

EEG Molecular Evolution Lectures

Andrew Leigh Brown ICAPB

Lecture 1

Diversity and divergence in proteins and DNA.

- Estimating diversity and divergence in proteins and DNA
- Using molecular sequence data in the construction of evolutionary trees
 - Method 1: UPGMA – assumes constant rate
 - Method 2: Fitch Margoliash – allows variation in rate of evolution

Lecture 1

Estimating nucleotide distances between 2 sequences

2 sequences 500 bases long differ at 15 sites

- How much evolution has occurred?

$$15/500 = 3\% ?$$

- Not necessarily
 - Only 4 nucleotides, so a site could change 2 x and return to original state
- Need to correct for “***multiple hits***”
- Probability of multiple hits is proportional to overall difference.

Lecture 1

Correction for multiple hits.

Assume an *evolutionary model*

2 examples: Jukes and Cantor (JC) and Kimura 2-parameter (K2P)

Jukes and Cantor (simplest)

Probability of all changes equal

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3} \lambda\right)$$

λ : observed proportion of sites where two sequences differ

K: estimated nucleotide distance.

for $\lambda = 5.0\%$, $K = 5.17\%$,

but for $\lambda = 25\%$, $K = 30.4\%$

Lecture 1

Kimura 2-parameter model (K2P).

- Because of their chemical structure the 4 nucleotides do not mutate at equal rates in each direction
 - bases are more likely to mutate to chemically similar ones
 - A (adenosine) and G (guanine) are purines
 - C (cytosine) and T (thymidine) are pyrimidines
 - A ↔ G and C ↔ T: **Transitions**
 - A or G ↔ C or T: **Transversions**

K2P model:
$$K = -\frac{1}{2} \ln[(1 - 2P - Q)\sqrt{(1 - 2Q)}]$$

P: proportion of sites at which sequences differ by a transition

Q: proportion at which they differ by a transversion

$$P + Q = \lambda$$

Lecture 1

Other evolutionary models.

- Both JC & K2P also assume
 - all sites along sequence have equal substitution rate
 - frequency of all 4 bases in sequence is 25%

Other models

1. Unequal base frequencies, JC model (“Tajima-Nei”)
2. Unequal base frequencies, K2P model (“Tamura-Nei”)
3. Different substitution rates at different sites
 - Protein coding DNA
 - Non-coding DNA

Lecture 1

Divergence in protein-coding DNA.

- In a protein coding region, position in the codon affects rate of evolution

		2nd base in codon				
		U	C	A	G	
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

3rd base in codon

Lecture 1

Synonymous and nonsynonymous sites

Synonymous nucleotide substitutions

- do not change the amino acid

Non-synonymous, (amino acid replacement) substitutions,

- do change the amino acid.

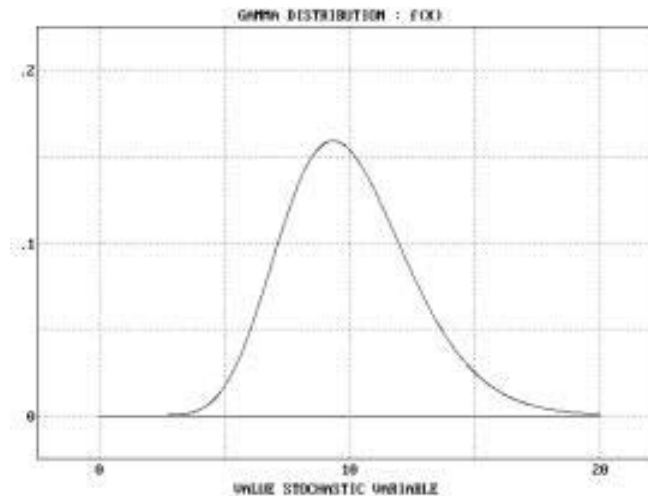
Lecture 1

Estimating synonymous and nonsynonymous rates

- Two main methods
- Both calculate 1 average rate for synonymous sites & another for nonsynonymous sites
 - Li-Wu-Luo
 - Uses a matrix of preferred substitution pathways to estimate number of synonymous changes
 - Nei-Gojobori
 - Estimates from direct comparison of sequences
- The distance is then estimated for each of the 2 categories using the JC model

Lecture 1

Different rates at different sites: non-coding DNA.



The gamma distribution

- Assuming nucleotide substitution rates vary across sites according to a γ distribution (above)
 - Shape of γ distribution described by a single parameter, α , which can be estimated directly from sequence datasets
 - JC, K2P and Tamura-Nei methods can all now be calculated using γ -distributed substitution rates

Lecture 1

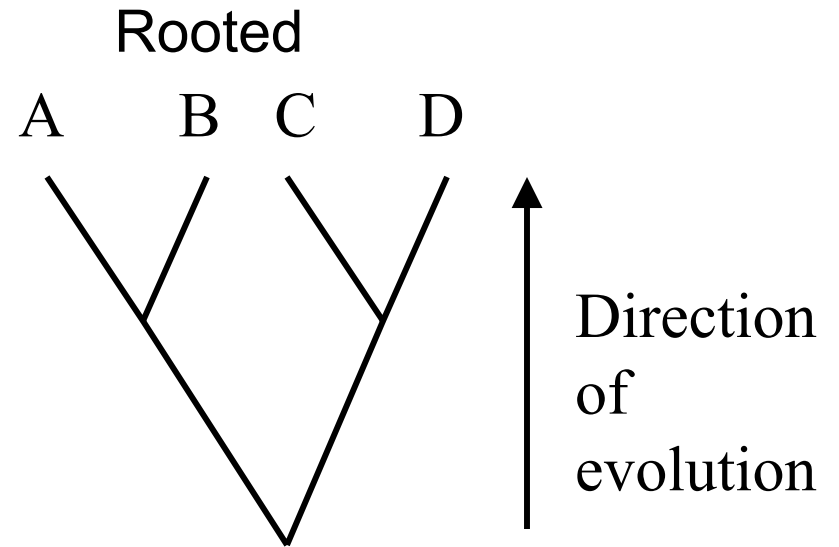
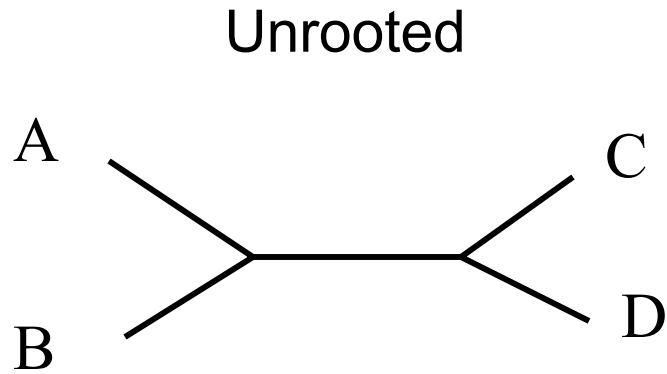
Phylogenies from distances.

Two basic types of tree – *rooted* and *unrooted*.

- Rooted
 - direction of evolution is known (UPGMA).
- Usually not the case with sequence data,
 - Most methods produce unrooted trees (Fitch-Margoliash, Neighbor-Joining)
- If a sequence known to be more distant than any other is included (“outgroup”), can use it to impose a root.

Lecture 1

Rooted and Unrooted Trees



Lecture 1

Rooted and Unrooted Trees.

Numbers of possible trees

- There are a lot

Number of taxa	No. rooted trees	No. unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

Lecture 1

UPGMA trees.

UPGMA

- Rooted tree
- Assumes equal rate of evolution in all branches

Use the following distance matrix to construct a UPGM tree

Species	A	B	C	D
A	-			
B	4	-		
C	9	7	-	
D	20	18	17	-

Lecture 1

UPGMA trees (2).

- Group closest relatives, calculate branch lengths as:

$$\frac{d_{AB}}{2} = 2$$

Species	A	B	C	D
A	-			
B	4	-		
C	9	7	-	
D	20	18	17	-

- Construct new matrix treating A+B as a single composite group:

$$d_{(AB)C} = \frac{(d_{AC} + d_{BC})}{2} = 8$$

Species	(AB)	C
C	8	
D	19	17

Lecture 1

UPGMA trees (3).

Calculate branch length from common ancestor to C

$$\frac{d_{(AB)C}}{2} = 4$$

Species	(AB)	C
C	8	
D	19	17

Now calculate the final distance, from [(AB)C] to D

$$d_{[(AB)C]D} = \frac{(d_{(AB)D} + d_{CD})}{2} = 18$$

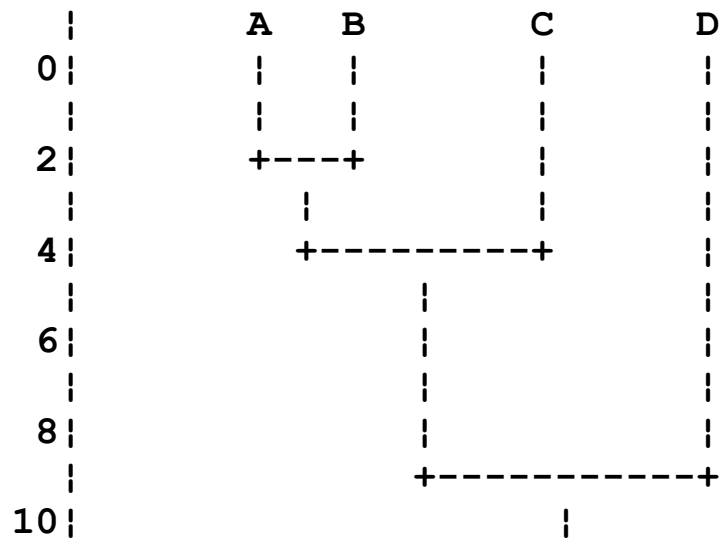
and the branch length to D

$$\frac{d_{[(AB)C]D}}{2} = 9$$

Lecture 1

UPGMA trees (4).

final tree:



Lecture 1

Fitch Margoliash Trees

- Unrooted
- Rate of evolution can vary across branches
 - estimate branch lengths from the data

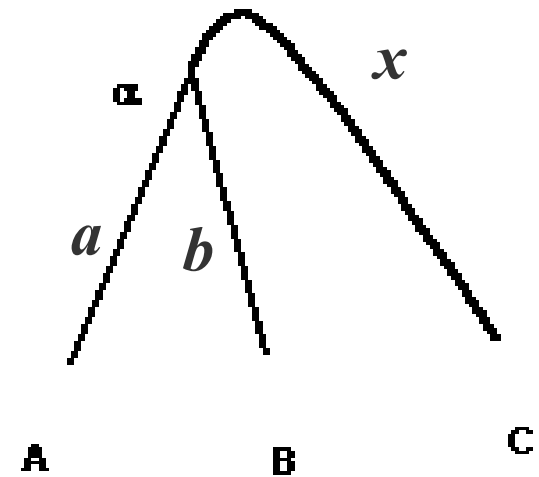
If a = distance from A to node α ,

b = distance from B to α ,

x = distance from α to C

we can use the distances from each of A and B to C, to estimate a and b , i.e. $(a + x)$ and $(b + x)$

	A	B	C	D
A	-			
B	4	-		
C	9	7	-	
D	20	18	17	-



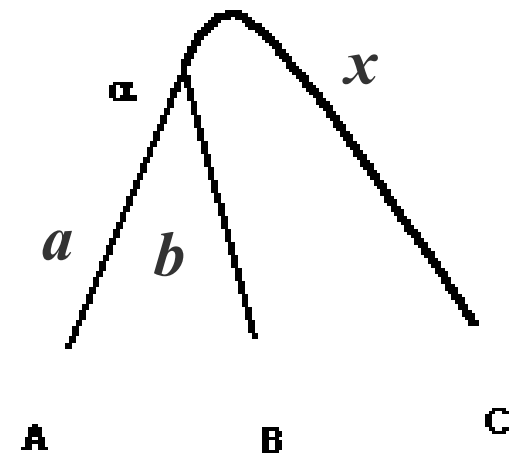
Lecture 1

Fitch Margoliash Trees (2).

Fitch Margoliash tree

The distances from each of A and B to C
are $(a + x)$ and $(b + x)$

	A	B	C	D
A	-			
B	4	-		
C	9	7	-	
D	20	18	17	-



From the data matrix we write

$$1) \quad a + b = 4$$

$$2) \quad a + x = 9$$

$$3) \quad b + x = 7$$

subtracting equ 3 from equ 2 gives: 4) $a - b = 2$

adding equ 1 + equ 4 gives: $2a = 6$

so $a = 3$, $b = 1$, $x = 6$

And the branch lengths a and b are not equal)

Lecture 1

Fitch Margoliash Trees (3).

