JTMS-19 Introduction to Bioinformatics

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

FINAL EXAM – PREVIEW 2

November 27, 2024

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

Your name: ______________________________

(1) Assume you have been given the joint probability $P(X, Y)$ of two events X and Y. What is the marginal probability of X and how do you obtain it?

[3 points]

$$P(x) = \sum_{y} P(x, y)$$

all possible states of variable (or 'event') y

(2) What is the score of the alignment below?

```
CGATC-CTGT
C-ATCGCCTT
| · | | | · | x x |
```

(a) in the following scoring model:
match $+1$, mismatch -1, gap-opening -3, ~~gap-extension -1~~.

(b) in the following scoring model:
match $+1$, mismatch 0, gap-opening 0, ~~gap-extension~~ 0.

[2 points]

(a) $+1 - 3 + 1 + 1 + 1 - 3 + 1 - 1 - 1 + 1 = -2$

(b) 6

(3) Markov models are an important class of probability models in bioinformatics. What are the model parameters in a Markov model? For which application did we use Markov models in class?

[3 points]

- transition probabilities
- application: CpG islands

[+ transitions from some initial ('zero') state optional!!]

(4) Give three examples of data types collected in the ENCODE project.

[3 points]

- RNA sequencing data
- chromatin modifications
- histone modifications
- various types of regulators

(5) How are the matrix elements F_{ij} defined in the Needleman-Wunsch algorithm for the optimal global alignment of two sequences? Explain in detail, how the entries of the substitution matrix and the gap penalty appear in the definition. ✓

[4 points]

$$F_{ij} = \max \begin{cases} F_{i-1,j-1} + s(x_i, y_j) \\ F_{i,j-1} - d \\ F_{i-1,j} - d \end{cases}$$

symbol with symbol ⇒ substitution matrix entry

symbol with gap ⇒ gap penalty

gap with symbol ⇒ gap penalty

(6) How do you define the entropy of a symbol sequence? Give a bioinformatics example, where the entropy can be helpful.

[2 points]

$X = x_1 x_2 \cdots x_n$, $x_i \in \Sigma$ (symbol space)

→ $p_a = \frac{\# a \text{ in } x}{n}$ relative frequency (≈ probability) of symbol $a \in \Sigma$ in x

$$H = -\sum_{a \in \Sigma} p_a \log p_a$$

- runtime of exact text matching
- data compression
- multiple sequence alignment

(7) What is a progressive multiple sequence alignment?

[2 points]

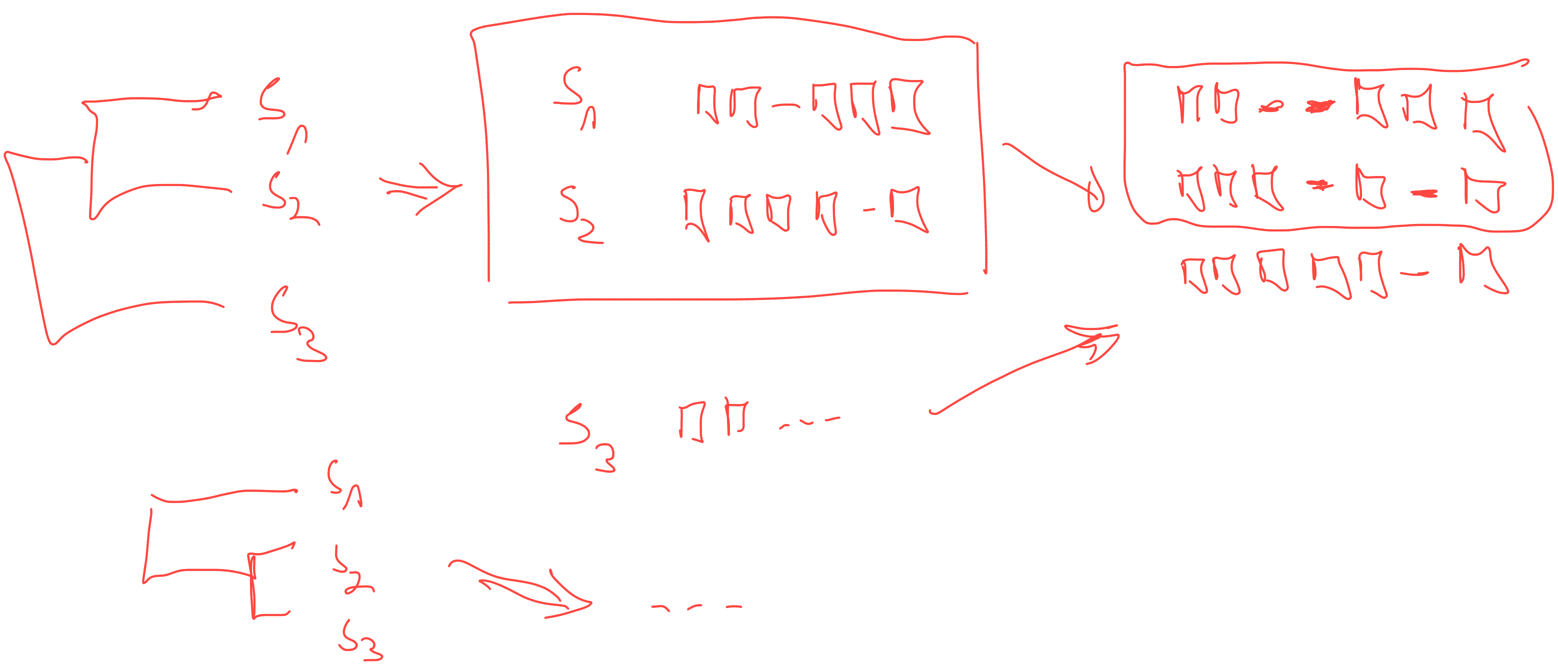
all pairwise alignments of sequences $S_1 \dots S_k$

