Sara Kemmler 5760949
Robin Bonkaß 5769588

1	2	3	Σ

Übungsblatt Nr. 10

(Abgabetermin 14.07.22)

Aufgabe 1

Theorem: $S_n = 2^{n-2} - 1$ für $n \geq 3$

Beweis. mittels Induktion:

Induktionsanfang:
n = 3

$$S_3 = 2^{3-2} - 1 = 1$$

n = 4

$$S_4 = 2^{4-2} - 1 = 3$$

Induktionsvoraussetzung:

Die Annahme $S_n = 2^{n-2} - 1$ gelte für ein beliebiges, aber festes $n \geq 3$.

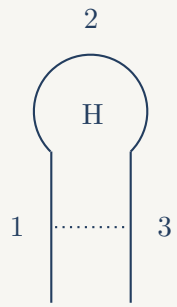
Induktionsschritt: $n \to n + 1$

$$\begin{aligned} S_{n+1} &= 2^{(n+1)-2} - 1 \\ &= 2^{n-1} - 1 \\ &= 2 \cdot 2^{n-2} - 1 \\ &\overset{I.V.}{=} 2 \cdot (S_n + 1) - 1 \\ &= 2 \cdot S_n + 2 - 1 \\ &= 2 \cdot S_n + 1 \end{aligned}$$

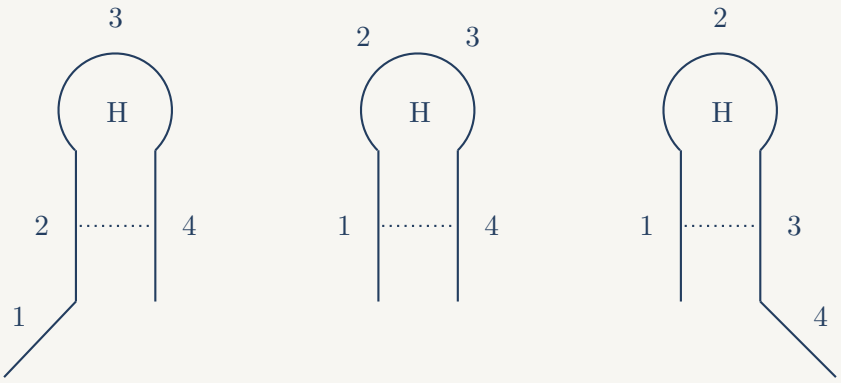
Damit wurde die Aussage bewiesen.

□

Alle Hairpin Strukturen für $n = 3$:



Alle Hairpin Strukturen für $n = 4$:



