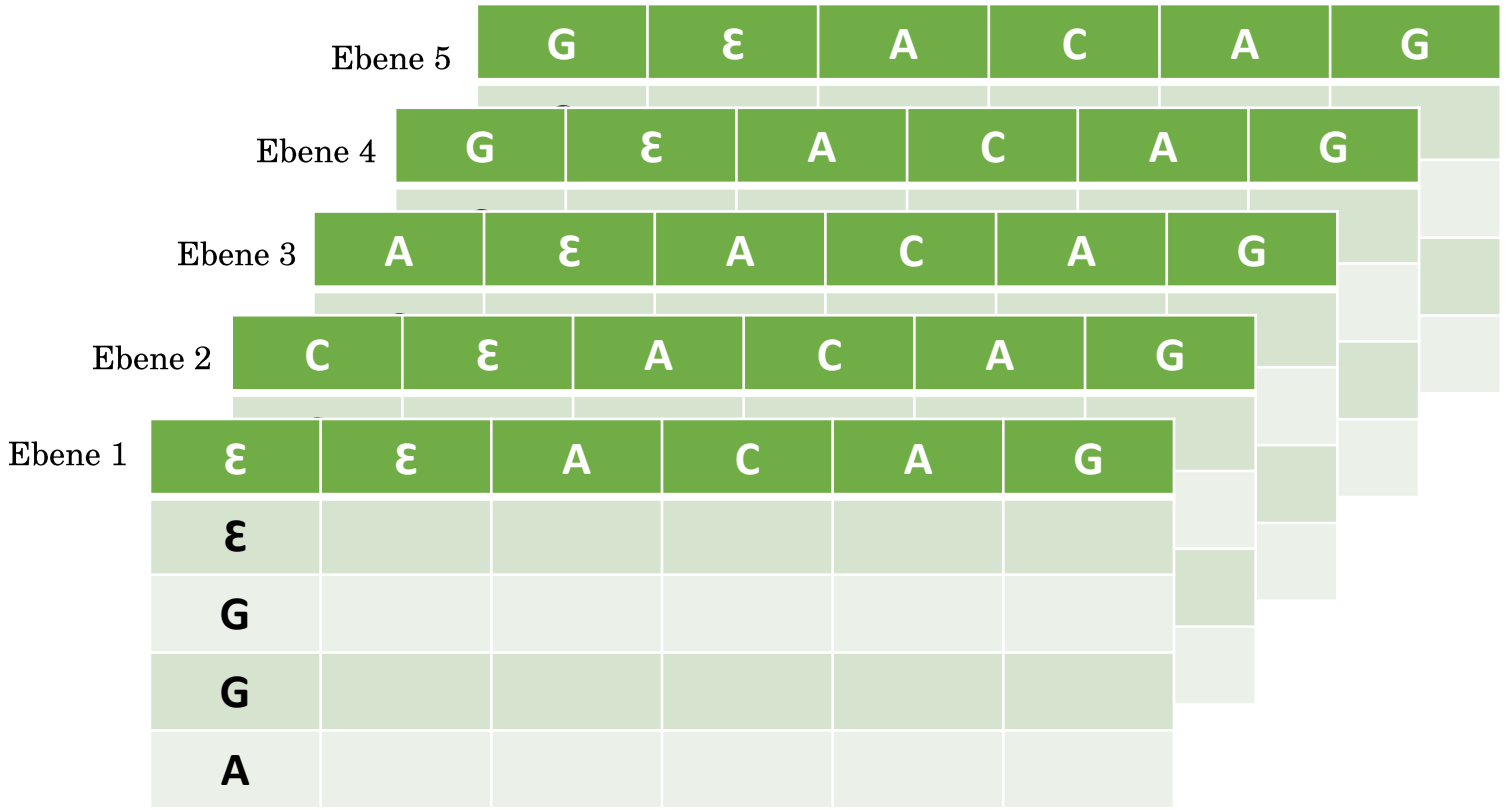
```
G T A A C A T C C A
A T G - C C - - G A
A T A - C C G G C T
A T G C C G - G A T
A T G C T C - G A T
```

2. für die Ähnlichkeitsfunktion S' mit
   - $S'(A, A) = 4$, $S'(C, C) = S'(G, G) = S'(T, T) = 2$,
   - $S'(C, G) = S'(G, C) = 1$, $S'(a, b) = -1$ für alle anderen $a \neq b$, und
   - $S'(a, -) = S'(-, b) = -1$, $S'(-, -) = 0$.