→ score matrix ($k \times k$)

→ distance matrix ($k \times k$)

UPGMA → guide tree (clustering tree)

→ progressively construct the multiple sequence alignment following the guide tree



(8) Write down the recursion relation for the backward variable in the posterior decoding of Hidden Markov models. What do k and i stand for in the backward variable $b_k(i)$? What does $b_k(i)$ mean in terms of probabilities (i.e. give a representation of $b_k(i)$ as a conditional probability)?

[6 points]

$x = x_1 \dots x_L$ observed sequence

transition p.

emission p.

$$b_k(i) = \sum_{l \in \Sigma_{HMM}} a_{kl} \, e_l(x_{i+1}) \, b_l(i+1)$$

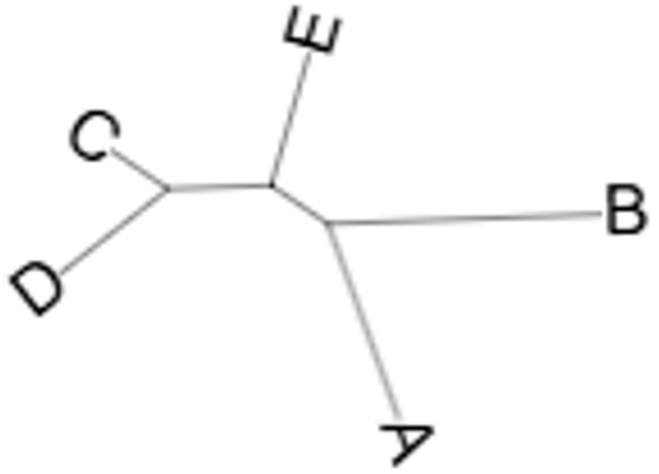
i : position in sequence x

k = HMM state

$$P(x_{i+1} x_{i+2} \dots x_L \mid \pi_i = k)$$

internal (hidden) path

(9) Given the following unrooted tree:



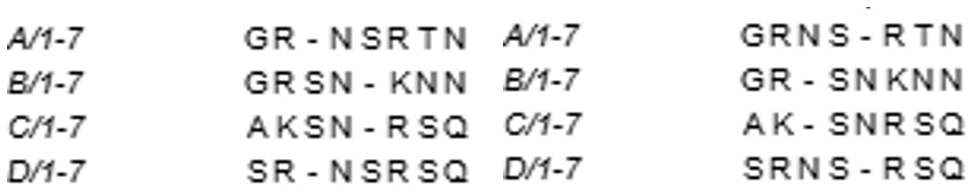
(A) Which algorithm was used to generate that tree? Please explain your answer.

(B) Draw the re-rooted tree using C as an Outgroup.

[4 points]

Prof. Jonas

(10) From biochemical experiments you know that a serine is critical for the function of a protein. Using two different MSA tools you generate two different multiple sequence alignments for five orthologs:



```
A/1-7    GR-NSRTN    A/1-7    GRNS-RTN
B/1-7    GRSN-KNN    B/1-7    GR-SNKNN
C/1-7    AKSN-RSQ    C/1-7    AK-SNRSQ
D/1-7    SR-NSRSQ    D/1-7    SRNS-RSQ
```

(A) Which of the two alignments do you think is more meaningful? Explain your answer in one sentence.

(B) Which other property of this sequence region do you think is functionally important? Explain your answer in one sentence.

[4 points]

Prof. Jonas