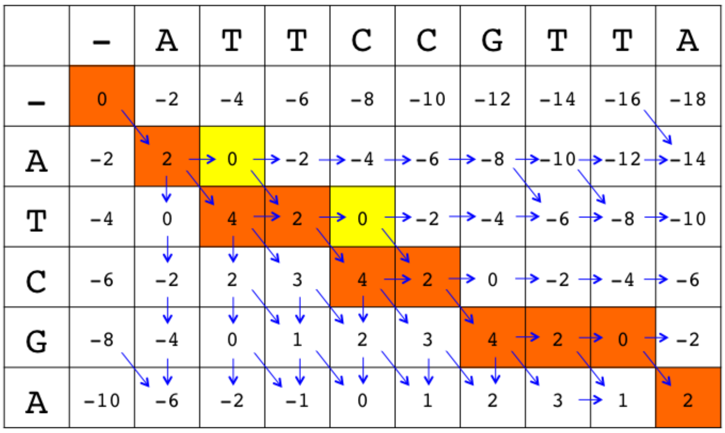
Genomics and Bioinformatics

Exam correction

October 30, 2012

Question 1 - Sequence Alignment

Linear gap penalty



	–	A	T	T	C	C	G	T	T	A
–	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
T	-4	0	4	2	0	-2	-4	-6	-8	-10
C	-6	-2	2	3	4	2	0	-2	-4	-6
G	-8	-4	0	1	2	3	4	2	0	-2
A	-10	-6	-2	-1	0	1	2	3	1	2

the four optimal alignments are below:

```
ATTCCGTTA   ATTCCGTTA   ATTCCGTTA   ATTCCGTTA
AT--CG--A   AT-C-G--A   A-T-CG--A   A-TC-G--A
```

Affine gap penalty

All alignments have 5 matches and 4 gaps. The first has 2 gap openings, hence a score of,

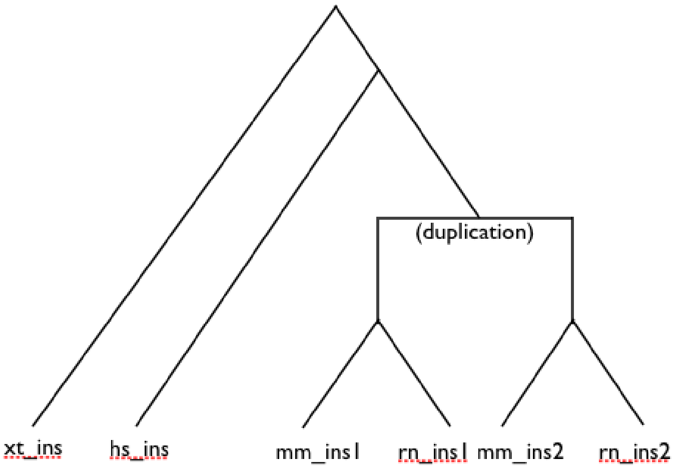
$$5 \times 2 - 4 \times 1 - 2 \times 2 \;=\; 2\;,$$

the other three have 3 gap openings, with a score of:

$$5 \times 2 - 4 \times 1 - 3 \times 2 \;=\; 0\;.$$

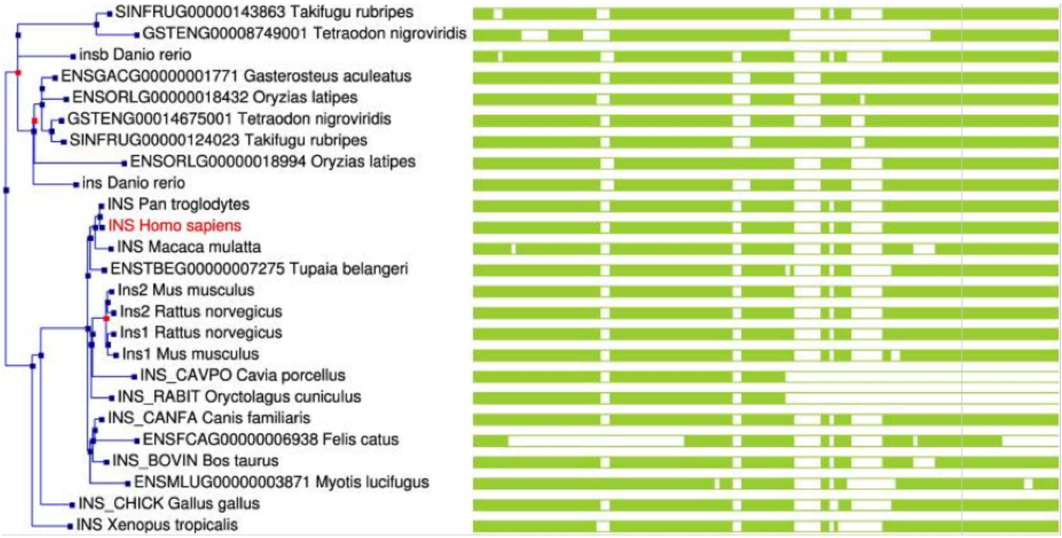
Question 2 - Phylogenetic trees

1. The tree:



2. - A pair of orthologs: mm_ins1, rn_ins1
   - A pair of parologs from the same species: mm_ins1, mm_ins2
   - A pair of parologs from different species: rn_ins1, mm_ins2

