**Student name _____**

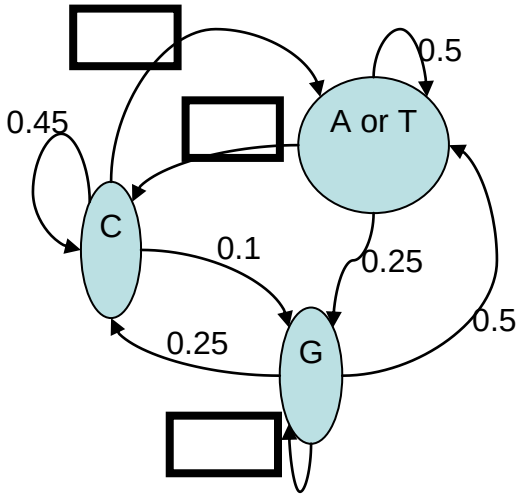
**Student number_____**

# CSc 428/589B
# Algorithms in bioinformatics

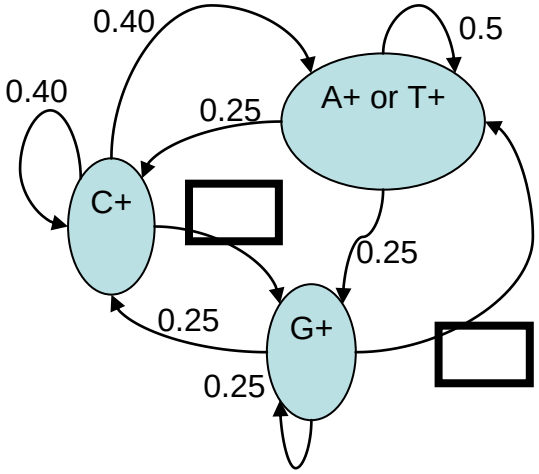
## Midterm exam 2

| | Max points | Results |
|---|---|---|
| **Question 1** | **10** | |
| **Question 2** | **10** | |
| **Question 3** | **10** | |
| **Question 4** | **10** | |
| **Question 5** | **10** | |
| **Question 6** | **10** | |
| **Question 7** | **10** | |
| **Question 8** | **10** | |
| **Question 9** | **10** | |
| **Question 10** | **10** | |
| **Total:** | **100** | |

**Question 1.**
Fill in the missing probabilities for the following Markov models




Normal DNA




CpG island

**Question 2.**
Use these models in order to answer whether or not the following DNA sequence is from a CpG island.

Assume that the probability of getting C in the first position equals 1 (no begin state)

CCGCACG

**Question 3.**
Estimate the set of emission and transition probabilities from the following labelled sequence of symbols and associated states for a fair (F) and a loaded (L) 6-sided die:

| Symbols | 2 | 6 | 6 | 3 | 4 | 6 | 5 | 6 | 6 |
|---|---|---|---|---|---|---|---|---|---|
| States | F | F | F | F | F | L | L | L | L |

eF(1)= eF(4)= eL(1)= eL(4)=

eF(2)= eF(5)= eL(2)= eL(5)=

eF(3)= eF(6)= eL(3)= eL(6)=

---

aFF= aFL= aLF= aLL=

**Question 4.**
Which sequence out of the following sequences will be chosen as a center sequence by the SP-star algorithm for multiple sequence alignment?

| |
|---|
| S1=ACG |
| S2=AGT |
| S3=ACGT |
| S4=ACT |

## Question 5.

Produce the matrix of edit distances for each pair of the following strings. This distance matrix will be used for question 6.

| | |
|---|---|
| English | father |
| Spanish | padre |
| Italian | padre |
| German | vater |
| Portuguese | pai |

| | E | S | I | G | P |
|---|---|---|---|---|---|
| E | 0 | | | | |
| S | | 0 | | | |
| I | | | 0 | | |
| G | | | | 0 | |
| P | | | | | 0 |

**Question 6**. Build a phylogenetic tree of languages using the UPGMA algorithm for the distance matrix from question 5. You are NOT required to label the length of all tree edges. Give only a tree topology and an updated distance matrix for each clustering step.

**Question 7.**
You are given 2 binary matrices representing 5 objects (rows) in terms of 5 characters or traits (columns).
**Part 1.** Which of the two has a perfect phylogenetic tree?

