CB2-201: Problem set

Malay (malay@uab.edu)

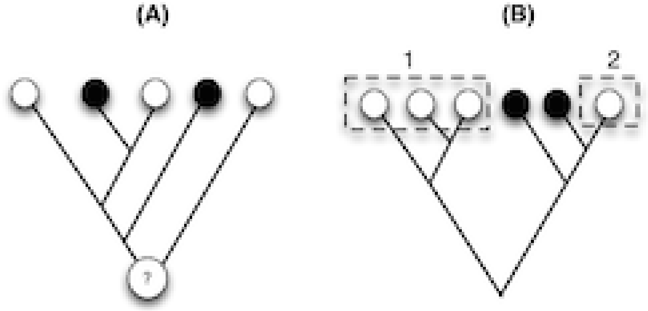
February 27, 2015

Contents

1 Q1

(A) The principle of parsimony defines the correct tree topology as the one with the minimum number of events (e.g. gain or loss). In the figure, there are two characters, white and black, represented in a tree structure. Following the rule of parsimony can you guess the color of the last universal common ancestor (LUCA), or the root node of the tree (marked with a question mark)?

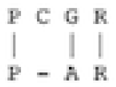
(B) “Homology” is defined as genes that descend from a common ancestor (vertical descent). “Homoplasy” is defined as genes with similarity to other genes but are not related by descent from a common ancestor. In the figure there are two groups (1 & 2) marked by dotted boxes. Identify each group either as homology or homoplasy.



2 Q2

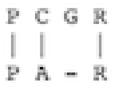
Score each alignment in the following figure using BLOSUM62 matrix. What are the scores? And which alignment is better? Hint: you may ignore insertion and deletion.

1

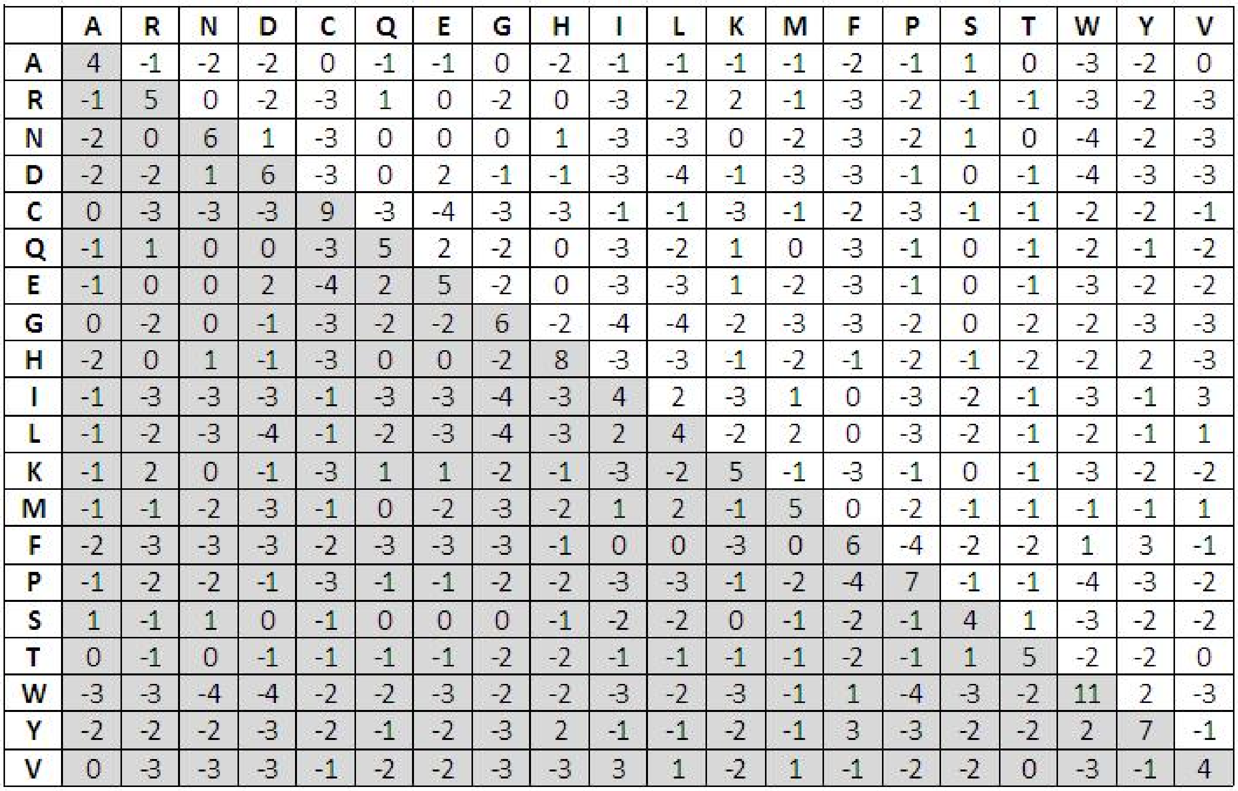


```
P C G R
|   | |
P - A R
```

2



```
P C G R
| |   |
P A - R
```



	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	-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

Figure 1: BLOSUM62 matrix.

3 Q3

The most frequent way to find orthologs is to use the method of reciprocal best BLAST hits (RBBH). Paralogs are defined as genes that came about by duplication within the same species and have stronger similarities than orthologs. Given two genomes A & B, and there genes represented as a and b, we have created these four BLAST results:

A vs B			A vs A			B vs A			B vs B		
Q	H	e-val	Q	H	e-val	Q	H	e-val	Q	H	e-val
a3	b4	10e[-10]	a1	a4	0.0	b1	a2	0.0	b3	b1	10.0
a1	b2	10e[-16]	a3	a1	10e[-20]	b3	a3	10e[-20]	b4	b2	5.0
a3	b1	10.0	a3	a4	10e[-16]	b4	a1	10[e-5]	b4	b1	0.0
a2	b3	5.0	a4	a2	10e[-5]	b4	a3	10e[-10]	b2	b3	10.0

Q and H indicate “Query” and “Hit” respectively.

Could you find the ortholog of gene “a3” from genome A in genome B? List the paralogs in individual genomes?

4 Q4

There are these fastq files (http://cmb.path.uab.edu/training/docs/CB2-201-2015/q4_data.tar.bz2). We do not know which organism it came form. Can you tell, given a mouse and human genome, which one is more likely to be the source? Calculate a log-likelihood score of the sequence coming from human genome compared to mouse.

5 Q5

Write a software that will calculate the best orthologs (also called indexed ortholog) and paralogs given all.vs.all BLAST result file and two genomes. You can use InParanoid algorithm (http://www.sciencedirect.com/science/article/pii/S0022283600951970; Remm et al. (2001)). Or, you can derive your own novel algorithm. Example all.vs.all files are available here (http://cmb.path.uab.edu/training/docs/CB2-201-2015/yeast_data/).

Hint: The problem is not so simple. Two orthologs can be paralogous to each other. So after you have found the indexed orthologs, you need to merge two orthologs into one cluster, if they are paralogous to each other. The paralogs can also be clustered into two separate orthologous groups. In that case you need to resolve this conflict by putting the paralog to its closest orthologus group. For some of the problems that may arise look at the the methods of the InParanoid paper (Remm *et al.*, 2001) .

Bibliography

Remm,M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. *Journal of Molecular Biology*, **314**, 1041–1052.