

BIIN 200: Bioinformatics I

Dr. Craig Struble

Fall 2008, Final Exam

closed book, closed notes, two US Letter cheat sheets, calculators OK

100

8 pages

Name: _____

Instructions: Answer each question in the space provided. Do not use more space than provided. Write neatly. You may use the back of your test pages for scratch paper.

Section 1. **Phylogenetic Analysis**

1. (5 points) Using the 4 sequences,

GTTGG
GTTGA
ATTGG
ATTGA

briefly explain how the *maximum parsimony* algorithm constructs a phylogenetic tree. (You do not need to actually construct the tree. Just explain the steps needed to construct the tree.)

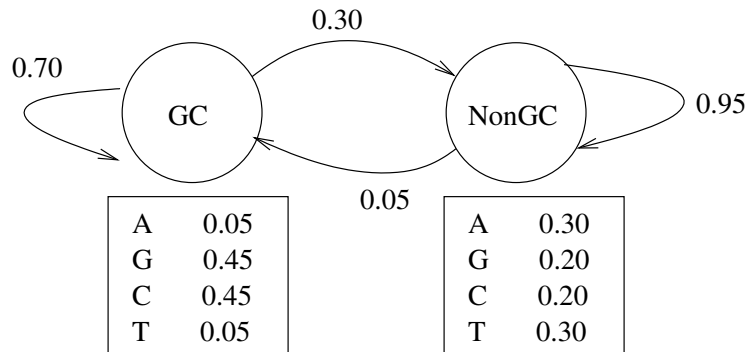
2. (5 points) Three models of evolution were discussed in class and the textbook: Jukes-Cantor, Kimura two-parameter, and HKY85. Which of these models distinguish between mutation types, but does not take into account base composition?

3. (5 points) Name one phylogenetic analysis algorithm that creates *ultrameric trees*.

Section 3. Hidden Markov Modeling

6. (10 points) Draw the HMM state diagram for two positions, u and $u + 1$ in a profile hidden Markov model.

7. (10 points) Use the HMM diagram below, and assume that the starting and stopping probabilities are unimportant (i.e., just assume they are 1 for calculation purposes). Calculate the probability of the DNA sequence $GCGTGA$ if the state sequence is (a) NonGC-GC-GC-GC-GC-NonGC and (b) NonGC-NonGC-NonGC-NonGC-NonGC-NonGC.



Section 4. Python and BioPython

8. (5 points) In a single simple sentence, explain what the following Python code is doing.

```
def foo(data, abet):
    n = len(data)
    m = len(data[0])
    c = zeros((m, len(abet)))
    for i in range(n):
        for j in range(len(data[i])):
            x = abet.index(data[i][j])
            c[j,x] = c[j,x]+1
    c = c / n
    return c

f = open("data.txt")
data = f.readlines()
f.close()
c = foo(data, "ACGT")
print c
```

9. (10 points) Write a program that reads a file containing FASTA formatted DNA sequences and prints out the name of each sequence that contains the motif GCGTG.

Section 6. Multiple Sequence Alignment

14. (10 points) Given the position specific scoring matrix PSSM containing log odds scores below, find the starting position and the ODDS score of the best match in the sequence ATGTGCTGC. Do scratch work on the back of the previous page.

Position	A	C	G	T
1	0.26	-0.32	0.26	-0.32
2	-1.32	-1.32	-1.32	1.49
3	-1.32	-1.32	1.49	-1.32
4	-1.32	1.0	0.26	-1.32

15. (10 points) Using the BLOSUM62 scoring matrix at the end of this test and a gap penalty of -5, what is the multiple sequence alignment score for the alignment

QSVIVK
QSCIVK
QGCLVK
QGVL-K

BLOSUM62 Scoring Matrix

	a	r	n	d	c	q	e	g	h	i	l	k	m	f	p	s	t	w	y	v
a	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-2	-1	1	0	-3	-2	0	
r	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
n	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
d	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
c	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
e	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
g	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
h	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
i	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
l	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
k	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
m	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
f	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
p	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
s	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
t	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
w	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
v	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4