CS-415
Assignment 2

Colby Blair

March 30th, 2011

1

a

e_{fish} for the next 3 symbols:

$$e_{fish_1,fish_2,fish_3} = .1 * .1 * .1 = .001$$

b

The 3x3 matrix is as follows:

	Red	Blue	Fish
Red	.8	.1	.1
Blue	.5	.1	.4
Fish	.8	.1	.1

c

Using the R language, we can see a quick convergence:

```
> m <- matrix(c(.8,.1,.1,.5,.1,.4,.8,.1,.1),nrow=3)
>
> m1 <- m %*% m #matrix multiply m times itself
> print(m1)
     [,1] [,2] [,3]
[1,] 0.77 0.77 0.77
[2,] 0.10 0.10 0.10
[3,] 0.13 0.13 0.13
>
> m2 <- m1 %*% m1
> print(m2)
     [,1] [,2] [,3]
[1,] 0.77 0.77 0.77
[2,] 0.10 0.10 0.10
[3,] 0.13 0.13 0.13
```

d

$$\begin{aligned} P(x) &= .77 * .1 * .1 * .4 \\ &= .00308 \\ &= .308\% \\ &\exists x = \{Red, Fish, Blue, Fish\} \text{ (respectively)} \end{aligned}$$

2

a

Red and Blue can never be emitted.

b

This is a bit of long hand, but is nice for error checking:

Step 1:

$$
\begin{aligned}
P(x) \quad &= \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
&+\{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
&+\{P(Red)\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
&+\{P(Red)\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
&+\{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
&+\{P(Red)\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
&+\{P(Red)\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
&+\{P(Red)\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
&+\{P(Red)\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\}
\end{aligned}
$$

Step 2:

$$
\begin{aligned}
&= \{.1\} * \{.8 * .1\} * \{.8 * .1\} \\
&+\{.1\} * \{.8 * .1\} * \{.1 * .9\} \\
&+\{.1\} * \{.1 * .9\} * \{.5 * .1\} \\
&+\{.1\} * \{.1 * .9\} * \{.1 * .9\} \\
&+\{.1\} * \{.8 * .1\} * \{.5 * .1\} \\
&+\{.1\} * \{.5 * .1\} * \{.8 * .1\} \\
&+\{.1\} * \{.5 * .1\} * \{.5 * .1\} \\
&+\{.1\} * \{.1 * .9\} * \{.5 * .4\} \\
&+\{.1\} * \{.5 * .1\} * \{.1 * .9\}
\end{aligned}
$$

Step 3:

$$
\begin{aligned}
&= .00064 \\
&+.00072 \\
&+.00045 \\
&+.00081 \\
&+.0004 \\
&+.0004 \\
&+.00025 \\
&+.0018 \\
&+.00045 \\
\\
&= .00592 \\
&= .592\%
\end{aligned}
$$

c

This is even longer hand, but the only way I know how to do it:

$$P_{Red|Cat} = P(Cat|Red)P(Fish)/P(Cat) = .1 * .33/.495 = .06$$
$$P_{Blue|Cat} = P(Cat|Blue)P(Fish)/P(Cat) = .9 * .33/.495 = .60$$
$$P_{Fish|Cat} = P(Cat|Fish)P(Fish)/P(Cat) = .5 * .33/.495 = .33$$

Step 1:

$$
\begin{aligned}
P(x) \quad & = \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Red)\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Blue)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Red) * e_{Red}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Red) * e_{Red}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Blue) * e_{Blue}(Cat)\} * \{P(Fish) * e_{Fish}(Cat)\} \\
& + \{P(Fish)\} * \{P(Red * e_{Red}(Cat))\} * \{P(Fish) * e_{Fish}(Cat)\} * \{P(Blue) * e_{Blue}(Cat)\}
\end{aligned}
$$

Step 2:

$$
\begin{aligned}
& = \{.06\} * \{.8 * .1\} * \{.8 * .1\} * \{.8 * .1\} \\
& + \{.06\} * \{.8 * .1\} * \{.8 * .1\} * \{.1 * .9\} \\
& + \{.06\} * \{.8 * .1\} * \{.1 * .9\} * \{.5 * .1\} \\
& + \{.06\} * \{.8 * .1\} * \{.1 * .9\} * \{.1 * .9\} \\
& + \{.06\} * \{.8 * .1\} * \{.8 * .1\} * \{.5 * .1\} \\
& + \{.06\} * \{.8 * .1\} * \{.5 * .1\} * \{.8 * .1\} \\
& + \{.06\} * \{.8 * .1\} * \{.5 * .1\} * \{.5 * .1\} \\
& + \{.06\} * \{.8 * .1\} * \{.1 * .9\} * \{.5 * .4\} \\
& + \{.06\} * \{.8 * .1\} * \{.5 * .1\} * \{.1 * .9\} \\
& = \{.60\} * \{.8 * .1\} * \{.8 * .1\} * \{.8 * .1\} \\
& + \{.60\} * \{.8 * .1\} * \{.8 * .1\} * \{.1 * .9\} \\
& + \{.60\} * \{.8 * .1\} * \{.1 * .9\} * \{.5 * .1\} \\
& + \{.60\} * \{.8 * .1\} * \{.1 * .9\} * \{.1 * .9\} \\
& + \{.60\} * \{.8 * .1\} * \{.8 * .1\} * \{.5 * .1\} \\
& + \{.60\} * \{.8 * .1\} * \{.5 * .1\} * \{.8 * .1\} \\
& + \{.60\} * \{.8 * .1\} * \{.5 * .1\} * \{.5 * .1\} \\
& + \{.60\} * \{.8 * .1\} * \{.1 * .9\} * \{.5 * .4\} \\
& + \{.60\} * \{.8 * .1\} * \{.5 * .1\} * \{.1 * .9\} \\
& = \{.33\} * \{.8 * .1\} * \{.8 * .1\} * \{.8 * .1\} \\
& + \{.33\} * \{.8 * .1\} * \{.8 * .1\} * \{.1 * .9\} \\
& + \{.33\} * \{.8 * .1\} * \{.1 * .9\} * \{.5 * .1\} \\
& + \{.33\} * \{.8 * .1\} * \{.1 * .9\} * \{.1 * .9\} \\
& + \{.33\} * \{.8 * .1\} * \{.8 * .1\} * \{.5 * .1\} \\
& + \{.33\} * \{.8 * .1\} * \{.5 * .1\} * \{.8 * .1\} \\
& + \{.33\} * \{.8 * .1\} * \{.5 * .1\} * \{.5 * .1\} \\
& + \{.33\} * \{.8 * .1\} * \{.1 * .9\} * \{.5 * .4\} \\
& + \{.33\} * \{.8 * .1\} * \{.5 * .1\} * \{.1 * .9\}
\end{aligned}
$$

Step 3:

= .00064
+.00072
+.00045
+.00081
+.0004
+.0004
+.00025
+.0018
+.00045
+.00729
+.00648
+.00648
+.00576
+.0162
+.009
+.0036
+.0144
+.06921
+.00125
+.00125
+.002
+.002
+.0032
+.0025
+.009
+.00405
+.0036
+.0036
+.0312

= .0433
= 4.33%

d

To find the most likely path, just select the best probability from one of the additions in b. Step 2. We would use the Vertirbi Algorithm to do this. In that case, we would select the path .0018 ($\{P(Red)\}*\{P(Blue)*e_{Blue}(Cat)\}*\{P(Fish)*e_{Fish}(Cat)\}$), or Red Blue Fish . The matrix doesn't change at all, but the algorithm changes.

e

You select the second position,use the Forward Algorithm before that postion, and use the Backward Algorithm after it.

3

	Red	Blue	Fish
Red	.11	.16	.16
Blue	.16	.05	.16
Fish	.16	.16	.05

$e_{Red}(Cat) = .57$ $e_{Red}(Dog) = .42$	$e_{Blue}(Cat) = .25$ $e_{Blue}(Dog) = .75$	$e_{Fish}(Cat) = .6$ $e_{Fish}(Dog) = .4$

4

a

Sequence	Family	Desc	Entry Type	Bit score	E-value
seq	PsaA PsaB	Photosystem I psaA/psaB protein	Family	242.4	5.9e-72
seq50	PsaA PsaB	Photosystem I psaA/psaB protein	Family	107.7	3.5e-31
seq50qrs	PsaA PsaB	Photosystem I psaA/psaB protein	Family	119.4	1e-34
seq30	PsaA PsaB	Photosystem I psaA/psaB protein	Family	154.1	3.2e-45
seq30qrs	PsaA PsaB	Photosystem I psaA/psaB protein	Family	155.0	1.7e-45

All of these proteins belong to the PsaA/PsaB family. The original E-value of the protein is very small, meaning that it is very unique. As we increase decomposition from residues within the protein, it becomes much more generic, with much higher E-values. The 50% decomposition decomposes the protein much more than 30%, no suprise. This results in a much more generic protein sequence, however, with the 50% decomposition, as their E-values are much higher than the 30% decomps.

b

seq.txt NCBI BLAST results:

Accession	Description	Max score	Total score	Query coverage	E value
NP_876063.1	P700 chlorophyll a apoprotein A1	487	487	100%	$4e-136$
CAB64198.2	**subunit PsaA**	**486**	**486**	**100%**	$7e-136$
YP_001551524.1	P700 chlorophyll a apoprotein A1	475	475	100%	$2e-132$
YP_001015777.1	P700 chlorophyll a apoprotein A1	380	380	100%	$7e-104$

Our original PFAM results:

Sequence	**Family**	**Desc**	**Entry Type**	**Bit score**	**E-value**
seq	PsaA PsaB	Photosystem I psaA/psaB protein	Family	242.4	5.9e-72

The second result from the top in our BLAST shows us results for the PasA family. The score in BLAST is much higher, as the E-value is much lower. The BLAST algorithm expects much the sequence to be much less common.

BLAST and PFAM scores are different, perhaps because NCBI uses standard matching algorithms, and PFAM emphasises more on frequency of reoccuring groups in HMMs, what they call clans. The latter is a bigger grouping of sub-sequence proteins, so they of course have a lower score.