Marvin Doebel
Rabanus Derr

1	2	3	Σ

Übungsblatt Nr. 3

(Abgabetermin 07.05.2018)

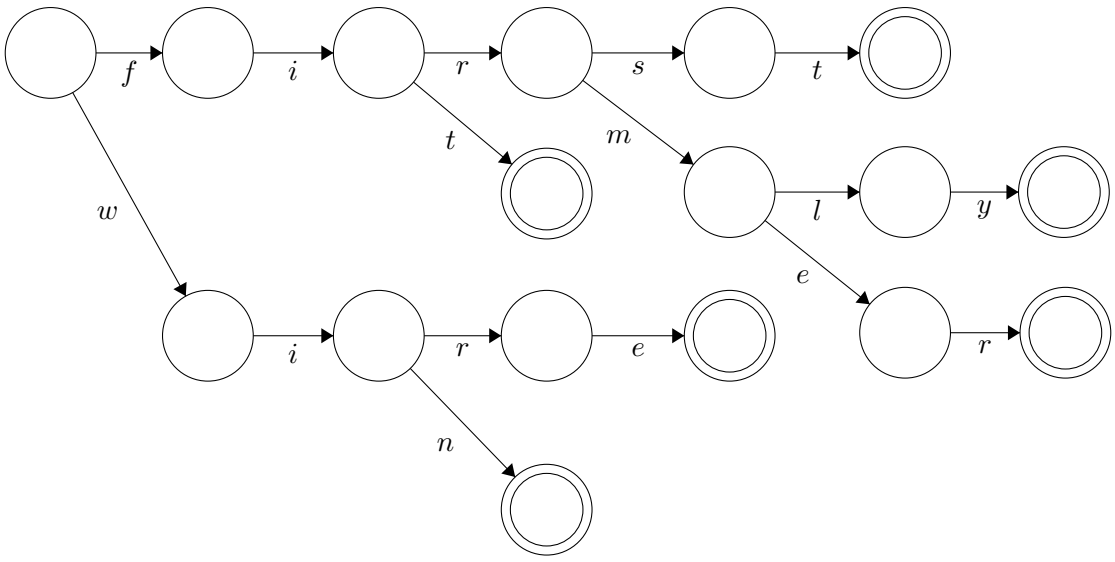
Aufgabe 1

Ein keyword tree ist ein gerichteter Graph, oder ein Baum für die Menge S, die alle Seeds enthält. Er hat folgende Eigenschaften:

- Jede Kante zwischen zwei Knoten ist mit genau einem Zeichen aus einem Element aus S beschriftet
- Jedes Element aus S bekommt vom Anfangsknoten für jedes Zeichen im Element aus S eine Transition auf einen neuen Knoten, ausser eine passend gelabelte Transition existiert bereits.
- Der Knoten, der am Ende einer Transitionsfolge steht bleibt ohne Transitionen und markiert das Ende eines Wortes. Ausnahme...
- ...Wenn Präfixe existieren werden unbeschriftete Transitionen so einführt, dass sich überlappende Wörter erkannt werden.