Fitch Margoliash tree

The matrix is recalculated with AB combined as

	(AB)	C
C	8	
D	19	17

From this data matrix we can write

$$5) \ c + d = 8$$

$$6) \ c + e = 19$$

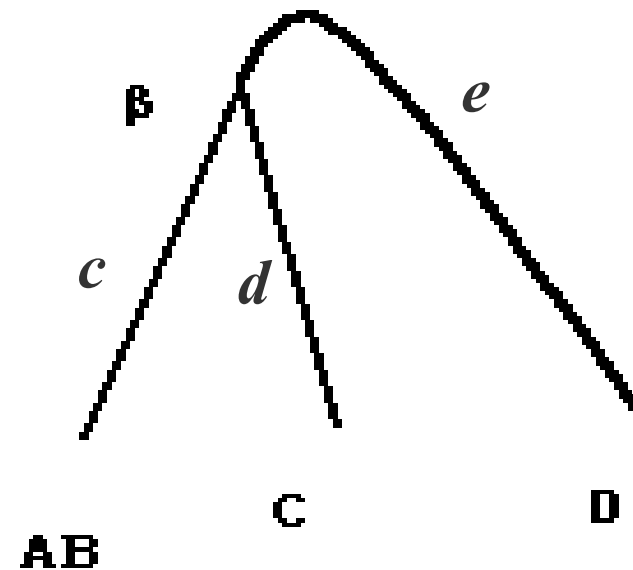
$$7) \ d + e = 17$$

$$\text{so: } c - d = 2$$

$$c = 5$$

$$d = 3$$

$$e = 14$$



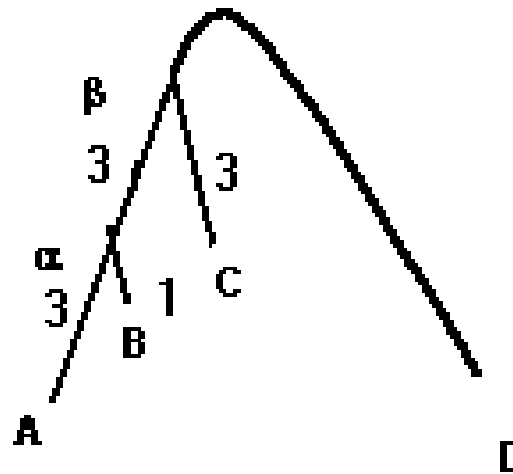
Lecture 1

Fitch Margoliash Trees (4).

Final tree

We can now place the two nodes, α and β on the tree as the distance between them is given by $x - d = 3$.

confirming that in this case the branch lengths below a common node are not equal, i.e. the rate of evolution was not the same in all branches



Lecture 1

Phylogenetic trees: character state data.

- Trees can also be constructed using the nucleotides at each positions as independent characters
- 2 approaches available:
 - Maximum parsimony (=cladistics)
 - Maximum Likelihood

Lecture 1

Maximum parsimony trees.

Using cladistics to build trees from sequence data

Species	Sequence						
a	T	A	C	T	C	G	G
b	T	T	A	C	A	C	G
c	G	T	A	C	A	C	G
d	G	A	A	C	T	C	G
Site:	1	2	3	4	5	6	7

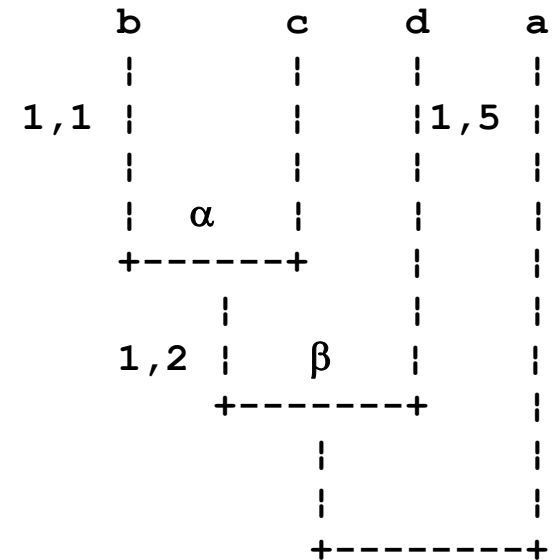
- Site 7: no variation, ignored
- Sites 3,4 and 6: only distinguish single sequences.
- Sites 1, 2 and 5 are *phylogenetically informative*.
 - **b** & **c** share nucleotides at 2 and 5,
 - **a** & **b** and **c** & **d** share nucleotides at 1 only.
- **b** & **c** grouped together to produce the tree with the smallest number of changes

Lecture 1

Maximum parsimony trees (2).

Species Sequence

a	T	A	C	T	C	G	G	
b	T	T	A	C	A	C	G	GTACACG
c	G	T	A	C	A	C	G	
d	G	A	A	C	T	C	G	
Site:	1	2	3	4	5	6	7	GAACA/TCG



Where did the T/G change at site 1 occur?

- Compare **b** & **c** with **d**; **d** has G at this site, this is “ancestral” state.
 - Sequence **b** is ancestral to **c**
- ancestral sequence at node β can be partly reconstructed. At site 2, sequences *b* and *c* both have T but *d* and *a* both have A.
 - Change at site 2 occurred in common ancestor of **b** & **c**

Lecture 1

Diversity and divergence in proteins and DNA.

Summary

1. Genetic distances based on nucleotide sequences must be corrected for "multiple hits".
 - several different models can be used
2. Tree construction methods are designed to identify the best tree out of a very large number of possible trees.
3. Tree construction methods can be based on genetic distance estimates (eg UPGM, Fitch-Margoliash) or on sequence data directly (eg Maximum Parsimony)
4. UPGM gives a rooted tree and assumes a **constant** rate of evolution. Fitch-Margoliash allows for variable rates of evolution and gives an unrooted tree.
5. Maximum parsimony is a character-based method that allows reconstruction of the ancestral states