Lösungen hier: ATACCC-GAA und ATACCCGGAA und ATACCCTGAA

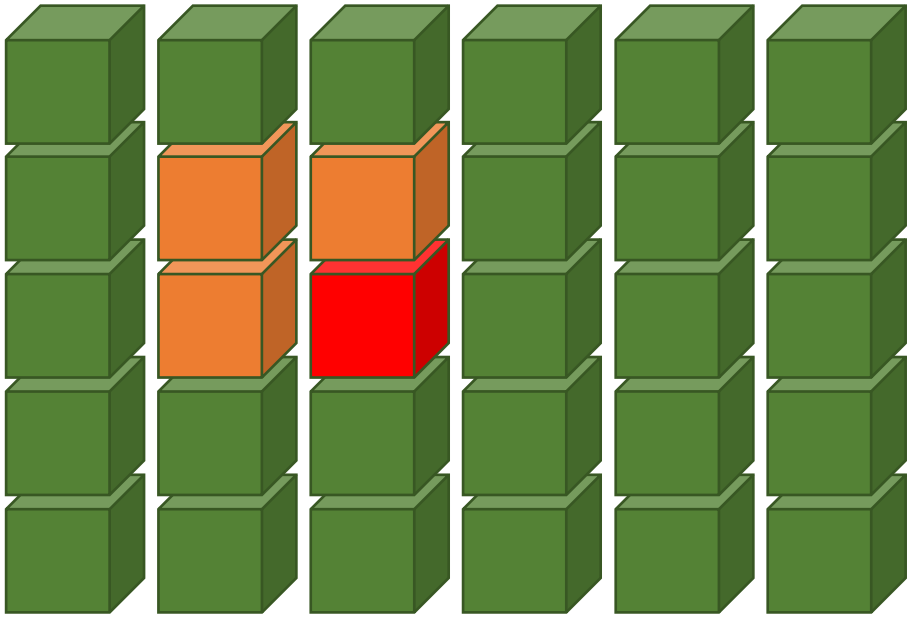
Aufgabe 3 (8 Punkte): Berechnen Sie die Matrix D für das lokale multiple Alignment der drei Sequenzen u =ACAG, v =GGA und w =CAGG für den Sum-of-Pairs-Score zur Ähnlichkeitsfunktion S'' mit $S''(a,a) = 3, S''(a,b) = -1$ für $a \neq b$, $S''(a,-) = S''(-,b) = -1$ und $S''(-,-) = 0$. Schreiben Sie die zweidimensionalen Matrizen $D[i,j,0], D[i,j,1], \ldots, D[i,j,4]$ einzeln auf. Was sind die optimalen Alignments, was ihre Ähnlichkeit?