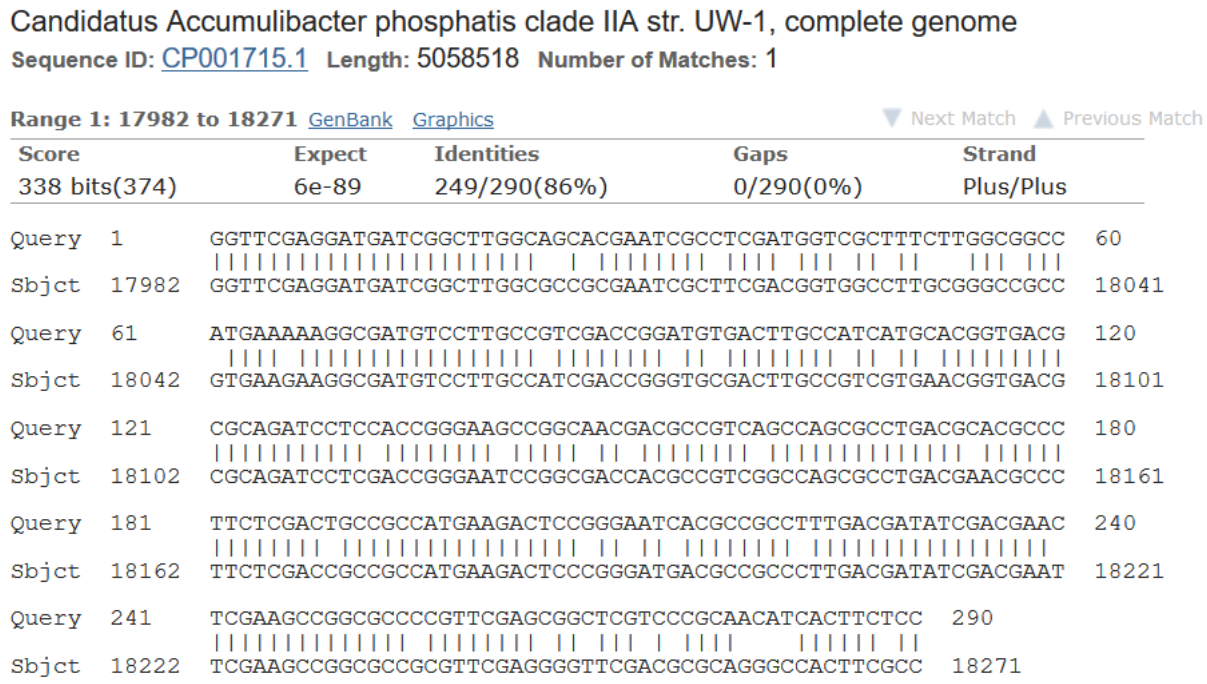
Wenn nun ein String gelesen wird, wird pro Buchstabe im String eine Transition mit dem entsprechenden Label weitergegangen. Wird ein Endzustand bzw. das Ende eines passenden Substrings erreicht so wird das Wort erkannt. Falls nun keine Transitionen mehr verfügbar sind wird wieder an den Anfang des Baumes bzw. des Graphen oder der Automatenstruktur gesprungen. Falls es keine passend gelabelte Transition von einem Knoten zu einem nächsten gibt, wird auch zum Anfang gesprungen (Ausnahme sind wieder die ungelabelten Knoten)

Aufgabe 2



Aufgabe 3

Die Suche der Sequenz auf BLAST ergab, dass Candidatus Accumulibacter phosphatis clade mit 100% Query-cover übereinstimmt. Hier ein Screenshot vom Suchergebnis:



Candidatus Accumulibacter phosphatis clade IIA str. UW-1, complete genome
Sequence ID: CP001715.1 Length: 5058518 Number of Matches: 1

