3: Project Write-up

Jacob Friedberg
frie2142@vandals.uidaho.edu

Garrett Wells
well1157@vandals.uidaho.edu

April 28, 2023

Abstract

This project used techniques presented in the course to estimate parameters for a Hidden Markov Model of the provided data set. We prepared a substitution matrix, emission probability table, and state transition table from the data. The methods employed here produced good results which we are moderately confident in.

1 Introduction ADD SOME STUFF HERE GARRETT

Something about substitution matrices and what they are used for. HMM and the definitions for emission and transition. In this paper we implement two algorithms: Generating a BLOSUM-like substitution matrix using a provided set of 200 sequences and calculating the emission and transition probabilities given a set of sequences that are annotated to indicate the locations where specific amino acids within the gene contribute to different characteristic of the *Silacus Soulas* insect.

2 Algorithm Descriptions

2.1 Substitution Matrix

2.2 Emission and Transition Probabilities

Emission and transition probabilities may be described as the following:

Emission a symbol occur(emitted) in a genome while in a specific state.

Transition a change from one state to another, representing a change in the purpose or type of the sequence.

In this case, a symbol refers to one of the 20 amino acids, represented by its single letter substitute. Examples can be seen in emission table 3. In the emission parameters are calculated individually for each symbol based on state. The equation used for these calculations is listed below. The term, $e_{s_x symbol_y}$, represents the number of occurrences of $symbol_y$ in state s_x. This term is divided by all occurrences of any symbol in state s_x to produce a probability of seeing $symbol_y$ in s_x.

$$\frac{e_{s_x symbol_y}}{\sum_{k=1}^{n} e_{s_x symbol_k}} \tag{1}$$

Transition parameters are described as the probability of "transitioning from state `X` to state `Y`"(see table 4). They are calculated as the number of observed transitions between two states,

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

Table 1: Generated BLOSUM-like scoring matrix

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

Table 2: Generated BLOSUM-like scoring matrix

s_x and s_y, divided by all observed transitions out of state s_x. Transitions between the symbols in the same state are counted to estimate the probability of remaining in the current state. The equation for this calculation is more formally stated below with t representing observed transitions between two states.

$$\frac{t_{s_x s_y}}{\sum_{k=0}^{n} t_{s_x s_k}} \tag{2}$$

As may be expected, all probabilities sum to 1, though in slightly different ways for each table. For the emission table, all probabilities for symbols seen in state 0(seen as one column of the table), for example will sum to 1. For the transition table, however, one horizontal row should be expected to sum to 1.

3 Results

Table 3: Emission By State

Amino Acid	**State**		
-	0	1	2
a	3.91	7.94	1.59
c	3.84	2.75	6.86
d	1.87	1.72	4.18
e	3.91	2.62	5.62
f	6.60	4.96	5.55
g	3.73	1.85	4.86
h	3.65	1.77	3.63
i	2.91	2.13	5.01
k	3.29	2.40	4.96
l	3.48	5.26	4.22
m	4.67	4.28	4.79
n	5.58	2.35	5.41
p	5.84	3.71	5.75
q	5.44	2.18	5.38
r	2.82	1.75	5.18
s	5.77	4.17	4.90
t	5.96	11.10	5.75
v	11.35	14.54	5.38
w	7.43	11.86	6.43
y	7.97	10.66	4.56

Table 4: Transitions Between States

From State	**To State**		
-	0	1	2
0	96.1	3.89	0.00
1	0.00	92.01	7.99
2	1.54	0.00	98.5