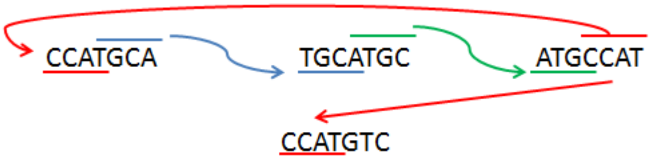
Here is the reference document and the complete tree (duplications in red):


SINFRUG00000143863 Takifugu rubripes
GSTENG00008749001 Tetraodon nigroviridis
insb Danio rerio
ENSGACG00000001771 Gasterosteus aculeatus
ENSORLG00000018432 Oryzias latipes
GSTENG00014675001 Tetraodon nigroviridis
SINFRUG00000124023 Takifugu rubripes
ENSORLG00000018994 Oryzias latipes
ins Danio rerio
INS Pan troglodytes
INS Homo sapiens
INS Macaca mulatta
ENSTBEG00000007275 Tupaia belangeri
Ins2 Mus musculus
Ins2 Rattus norvegicus
Ins1 Rattus norvegicus
Ins1 Mus musculus
INS_CAVPO Cavia porcellus
INS_RABIT Oryctolagus cuniculus
INS_CANFA Canis familiaris
ENSFCAG00000006938 Felis catus
INS_BOVIN Bos taurus
ENSMLUG00000003871 Myotis lucifugus
INS_CHICK Gallus gallus
INS Xenopus tropicalis

`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652215/`

Question 3 - Genome Assembly

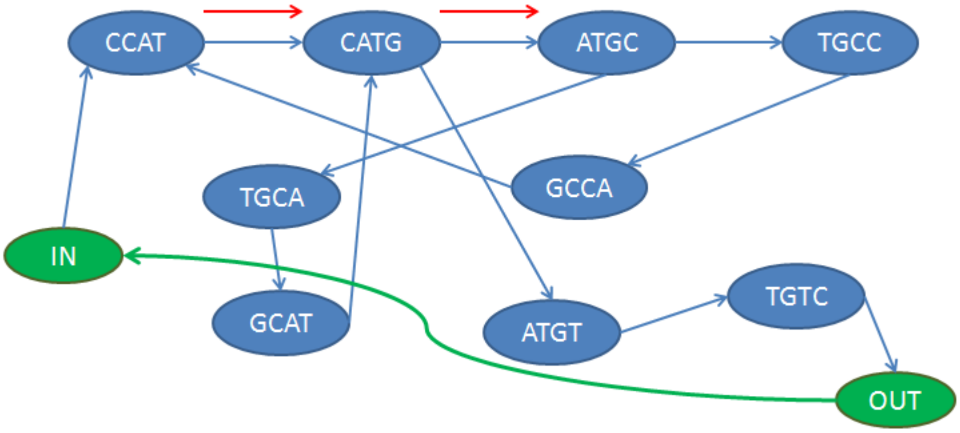
1. The Hamiltonian path starts with the blue arrow and generates the following contig: CCATGCATGCCATGTC



2. The 4-mers and 5-mers sets are:

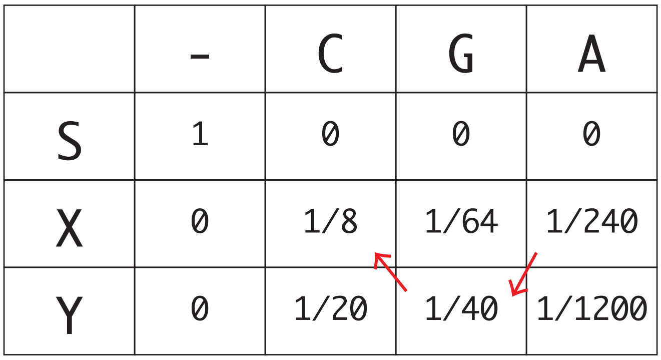
$$\begin{aligned} S_4 &= \{\text{CCAT,CATG,ATGC,TGCA,GCAT,TGCC,GCCA,ATGT,TGTC}\}. \\ S_5 &= \{\text{CCATG,CATGC,ATGCA,} \\ &\quad \text{TGCAT,GCATG,CATGC,} \\ &\quad \text{ATGCC,TGCCA,GCCAT,} \\ &\quad \text{CCATG,CATGT,ATGTC}\}. \end{aligned}$$

3. The de Bruijn graph is as below:



4. The green edge in the figure above makes the graph eulerian, possible paths are
CCAT-CATG-ATGC-TGCA-GCAT-CATG-ATGC-TGCC-GCCA-CCAT-CATG-ATGT-TGTC
CCAT-CATG-ATGC-TGCC-GCCA-CCAT-CATG-ATGC-TGCA-GCAT-CATG-ATGT-TGTC
leading to two possible contigs:
CCATGCATGCCATGTC
CCATGCCATGCATGTC
The second contig does not contain the read TGCATGC (for example).

Question 4 - HMMs



	-	C	G	A
S	1	0	0	0
X	0	1/8	1/64	1/240
Y	0	1/20	1/40	1/1200

The most probable hidden sequence associated to CGA is XYX.