JTMS-19 Introduction to Bioinformatics

Prof. Dr. Marc-Thorsten Hütt
Prof. Dr. Felix Jonas

FINAL EXAM – PREVIEW 3

December 4, 2024

Note: This preview is approximately 50 percent the size of the actual final; thus you should need below 60 minutes.

Your name: _____

(1) Write down the joint probability $P(x, \pi)$ of a sequence x and an internal path π in a Hidden Markov Model (HMM) in terms of the parameters of the HMM.

[4 points]

(2) Describe in 2–3 sentences what a genome browser is and give examples of information accessible through it.

[3 points]

(3) Which preprocessing of a pattern sequence is required in order to run the Boyer-Moore algorithm?

[2 points]

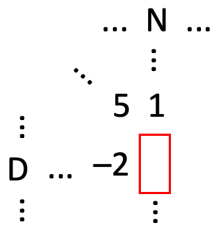
(4) How do you define distances between clusters in UPGMA? How do these distances transform, when clusters merge during the iterative clustering procedure?

[4 points]

(5) Let $G(V, E)$ be a graph. Explain the notation: What does V and E stand for?

[2 points]

(6) In the Needleman-Wunsch algorithm, you encounter the situation depicted below.



What is the next F-matrix entry and what is the corresponding entry of the pointer variable? Additional information: scoring matrix entry for `N` and `D`: $+2$, gap penalty: -2.

[3 points]

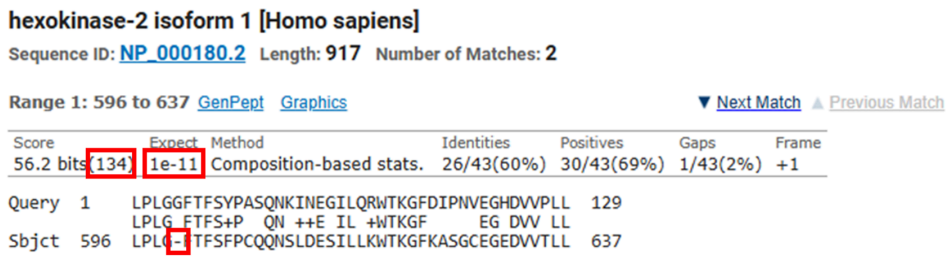
(7) Given a result (e.g., an alignment score) and a set of corresponding random observations derived from a null model (e.g., shuffled sequences), how are the z-score and the p-value defined?

[4 points]

(8) The word length and the score threshold are two important parameters of BLAST. Describe how they work.

[2 points]

(9) Below you can find a screenshot of the output of a BLAST search:



hexokinase-2 isoform 1 [Homo sapiens]

Sequence ID: NP_000180.2 Length: 917 Number of Matches: 2

Range 1: 596 to 637 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
56.2 bits(134)	1e-11	Composition-based stats.	26/43(60%)	30/43(69%)	1/43(2%)	+1

```
Query  1    LPLGGFTFSYPASQNKINEGILQRWTKGFDIPNVEGHDVVPLL  129
            LPLG FTFS+P  QN ++E IL +WTKGF      EG DVV LL
Sbjct  596  LPLG-FTFSFPCQQNSLDESILLKWTKGFKASGCEGEDVVTLL  637
```

(A) Identify the used blast flavour and explain your choice.

(B) Define the meaning of the ”–” in the bottom line of the alignment

(C) Identify the maximum ”Word size” that could be used to identify this alignment.

(D) Outline how the raw score and expected value (both highlighted in red) would change if (1) a different scoring matrix was used, and (2) a larger search database would be used.

[5 points]

(10) Below you can find the cumulative number of occurrences of hydrophobic and hydrophilic in the transmembrane domains and soluble domain of many transmembrane proteins:

	hydrophobic	hydrophilic
transmembrane	1000	950
soluble	1000	4000

(A) Draw the structure of a simple hidden Markov model that could describe these proteins.

(B) Calculate and indicate the emission probabilities on your model.

(C) Outline how the average size of a transmembrane / soluble domain could be reflected in your model.

(D) Aquaporins are characterized by transmembrane alpha helix bundles with an hydrophilic inside (to allow for the passage of water molecules). Explain if you think those helixes would be identified as transmembrane domain proteins.

[5 points]