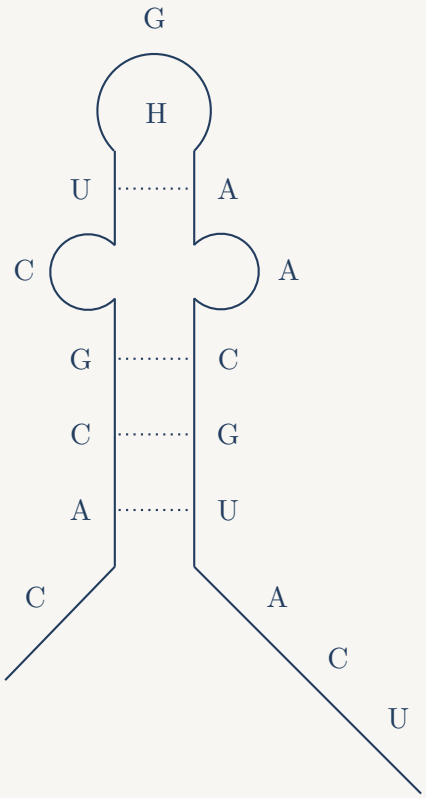
Aufgabe 2

Hairpin Structure

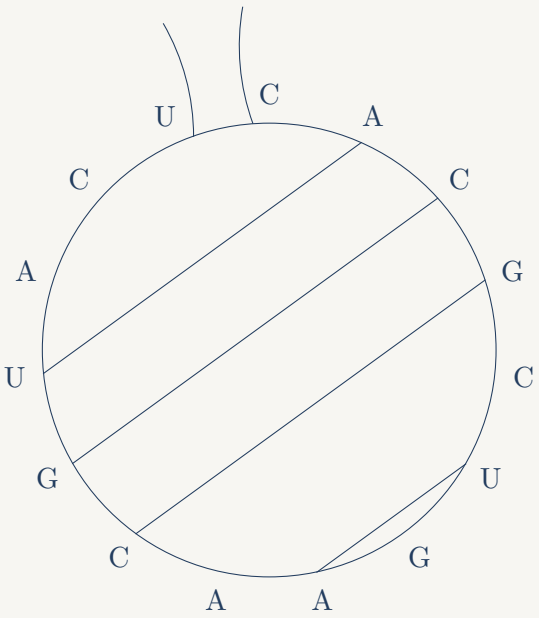
Dot-bracket notation:

C	A	C	G	C	U	G	A	A	C	G	U	A	C	U
.	(	(	(	.	(	.	)	.	)	)	)	.	.	.

Secondary structure as a graph:



Arc representation:

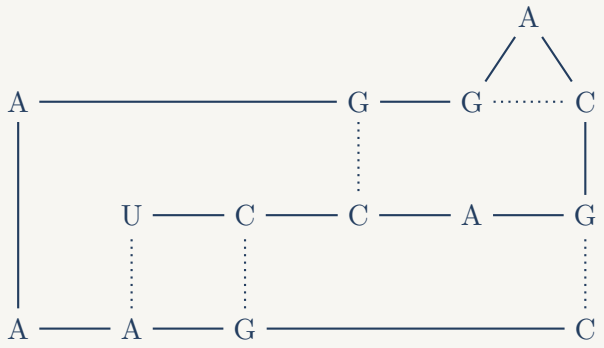


Pseudoknot Structure

Dot-bracket notation:

U	C	C	A	G	C	A	G	G	A	A	A	G	C
(	(	(	.	(	(	.	)	)	.	.	)	)	)

Secondary structure as a graph:



Arc representation:

Eine arc representation kann nicht gezeichnet werden, da es Überschneidungen der Verbindungen gibt.

Aufgabe 3

1. Part

Die Ausgabe von *PROKKA* befindet sich in der Datei `prokka.gff`. Die Ausgabe von *GeneMark* befindet sich in der Datei `genemark.gff`.

2. Part

Der benutzte `.gff`-Parser wurde selbst geschrieben und befindet sich in der Datei `gff_parser.py`. Der Code, welcher jeweils zwei `.ggf`-Dateien vergleicht befindet sich in der Datei `Robin_Bonkass_Sara_Kemmler_A10.py`. Der Code kann mit dem Befehl

```
python3 Robin_Bonkass_Sara_Kemmler_A10.py PAO1_annotation.gff genemark.gff prokka.gff
```

ausgeführt werden. Das Ergebnis wird in der Kommandozeile ausgegeben und eine Besipielausgabe wird im folgenden angegeben und diskutiert.

Ergebnisse

```
1
2  Values for file PAO1_annotation.gff as ground truth and genemark.gff as prediction
3
4  The calculated values are:
5
6  tp = 5512316
7  tn = 619795
8  fp = 57652
9  fn = 74599
10
11 sensitivity = 0.9866475505712903
12 specificity = 0.9148981396330635
13 accuracy = 0.9788885091392402
14
15
16 Values for file PAO1_annotation.gff as ground truth and prokka.gff as prediction
17
18 The calculated values are:
19
20 tp = 5542389
21 tn = 618688
22 fp = 58759
23 fn = 44526
24
25 sensitivity = 0.9920303065287372
26 specificity = 0.9132640634617911
27 accuracy = 0.9835124444456506
```

Zunächst fällt auf, dass die Ergebnisse bei dem Vergleich mit PROKKA und mit GeneMark jeweils um weniger als 0.01 Prozentpunkte abweichen. Sie sind also zueinander hochsignifikant gleich.

Die Sensitivität ist hochsignifikant. Die Genauigkeit ist ebenfalls sehr hoch, jedoch nur signifikant. Die Spezifität ist mit ungefähr 91.33% am geringsten. Es wurden also fast alle CDS gefunden. Es wurden dazu jedoch auch mehr CDS erkannt, als eigentlich existieren.