

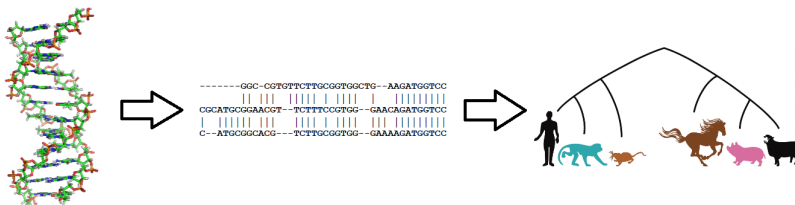
Lecture 4: Tree Construction

Rob Powell

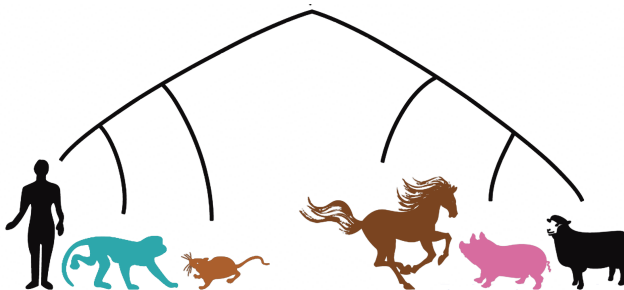
`robert.powell@dur.ac.uk`

Tree Construction

- 1 Extract DNA sequence from each species.
- 2 Align the sequences.
- 3 Compute the inter-species distances.
- 4 Build the tree from the distance matrix.









Phylogenetics



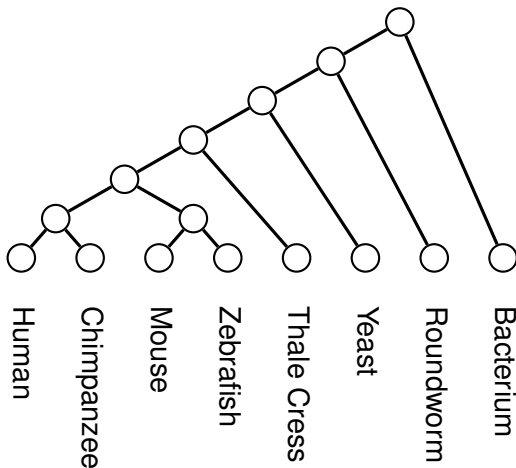
- The study of the evolutionary relationships between species.
- We can ask how similar (or how different) certain species are today.








From Sequences to Phylogenies

- We share 98% of our genes with chimps, but only 85% with zebrafish.
- We ought to be more closely related (on the phylogenetic tree) to chimps than zebrafish.
- Aligned sequences can be converted into a distance measure between species (e.g. % difference).
- We can use the distance matrix to build a tree.

	Chimpanzee, <i>Pan troglodytes</i>	98%
	Mouse, <i>Mus musculus</i>	85%
	Zebrafish, <i>Danio rerio</i>	85%
	Thale cress, <i>Arabidopsis thaliana</i>	26%
	Yeast, <i>Saccharomyces cerevisiae</i>	23%
	Roundworm, <i>Caenorhabditis elegans</i>	21%
	Bacterium, <i>Escherichia coli</i>	7%

From Sequences to Phylogenies



	Chimpanzee, <i>Pan troglodytes</i>	98%
	Mouse, <i>Mus musculus</i>	85%
	Zebrafish, <i>Danio rerio</i>	85%
	Thale cress, <i>Arabidopsis thaliana</i>	26%
	Yeast, <i>Saccharomyces cerevisiae</i>	23%
	Roundworm, <i>Caenorhabditis elegans</i>	21%
	Bacterium, <i>Escherichia coli</i>	7%

Building a phylogenetic tree

	Dog	Bear	Skunk	Weasel	Seal	Sealion	Cat	Monkey
Dog	0	32	48	51	50	48	98	148
Bear		0	26	34	29	33	84	136
Skunk			0	42	44	44	92	152
Weasel				0	44	38	86	142
Seal					0	24	89	142
Sealion						0	90	142
Cat							0	148
Monkey								0

Construct a plausible phylogenetic tree for these species.

The WPGMA algorithm

- 1 Pick the closest pair of species "a" and "b". i.e.

$$\min_{a,b} d(a,b)$$

- 2 Cluster these two species together, and replace with a new species "ab".
- 3 Calculate the data entries for the new column "ab" using the formula:

$$d(ab, c) = \frac{d(a, c) + d(b, c)}{2}$$

- 4 Repeat until tree is complete.

