Making Markov matrices from phylogenetic trees

Julia A Shore, Barbara R Holland, Jeremy G Sumner, Kay Nieselt, Alex Popinga and Peter R Wills

February 13, 2019

UNIVERSITY of TASMANIA

Summary of talk today

Summary of talk today

- Phylogenetic trees: what are they?

Summary of talk today

- Phylogenetic trees: what are they?
- A brief biological overview

Summary of talk today

- Phylogenetic trees: what are they?
- A brief biological overview
- What we want to test

Summary of talk today

- Phylogenetic trees: what are they?
- A brief biological overview
- What we want to test
- Building Markov matrices from phylogenetic trees

Summary of talk today

- Phylogenetic trees: what are they?
- A brief biological overview
- What we want to test
- Building Markov matrices from phylogenetic trees
- Using Markov matrices built from trees

Summary of talk today

- Phylogenetic trees: what are they?
- A brief biological overview
- What we want to test
- Building Markov matrices from phylogenetic trees
- Using Markov matrices built from trees
- The results

Phylogenetics: the aim

In phylogenetics, we are interested in mapping evolutionary history. For example, we may be interested in finding out how long ago there lived the common ancestor of horses and zebras.

What is a phylogenetic tree?

Phylogenetic trees are diagrams which map evolutionary history.

A (very brief) biological overview

A (very brief) biological overview

- Proteins do things in cells: they are borne from DNA and they enact DNA's instructions

A (very brief) biological overview

- Proteins do things in cells: they are borne from DNA and they enact DNA's instructions
- Proteins are chains of amino acids, there are twenty unique amino acids present in proteins

A (very brief) biological overview

- Proteins do things in cells: they are borne from DNA and they enact DNA's instructions
- Proteins are chains of amino acids, there are twenty unique amino acids present in proteins
- The twenty amino acids present in proteins are often represented by single letters: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V

Evolution of proteins

Evolution of proteins

Over time, changes in proteins can be observed.

Evolution of proteins

Over time, changes in proteins can be observed.

Of these changes, one field of study is to observe when one amino acid is replaced by another in a protein.

Rates of change between amino acids

Rates of change between amino acids

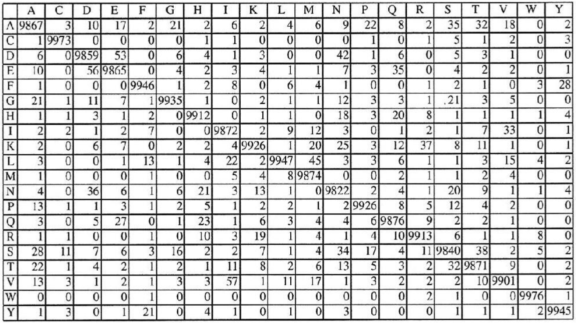
- Rates of change between amino acids are observed and measured

Rates of change between amino acids

- Rates of change between amino acids are observed and measured
- Empirical amino acid substitution matrices are 20×20 matrices whose entries represent the observed rate of change between two amino acids

Rates of change between amino acids

- Rates of change between amino acids are observed and measured
- Empirical amino acid substitution matrices are 20×20 matrices whose entries represent the observed rate of change between two amino acids



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	9867	3	10	17	2	21	2	6	2	4	6	9	22	8	2	35	32	18	0	2
C	1	9973	0	0	0	0	1	1	0	0	0	0	1	0	1	5	1	2	0	3
D	6	0	9859	53	0	6	4	1	3	0	0	42	1	6	0	5	3	1	0	0
E	10	0	56	9865	0	4	2	3	4	1	1	7	3	35	0	4	2	2	0	1
F	1	0	0	0	9946	1	2	8	0	6	4	1	0	0	1	2	1	0	3	28
G	21	1	11	7	1	9935	1	0	2	1	1	12	3	3	1	.21	3	5	0	0
H	1	1	3	1	2	0	9912	0	1	1	0	18	3	20	8	1	1	1	1	4
I	2	2	1	2	7	0	0	9872	2	9	12	3	0	1	2	1	7	33	0	1
K	2	0	6	7	0	2	2	4	9926	1	20	25	3	12	37	8	11	1	0	1
L	3	0	0	1	13	1	4	22	2	9947	45	3	3	6	1	1	3	15	4	2
M	1	0	0	0	1	0	0	5	4	8	9874	0	0	2	1	1	2	4	0	0
N	4	0	36	6	1	6	21	3	13	1	0	9822	2	4	1	20	9	1	1	4
P	13	1	1	3	1	2	5	1	2	2	1	2	9926	8	5	12	4	2	0	0
Q	3	0	5	27	0	1	23	1	6	3	4	4	6	9876	9	2	2	1	0	0
R	1	1	0	0	1	0	10	3	19	1	4	1	4	10	9913	6	1	1	8	0
S	28	11	7	6	3	16	2	2	7	1	4	34	17	4	11	9840	38	2	5	2
T	22	1	4	2	1	2	1	11	8	2	6	13	5	3	2	32	9871	9	0	2
V	13	3	1	2	1	3	3	57	1	11	17	1	3	2	2	2	10	9901	0	2
W	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	1	0	0	9976	1
Y	1	3	0	1	21	0	4	1	0	1	0	3	0	0	0	1	1	1	2	9945