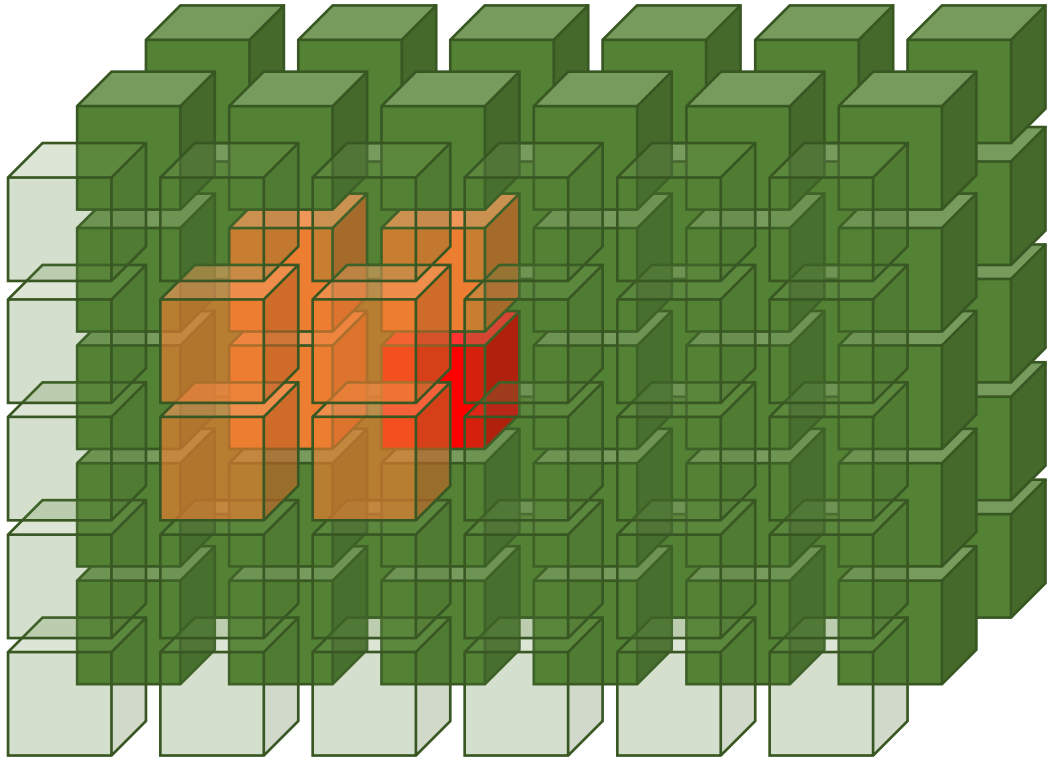
3-dimensionale Matrix D:



D. h. um eine Zelle in D auszufüllen, müssen wir jetzt bis zu 7 verschiedene Möglichkeiten betrachten (also in bis zu 7 Richtungen innerhalb der Matrix schauen).

Aufgabe 3 (8 Punkte): Berechnen Sie die Matrix D für das lokale multiple Alignment der drei Sequenzen u =`ACAG`, v =`GGA` und w =`CAGG` für den Sum-of-Pairs-Score zur Ähnlichkeitsfunktion S'' mit $S''(a,a) = 3, S''(a,b) = -1$ für $a \neq b$, $S''(a,-) = S''(-,b) = -1$ und $S''(-,-) = 0$. Schreiben Sie die zweidimensionalen Matrizen $D[i,j,0], D[i,j,1], \ldots, D[i,j,4]$ einzeln auf. Was sind die optimalen Alignments, was ihre Ähnlichkeit?





D. h. um eine Zelle in D auszufüllen, müssen wir jetzt bis zu 7 verschiedene Möglichkeiten betrachten (also in bis zu 7 Richtungen innerhalb der Matrix schauen).