Range 1: 17982 to 18271 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
338 bits(374)	6e-89	249/290(86%)	0/290(0%)	Plus/Plus

```
Query  1      GGTTCGAGGATGATCGGCTTGGCAGCACGAATCGCCTCGATGGTCGCTTTCTTGGCGGCC  60
              ||||||||||||||||||||||  | ||||||| |||| ||| || ||   ||| |||
Sbjct  17982  GGTTCGAGGATGATCGGCTTGGCGCCGCGAATCGCTTCGACGGTGGCCTTGCGGGCCGCC  18041

Query  61     ATGAAAAAGGCGATGTCCTTGCCGTCGACCGGATGTGACTTGCCATCATGCACGGTGACG  120
               ||||  |||||||||||||||||  |||||||| || |||||||| || || |||||||||
Sbjct  18042  GTGAAGAAGGCGATGTCCTTGCCATCGACCGGGTGCGACTTGCCGTCGTGAACGGTGACG  18101

Query  121    CGCAGATCCTCCACCGGGAAGCCGGCAACGACGCCGTCAGCCAGCGCCTGACGCACGCCC  180
              ||||||||||| |||||||| |||||  || ||||||||  ||||||||||||||  ||||||
Sbjct  18102  CGCAGATCCTCGACCGGGAATCCGGCGACCACGCCGTCGGCCAGCGCCTGACGAACGCCC  18161

Query  181    TTCTCGACTGCCGCCATGAAGACTCCGGGAATCACGCCGCCTTTGACGATATCGACGAAC  240
              ||||||||  ||||||||||||||||||  ||  ||  ||||||||  ||||||||||||||||||
Sbjct  18162  TTCTCGACCGCCGCCATGAAGACTCCCGGGATGACGCCGCCCTTGACGATATCGACGAAT  18221

Query  241    TCGAAGCCGGCGCCCCGTTCGAGCGGCTCGTCCCGCAACATCACTTCTCC  290
              |||||||||||||  ||||||||  ||  |||  |  ||||     ||||||  ||
Sbjct  18222  TCGAAGCCGGCGCCGCGTTCGAGGGGTTCGACGCGCAGGGCCACTTCGCC  18271
```

a)

```
>gnl|SRA|SRR172902.1 USI-EAS376:1:1:1:1204
CGGTCGTACTCGCGCTGGGCCNCGCCCAGCGCCAGCCGCGAGTNGATTTCTAACGCCTGC
CGGGANGCATGGCCG
```

This read is part of a microbiome analysis. It was sequenced in the Human Microbiome Project as a mock pilot:

Human Microbiome Project (HMP) Metagenomic WGS Projects, deeper sequencing of the human microbiome samples: Mock Pilot
656207 reads were produced in this project.

b)

```
>WP_ 104992252.1 MULTISPECIES: integrase [Deinococcus] PTA66341.1 integrase
[Deinococcus sp. OD32]
```

This line shows the meta-information of the BLAST match. The DNA-sequence could be aligned to the protein integrase of the genus *Deinococcus*.

```
Length=236
```

The original aminoacid sequence of the integrase is 236 aminoacids long.

```
Score = 44.7 bits (104), Expect = 3e-04, Method: Compositional matrix adjust.
```

The first number is the bit score, which is calculated on the raw score of the alignment, which is written behind in brackets (104). The bit score helps comparing blast results, while they have a close relationship to the E-value, just making it more userfriendly. The E-value, which is called 'Expect' in this case, is the amount of high-similar sequence pairs with a score greater than $S = 104$ (in this case), that could be expected by random chance alignment. The last part determines the method for the score calculation. In this case a modified substitution matrix is used, which is called compositional matrix. This matrix is adjusted in an appropriate way to the compared sequences.

```
Identities = 21/24 (88%), Positives = 22/24 (92%), Gaps = 0/24 (0%)
```

This line summarize exactly how many amino acids were identical in the comparison of the query to the database 21/24. The amount of positives 22/24. These are identical or quiet similar (in a structural meaning) amino acids. Furthermore there can not be found any gap in the alignment.
7

```
Frame = -2
```

The frame is -2. This means, that the DNA sequence was translated from the complement opposite end, like it was sequenced and shifted from the first to the second base.

```
Query 74 GHASRQALEIXSRLALGXAQREYD 3
```

This line shows the query, thus the translated DNA sequence. The switch of the ends and the shift can be noticed by the numbers. From the 74 base to the 3 base.

```
GHASRQ+LEI SRLALG AQREYD
```

These are the identical amino acids. The mismatches are gaps in this line and the plus represent nearly realted aminoacids of the two sequences.

```
Sbjct 205 GHASRQSLEIYSRLALGEAQREYD 228
```

This is the part of the database protein sequence, which was aligned. Its starts at the 205 amino acids and ends at the 228 amino acid.
For that comparison BLASTx was used, because a given DNA sequence was localy aligned to a protein sequence out of a database. Probably the **nr**-Database was used:
All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects