Genomics and Bioinformatics

Exam solutions

October 31, 2011
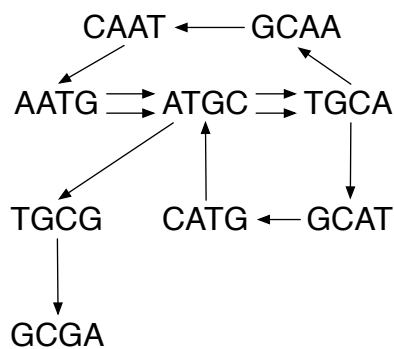
Question 1

1. AATGCATGCAATGCGA

2.

$$S_4 = \{\text{AATG, ATGC, TGCA, GCAT, CATG, GCAA, CAAT, TGCG, GCGA}\}.$$

$$S_5 = \{\text{AATGC, ATGCA, TGCAT,}$$
$$\text{GCATG, CATGC, ATGCA,}$$
$$\text{TGCAA, GCAAT, CAATG,}$$
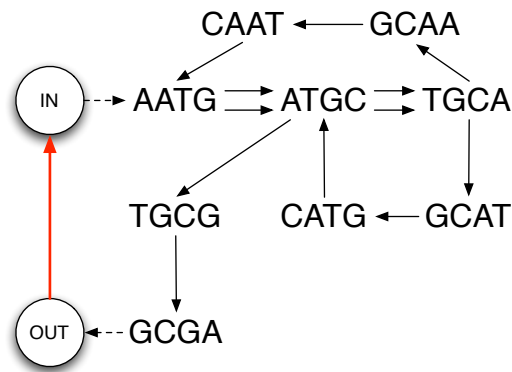$$\text{AATGC, ATGCG, TGCGA}\}.$$

3.

4. Possible paths are

AATG-ATGC-TGCA-GCAT-CATG-ATGC-TGCA-GCAA-CAAT-AATG-ATGC-TGCG-GCGA
AATG-ATGC-TGCA-GCAA-CAAT-AATG-ATGC-TGCA-GCAT-CATG-ATGC-TGCG-GCGA

leading to two possible maximum length contigs:

AATGCATGCAATGCGA
AATGCAATGCATGCGA

1

                    CAAT ◄——— GCAA

        (IN) ┈► AATG ⇄ ATGC ⇄ TGCA

              TGCG   CATG ◄— GCAT

        (OUT) ◄┈ GCGA

Question 2

Sequence alignment

```
ACGTATAGGC
AC-TA-A-GC
```

	-	A	C	G	T	A	T	A	G	G	C
-	0	−2	−4	−6	−8	−10	−12	−14	−16	−18	−20
A	−2	1	−1	−3	−5	−7	−9	−11	−13	−15	−17
C	−4	-1	2	0	−2	−4	−6	−8	−10	−12	−14
T	−6	-3	0	1	1	−1	−3	−5	−7	−9	−11
A	−8	-5	-2	-1	0	2	0	−2	−4	−6	−8
A	−10	-7	-4	-3	-2	1	1	1	-1	−3	−5
G	−12	-9	-6	-3	-4	-1	0	0	2	0	−2
C	−14	-11	-8	-5	-4	-3	-2	-1	0	1	1

Modified gap penalty

This above alignment has **7** matches, **3** gaps and **3** gap openings. The score is therefore

$$7 \times 1 - 3 \times 1 - 3 \times 2 = -1 \ .$$

Below is another alignment:

```
ACGTATAGGC
AC---TAAGC
```

It has **6** matches, **1** mismatch, **3** gaps and **1** gap opening and a score of:

$$6 \times 1 - 1 \times 1 - 3 \times 1 - 1 \times 2 = 0 \ .$$

2

Question 3

$$
\mathcal{M} \;=\;
\begin{array}{c}
\\
I_0 \\
M_1 \\
I_1 \\
M_2 \\
I_2 \\
M_3 \\
I_3
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
I_0 & M_1 & I_1 & M_2 & I_2 & M_3 & I_3
\end{array}\\
\left(
\begin{array}{ccccccc}
0.2 & 0.8 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \\
0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.2 & 0.8 & 0 \\
0 & 0 & 0 & 0 & 0.2 & 0.8 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}
\right)
\end{array},
$$

$$
\mathcal{E} \;=\;
\begin{array}{c}
\\
A \\
C \\
G \\
T
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
I_0 & M_1 & I_1 & M_2 & I_2 & M_3 & I_3
\end{array}\\
\left(
\begin{array}{ccccccc}
0.25 & 0.6 & 0.25 & 0.2 & 0.25 & 0 & 0.25 \\
0.25 & 0 & 0.25 & 0.8 & 0.25 & 0.4 & 0.25 \\
0.25 & 0 & 0.25 & 0 & 0.25 & 0 & 0.25 \\
0.25 & 0.4 & 0.25 & 0 & 0.25 & 0.6 & 0.25
\end{array}
\right)
\end{array},
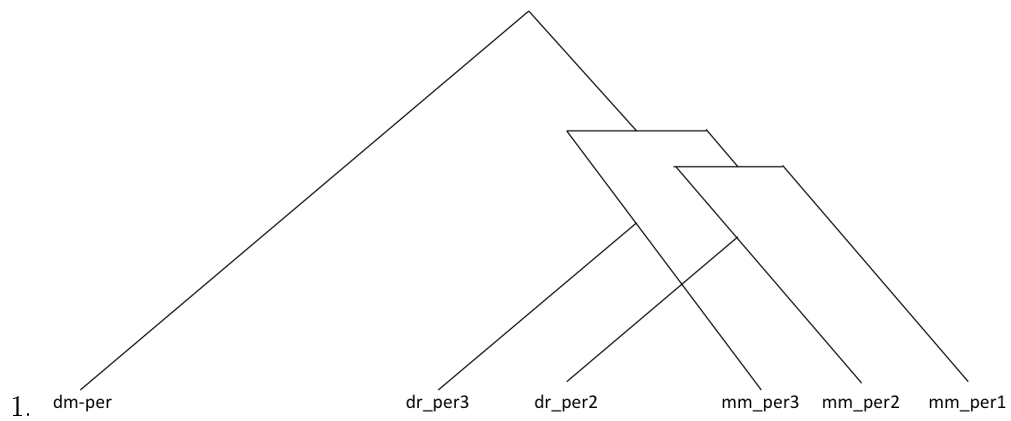$$
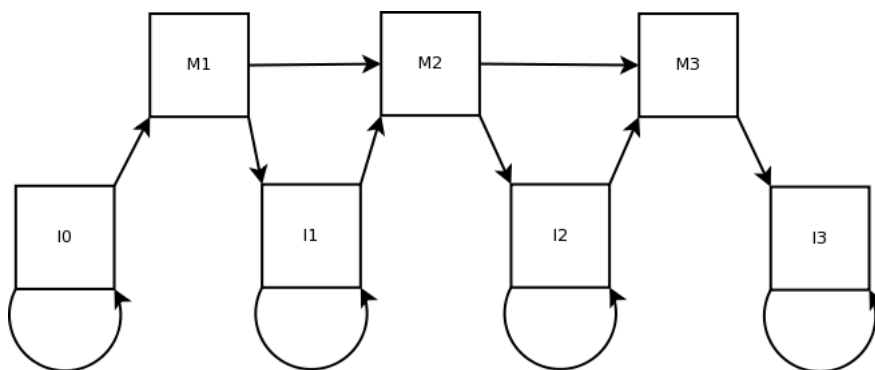
	-	G	A	C	A	T
I_0	1	5/4	25/16	125/64	625/256	3125/1024
M_1	0	0	15	0	375/16	625/32
I_1	0	0	0	75/4	375/16	1875/32
M_2	0	0	0	240	75	0
I_2	0	0	0	0	300	375
M_3	0	0	0	0	0	3600
I_3	0	0	0	0	0	0

The corresponding multiple alignment is:

```
-AC-T
-AA-T
-AC-C
-TC-T
-TC-C
GACAT
```

Question 4

1.

dm-per   dr_per3   dr_per2   mm_per3   mm_per2   mm_per1

The gene tree

2.
- An orthologous pair: **dr_per2** and **mm_per2**
- A paralogous pair in the same species: **mm_per1** and **mm_per2**
- A paralogous pair in different species: **mm_per1** and **dr_per2**