Building a phylogenetic tree

	Dog	Bear	Skunk	Weasel	Seal	Sealion	Cat	Monkey
Dog	0	32	48	51	50	48	98	148
Bear		0	26	34	29	33	84	136
Skunk			0	42	44	44	92	152
Weasel				0	44	38	86	142
Seal					0	24	89	142
Sealion						0	90	142
Cat							0	148
Monkey								0

24 is the minimum score in the distance matrix. So we will group Seal and Sealion together.

Building a phylogenetic tree

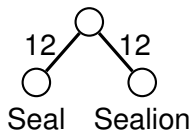
	Dog	Bear	Skunk	Weasel	Seal	x	Sealion	Cat	Monkey
Dog	0	32	48	51	50	49	48	98	148
Bear		0	26	34	29	31	33	84	136
Skunk			0	42	44	44	44	92	152
Weasel				0	44	41	38	86	142
Seal					0		24	89	142
x								89.5	142
Sealion							0	90	142
Cat								0	148
Monkey									0

Call the new species 'x' and calculate the new distances from 'x' to all other species. Now we can remove the rows and columns for Seal and Sealion.

Building a phylogenetic tree

	Dog	Bear	Skunk	Weasel	x	Cat	Monkey
Dog	0	32	48	51	49	98	148
Bear		0	26	34	31	84	136
Skunk			0	42	44	92	152
Weasel				0	41	86	142
x					0	89.5	142
Cat						0	148
Monkey							0

Building a phylogenetic tree



Building a phylogenetic tree

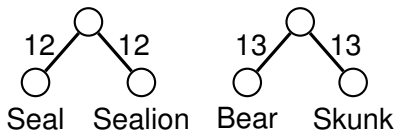
	Dog	Bear	Skunk	Weasel	x	Cat	Monkey
Dog	0	32	48	51	49	98	148
Bear		0	26	34	31	84	136
Skunk			0	42	44	92	152
Weasel				0	41	86	142
x					0	89.5	142
Cat						0	148
Monkey							0

26 is the minimum score in the distance matrix. So we will group Skunk and Bear together.

Building a phylogenetic tree

	Dog	y	Weasel	x	Cat	Monkey
Dog	0	40	51	49	98	148
y		0	38	37.5	88	144
Weasel			0	41	86	142
x				0	89.5	142
Cat					0	148
Monkey						0

Building a phylogenetic tree



Building a phylogenetic tree

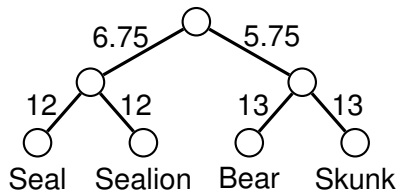
	Dog	y	Weasel	x	Cat	Monkey
Dog	0	40	51	49	98	148
y		0	38	37.5	88	144
Weasel			0	41	86	142
x				0	89.5	142
Cat					0	148
Monkey						0

37.5 is the minimum score in the distance matrix. So we will group the clusters 'x' and 'y' together.

Building a phylogenetic tree

	Dog	Weasel	z	Cat	Monkey
Dog	0	51	44.5	98	148
Weasel		0	39.5	86	142
z			0	88.75	143
Cat				0	148
Monkey					0

Building a phylogenetic tree



Building a phylogenetic tree

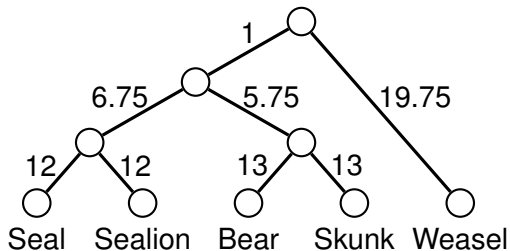
	Dog	Weasel	z	Cat	Monkey
Dog	0	51	44.5	98	148
Weasel		0	39.5	86	142
z			0	88.75	143
Cat				0	148
Monkey					0

39.5 is the minimum score in the distance matrix. So we will group Weasel and the cluster 'z' together.

