Grundlagen der Bioinformatik

SoSe 2022

Tutor: Theresa/ Mathias

1	2	3	4	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Blatt 8

(Abgabe am 30.06.2022)

Theoretical Assignments

Task 1: Sequencing approaches (in a nutshell) (4)

The Illumina Sequencing workflow consists of four basic steps [2]:

- **Sample prep**: Additional motifs are added to the DNA: Regions complimentary to the flow cell oligose, sequencing binding site and indices [1].
- In the **cluster generation** step, DNA is fixed in an extremely high density on the flow cell while facilitating enzyme access and then amplified using solid-phase amplification [2].
- **Sequencing**: In each cycle, a single labeled deoxynucleoside triphosphate is added to the DNA nucleotide chain and emits a light signal, afterwards being cleaved off [2].
- In **data analysis**, accompanying software enables researchers to peform alignment to a reference, working with extremely accurate data and being able to streamline collection and analysis of data [2].

Task 2: HMM for RNA sequence structure prediction(3)

Emission alphabet:
$\sum = \{\text{A, G, C, U}\}$
Set of states:
$Q = \{\text{Coil, Steam, Loop}\}$
Transitions probabilities:
$p = \{p_{bC}, p_{eC}, p_{CC}, p_{SS}, p_{LL}, p_{SC}, p_{CS}, p_{SL}, p_{LS}\}$
Emission probabilities:
$e = \{e_{CA}, e_{CG}, e_{CC}, e_{CU}, e_{SA}, e_{SG}, e_{SC}, e_{SU}, e_{LA}, e_{LG}, e_{LC}, e_{LU}\}$
The number of total parameters is 21 (12 emission probabilities and 9 transition probabilities).

The HMM graph for our hairpin loop structure looks as follows:

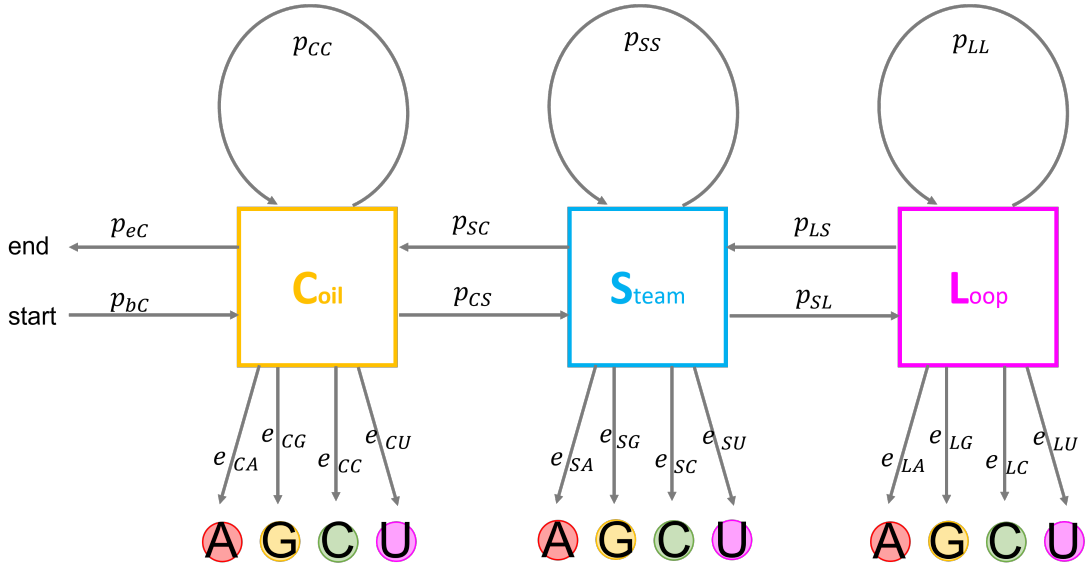

Figure 1: Hidden and Emission states of HMM for the hairpin loop structure.

Task 3: Transition matrix computation by hand (2)

Given the following sequences of exonic regions, compute the transition matrix P_{exonic} by hand for the alphabet $\sum = \{b, e, A, T, C, G\}$

`seq1 = CTTCTTGTGT  seq2 = GTTGGACACTTTCGGG  seq3=TTGCTGTCGTA`

`seq4 = CAGACGTAAGTCG  seq5 = GCCCGTATAGGGC  seq6=CCTGTG`

The number of observed frequencies matrix P_{exonic}=

c_{st}	b	A	G	C	T	e
b	0	0	2	3	1	0
A	0	1	3	3	1	1
G	0	2	5	3	9	3
C	0	2	5	3	5	1
T	0	4	7	3	6	1
e	0	0	0	0	0	1

From observed frequencies to probabilities in the final transition matrix P_{exonic} =

c_{st}	b	A	G	C	T	e
b	0	0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$	0
A	0	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{9}$
G	0	$\frac{1}{11}$	$\frac{5}{22}$	$\frac{3}{22}$	$\frac{9}{22}$	$\frac{3}{22}$
C	0	$\frac{1}{8}$	$\frac{5}{16}$	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{1}{16}$
T	0	$\frac{4}{21}$	$\frac{1}{3}$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{1}{21}$
e	0	0	0	0	0	1

Practical Assignments

Task 4: Transition matrix and log-odds computation(11)

For Task 4 we used Python 3.8.8, and the following libraries **sys, getopt, Bio and numpy**. Enter the following code in the command line to run the file **auckenthaler_dittschar_train_hmm.py**:

```
python auckenthaler_dittschar_train_hmm.py -f "cds_set.fasta" -n "cds_p_matrix
python auckenthaler_dittschar_train_hmm.py -f "notcds_set.fasta" -n "notcds_p_matrix
```

With the parameters:
-f : input file
-n : name for output transition matrix.

This code outputs a .txt file with the transition matrix and the header as specified in the script.
Enter the following code in the command line to run the file **auckenthaler_dittschar_test_hmm.py**:

```
python auckenthaler_dittschar_test_hmm.py -a "auckenthaler_dittschar_cds_p_matrix.txt"
-b "auckenthaler_dittschar_notcds_p_matrix.txt" -i "contig.fasta"
```

With the parameters:
-a : transition matrix for the plus model
-b : transition matrix for the minus model
- i : input file with sequence to determine probability for
The log-odds ratio is 10.28. It is positive, therefore it is more likely that the sequence in question belongs to protein-coding regions.

References

[1] Illumina. Illumina sequencing by synthesis. `https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=3s`, 2016.

[2] Inc. Illumina. Illumina sequencing technology. https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf, 2010.