Initialisierung

Ebene 1

ε	ε	A	C	A	G
ε	0	0	0	0	0
G	0				
G	0				
A	0				

Ebene 2

C	ε	A	C	A	G
ε	0				
G	0				
G	0				
A	0				

Ebene 3

A	ε	A	C	A	G
ε	0				
G					
G					
A					

Ebene 4

G	ε	A	C	A	G
ε	0				
G					
G					
A					

Ebene 5

G	ε	A	C	A	G
ε	0				
G					
G					
A					

Ebene 1

Ɛ	Ɛ	A	C	A	G
Ɛ	0	0	0	0	0
G	0	0	0	0	1
G	0	0	0	0	1
A	0	1	0	1	0

Ebene 2

C	Ɛ	A	C	A	G
Ɛ	0	0	1	0	0
G	0	0	1	0	1
G	0	0	1	0	1
A	0	1	2	2	0

Ebene 3

A	Ɛ	A	C	A	G
Ɛ	0	1	0	2	0
G	0	1	0	2	3
G	0	0	1	0	1
A	1	9	7	10	8

Ebene 4

G	Ɛ	A	C	A	G
Ɛ	0	0	0	0	3
G	1	2	2	3	11
G	1	2	2	3	11
A	0	7	6	8	11

Ebene 5

G	Ɛ	A	C	A	G
Ɛ	0	0	0	0	1
G	1	1	1	1	9
G	2	3	3	4	12
A	0	5	4	6	10

Ebene 1

Ɛ	Ɛ	A	C	A	G
Ɛ	0	0	0	0	0
G	0	0	0	0	1
G	0	0	0	0	1
A	0	1	0	1	0

Ebene 2

C	Ɛ	A	C	A	G
Ɛ	0	0	1	0	0
G	0	0	1	0	1
G	0	0	1	0	1
A	0	1	2	2	0

Ebene 3

A	Ɛ	A	C	A	G
Ɛ	0	1	0	2	0
G	0	1	0	2	3
G	0	0	1	0	1
A	1	9	7	10	8

Ebene 4

G	Ɛ	A	C	A	G
Ɛ	0	0	0	0	3
G	1	2	2	3	11
G	1	2	2	3	11
A	0	7	6	8	11

Ebene 5

G	Ɛ	A	C	A	G
Ɛ	0	0	0	0	1
G	1	1	1	1	9
G	2	3	3	4	**12**
A	0	5	4	6	10

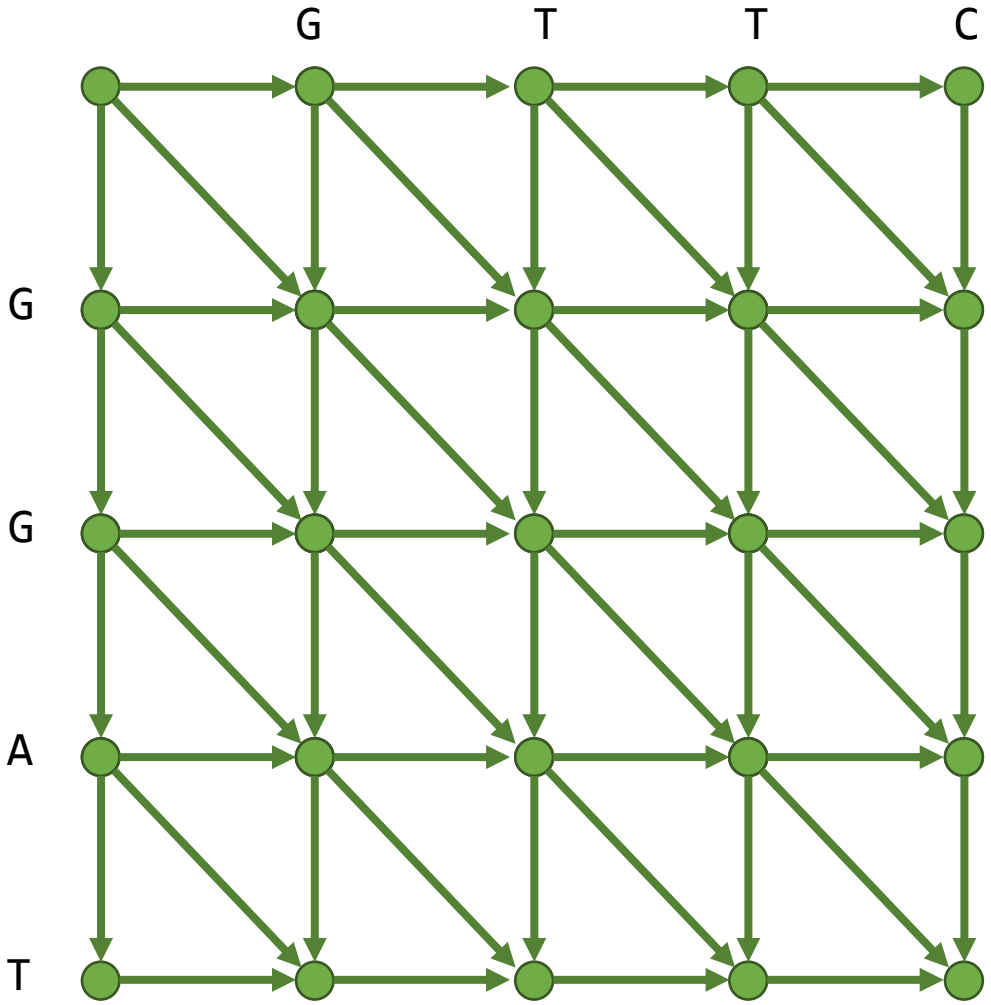
$$\begin{pmatrix} C & A & G & - \\ - & - & G & G \\ C & A & G & G \end{pmatrix}$$