Building a phylogenetic tree

	Dog	w	Cat	Monkey
Dog	0	47.75	98	148
w		0	87.375	142.5
Cat			0	148
Monkey				0

Building a phylogenetic tree



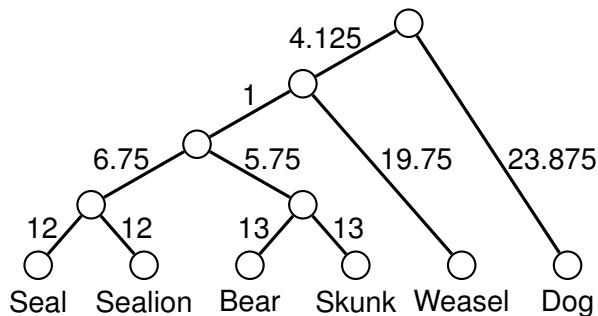
Building a phylogenetic tree

	Dog	w	Cat	Monkey
Dog	0	47.75	98	148
w		0	87.375	142.5
Cat			0	148
Monkey				0

47.75 is the minimum score in the distance matrix. So we will group Dog and the cluster 'w' together.

	v	Cat	Monkey
v	0	92.6875	145.25
Cat		0	148
Monkey			0

Building a phylogenetic tree



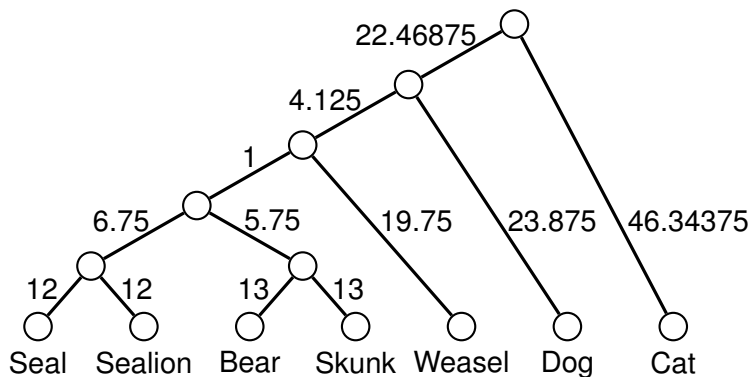
Building a phylogenetic tree

	v	Cat	Monkey
v	0	92.6875	145.25
Cat		0	148
Monkey			0

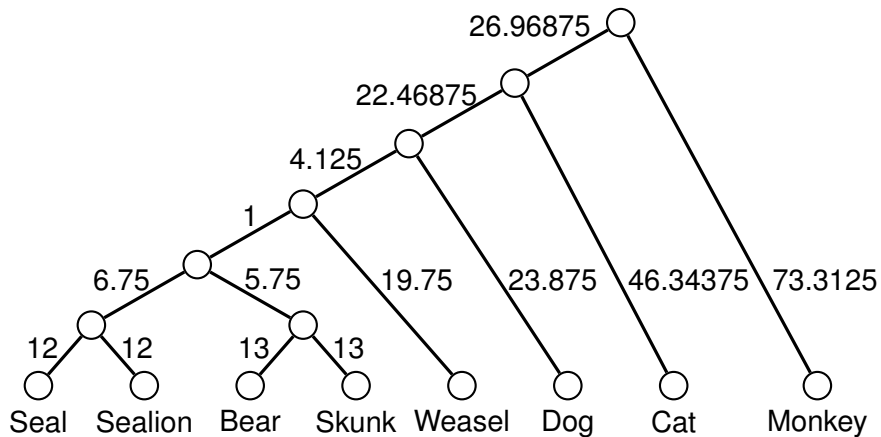
92.6875 is the minimum score in the distance matrix. So we will group Cat and the cluster 'v' together.

	u	Monkey
u	0	146.625
Monkey		0

Building a phylogenetic tree



Building a phylogenetic tree



The UPGMA algorithm

This algorithm differs ever so slightly from WPGMA in the way we calculate new values when we cluster species.

- 1 Pick the closest pair of species "a" and "b".
- 2 Cluster these two species together, and replace with a new species "ab".
- 3 Calculate the data entries for the new column "ab" using the formula:

$$d(ab, c) = \frac{|a| \cdot d(a, c) + |b| \cdot d(b, c)}{|a| + |b|}$$

- 4 Repeat until tree is complete.

Building a phylogenetic tree

- WPGMA and UPGMA both have the hidden assumption that all taxa are the same distance from the root.
- This assumes that the rate of evolution is constant over time - which is certainly not the case!
- Next week we'll look at a different algorithm which gives a better phylogenetic tree.