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 1 | 1 | 0 | 0 | 0 |
| B | 0 | 0 | 1 | 0 | 0 |
| C | 1 | 1 | 1 | 0 | 1 |
| D | 0 | 0 | 1 | 1 | 0 |
| E | 0 | 1 | 0 | 0 | 0 |

A

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 0 | 0 | 1 | 0 | 1 |
| B | 1 | 0 | 0 | 1 | 0 |
| C | 0 | 0 | 0 | 0 | 1 |
| D | 1 | 0 | 0 | 0 | 0 |
| E | 0 | 1 | 1 | 0 | 1 |

B

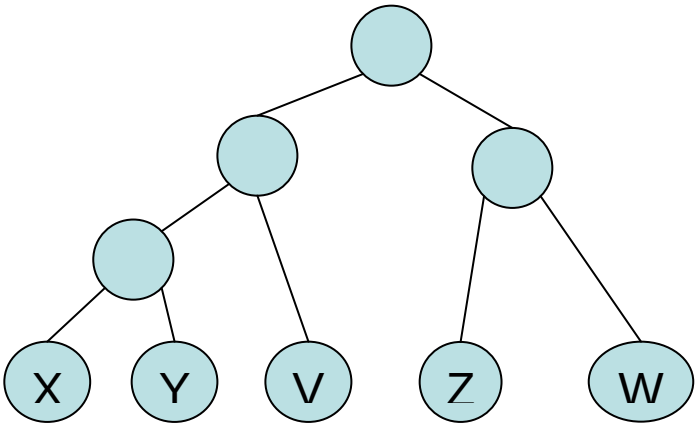
**Part 2.** Build this tree

**Question 8.**
Find the most parsimonious labeling of internal nodes of the following tree based on a given multiple alignment, using the Fitch algorithm.

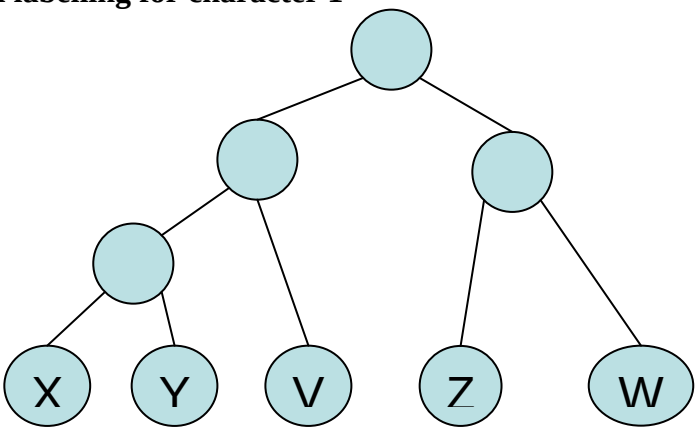
| | |
|---|---|
| **X** | **AAG** |
| **Y** | **ACG** |
| **Z** | **CGA** |
| **V** | **CCG** |
| **W** | **AGA** |

