# Grundlagen der Bioinformatik

SoSe 2022

Tutor: Theresa/ Mathias

| 1 | 2 | 3 | 4 | $\sum$ |
|---|---|---|---|---|
| | | | | |

Marina Dittschar & Clarissa Auckenthaler

## Blatt 9

(Abgabe am 07.06.2022)

## Theoretical Assignments

### Task 1: Viterbi Algorithm by hand (2)

| **V** | **b** | **R** | **Y** | **Y** | **e** |
|---|---|---|---|---|---|
| **b** | 1 | 0 | 0 | 0 | 0.005762 |
| **P** | 0 | 0.28 | 0.0378 | 0.057624 | 0 |
| **N** | 0 | 0.09 | 0.1372 | 0.019208 | 0 |

$$
\begin{aligned}
V_N(2) &= e_N(Y) * max(v_k(1) * p_{KN}) \\
&= 0.7 * max(0.09 * 0.2, 0.28 * 0.7) \\
&= 0.7 * max(0.018, 0.196) \\
&= 0.7 * 0.196 \\
&= 0.1372
\end{aligned}
$$

$$
\begin{aligned}
V_P(3) &= e_P(Y) * max(v_k(2) * p_{KP}) \\
&= 0.6 * max(0.00378 * 0.2, 0.1372 * 0.7) \\
&= 0.6 * max(0.00756, 0.09604) \\
&= 0.6 * 0.09604 \\
&= 0.057624
\end{aligned}
$$

$$
\begin{aligned}
V_N(3) &= e_N(Y) * max(v_k(2) * p_{KN}) \\
&= 0.7 * max(0.1372 * 0.2, 0.00378 * 0.7) \\
&= 0.7 * max(0.02744, 0.02646) \\
&= 0.7 * 0.02744 \\
&= 0.019208
\end{aligned}
$$

Termination:

$$P(x, \pi^*) = max_{k \in Q}(v_k(L)p_{ke}) = max(0.057624 * 0.1, 0.019208 * 0.1)$$
$$= max(0.0057624, 0.0019208)$$
$$= 0.0057624$$

Traceback:

| **V** | **b** | **R** | **Y** | **Y** | **e** |
|---|---|---|---|---|---|
| **b** | **1** | 0 | 0 | 0 | **0.005762** |
| **P** | 0 | **0.28** | 0.0378 | **0.057624** | 0 |
| **N** | 0 | 0.09 | **0.1372** | 0.019208 | 0 |

$\pi_3^* = argmax_{k \in Q}(v_k(L)p_{ke}) = Y_P$
$\pi_2^* = ptr_3(\pi_3^*) = ptr_3(Y_P) = Y_N$
$\pi_1^* = ptr_2(\pi_2^*) = ptr_2(Y_N) = R_P$
$\pi_1^* = ptr_1(\pi_1^*) = ptr_1(R_P) = b$
$\implies \pi^* = bR_PY_NY_P$ Final comparison:
$\mathbb{P}(RYY, R_P, Y_N, Y_P) = 0.28 * (0.7 * 0.7 * 0.6 * 0.7 * 0.01) = 0.0057624$

## Task 2: Profile HMMs (6)

$e_{m1}(A) = \frac{5}{10}$ $e_{m1}(T) = \frac{5}{10}$ $e_{m1}(X) = 0$ with X $\neq \{A, T\}$
$e_{m2}(A) = \frac{1}{10}$ $e_{m2}(C) = \frac{4}{10}$ $e_{m2}(G) = \frac{4}{10}$ $e_{m2}(X) = \frac{1}{10}$ with X $\neq \{A, C, G\}$
$e_{m3}(T) = \frac{2}{10}$ $e_{m3}(C) = \frac{2}{10}$ $e_{m3}(X) = \frac{6}{10}$ with X $\neq \{C, T\}$
$e_{m4}(C) = \frac{4}{10}$ $e_{m4}(G) = \frac{4}{10}$ $e_{m4}(T) = \frac{2}{10}$ $e_{m4}(X) = 0$ with X $\neq \{C, G, T\}$
$e_{m5}(C) = \frac{2}{10}$ $e_{m5}(G) = \frac{6}{10}$ $e_{m5}(X) = \frac{2}{10}$ with X $\neq \{C, G\}$

The third column refers to an insert state because more than 50% are gaps.

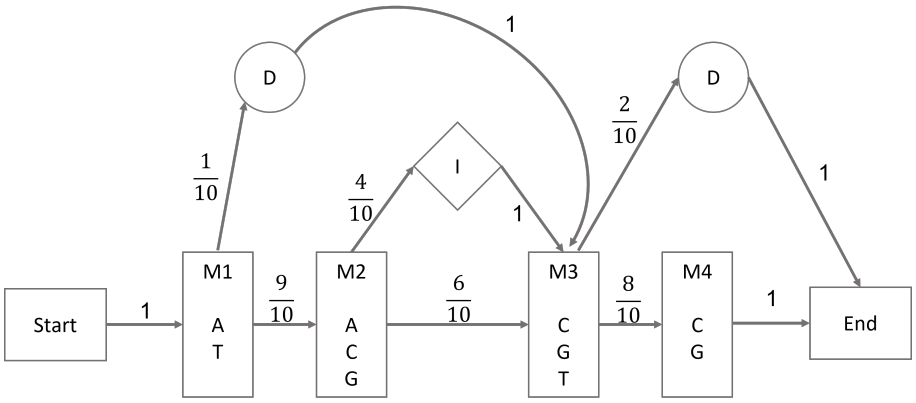


Figure 1: Our state diagram

### Task 3: Supervised training (2)

$$p_{kl} = \frac{P_{kl} + 1}{\sum_{q \in Q} P_{kq} + 1}$$

$$e_k(b) = \frac{E_k(b) + 1}{\sum_{s \in \sum} E_k(s) + 1}$$

## Practical Assignments

### Task 4: Decoding data using the Viterbi algorithm (10)

For Task 4 we used Python 3.8.8, and the following libraries **io, pandas, sys, getopt, Bio, collections and numpy**.

Enter the following code in the command line to run the file:

```
python auckenthaler_dittschar_hmm_handler.py -a cpg.hmm -s input_hmm.fasta
```

We exported the output as a txt-file and saved each Sequence in a Seperate txt-file named:

- auckenthaler_dittschar_viterbi_Seq1.txt
- auckenthaler_dittschar_viterbi_Seq2.txt
- auckenthaler_dittschar_viterbi_Seq3.txt
- auckenthaler_dittschar_viterbi_Seq4.txt