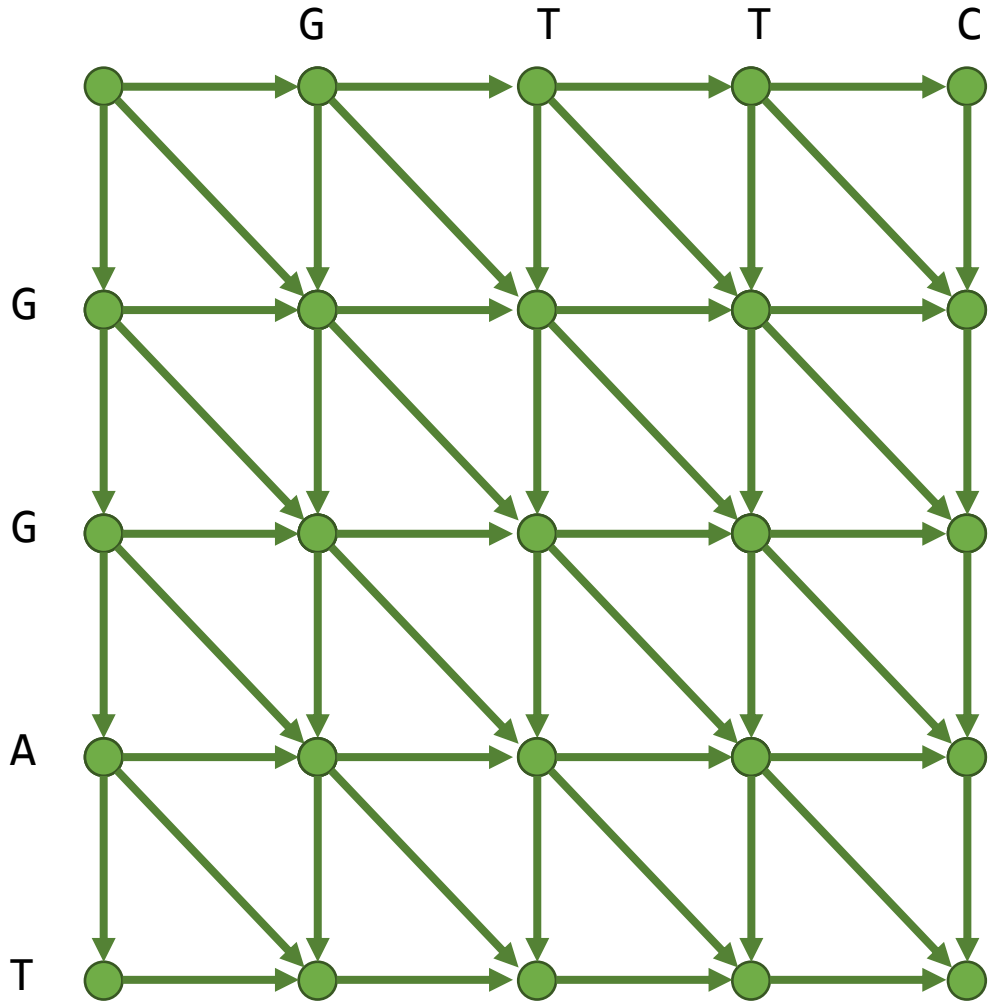
1+1+9+1

Aufgabe 4 (5 Punkte): Erstellen Sie den (vollständig beschrifteten) Edit-Graphen für die Sequenzen $u =$`GTTC` und $v =$`GGAT`

1. mit Einheitskosten
2. mit der Kostenfunktion

$$\delta(a,b) = \begin{cases} 0 & a = b \\ 1 & a,b \in \{\texttt{A},\texttt{G}\} \text{ und } a \neq b \\ & a,b \in \{\texttt{C},\texttt{T}\} \text{ und } a \neq b \\ 2 & \textit{sonst} \end{cases}$$





Aufgabe 4 (5 Punkte): Erstellen Sie den (vollständig beschrifteten) Edit-Graphen für die Sequenzen $u =$`GTTC` und $v =$`GGAT`

1. mit Einheitskosten
2. mit der Kostenfunktion

$$\delta(a,b) = \begin{cases} 0 & a = b \\ 1 & a,b \in \{\texttt{A},\texttt{G}\} \text{ und } a \neq b \\ & a,b \in \{\texttt{C},\texttt{T}\} \text{ und } a \neq b \\ 2 & \textit{sonst} \end{cases}$$