## For character 1:

**Candidate sets**

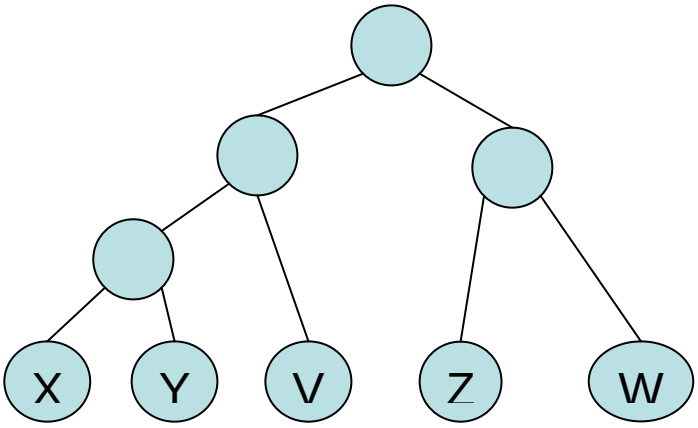



**Final labelling for character 1**

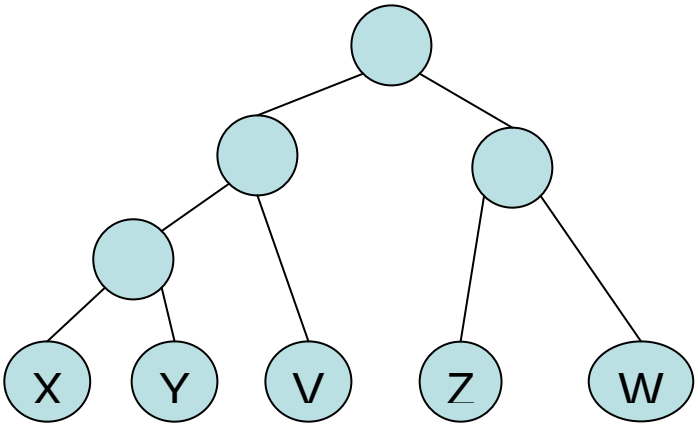


## For character 2

**Candidate sets**

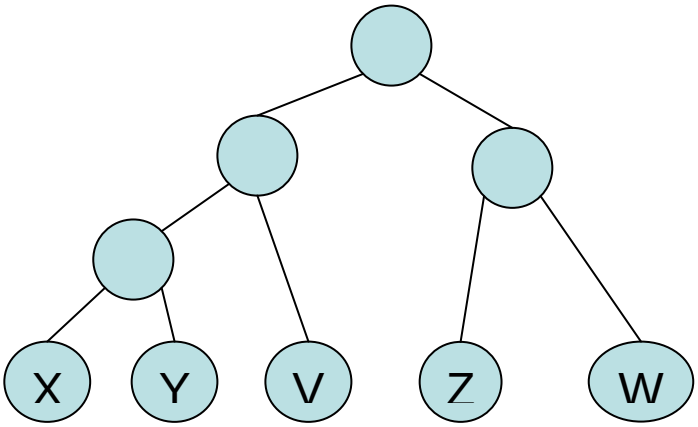



**Final labelling for character 2**

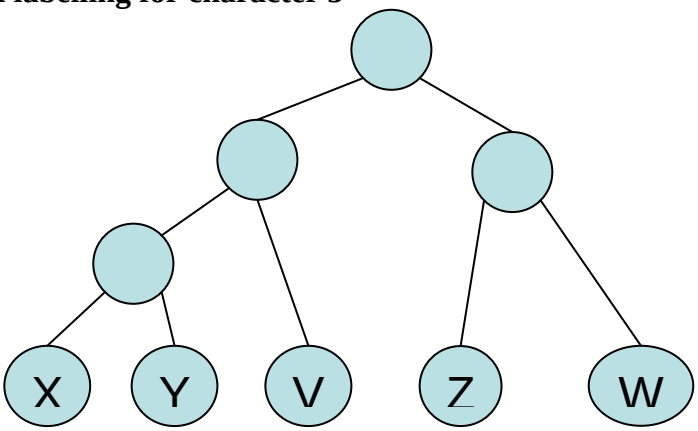


## For character 3:

**Candidate sets**

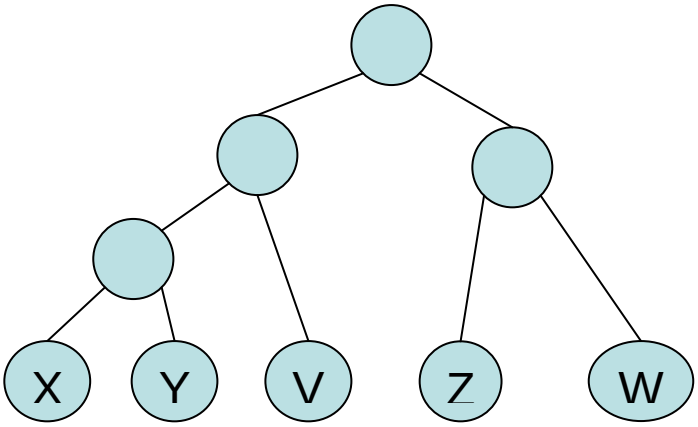



**Final labelling for character 3**



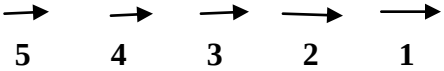

---

## The final labelling for all 3 characters:




The parsimony score of this tree:
**L(T)=**

**Question 9.**




For the signed permutation shown above answer the following questions

**Part 1.**

A. How many breakpoints does the permutation have?
B. Draw the reality and desire diagram.

|

**Part 2.**

How many reversals are required in order to sort this permutation? Justify your answer

**Question 10.** Describe the Branch and Bound optimization technique and explain how it is applied for solving the large parsimony problem. In your description, give an explicit answer to the following questions:

1. Does the B & B technique guarantee to find the optimal solution?
2. What is the main property of the search space which allows us to use the B & B technique?
3. Does B & B always (i.e. in the worst case) run in polynomial time?