The PAM1 matrix, Dayhoff et al. (1978).

A characteristic of amino acids

We are interested in a binary characteraristic of amino acids: aminoacyl-tRNA synthase (aaRS) class.

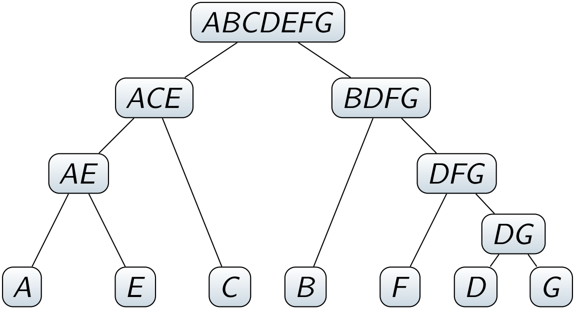
A characteristic of amino acids

We are interested in a binary characteraristic of amino acids: aminoacyl-tRNA synthase (aaRS) class.

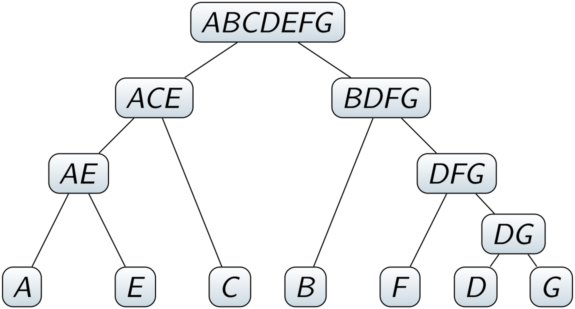
Class I: R, C, Q, E, I, L, M, W, Y, V
Class II: A, N, D, G, H, K, F, P, S, T

We assume that in early life forms, selecting the correct amino acid in building a protein was a crude process that became more refined as time progressed.

We assume that in early life forms, selecting the correct amino acid in building a protein was a crude process that became more refined as time progressed.



We assume that in early life forms, selecting the correct amino acid in building a protein was a crude process that became more refined as time progressed.

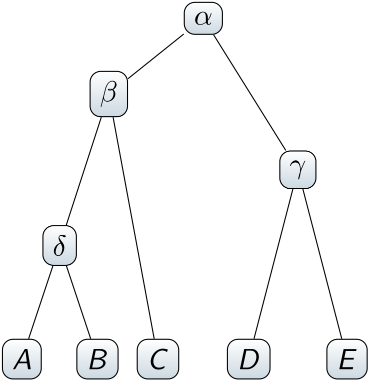


We are hypothesising that the first split was aaRS classes **I** and **II**.

Building Markov matrices from phylogenetic trees

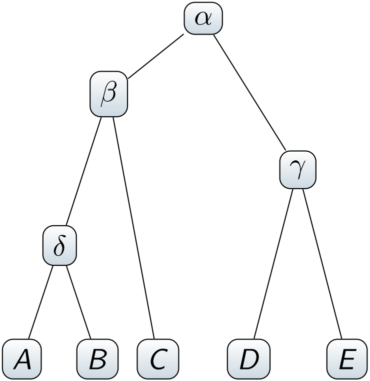
For each tree **node**, we assign a **rate** as a free parameter. The rate of change between two taxa is defined to be the rate associated to most recent common ancestor.

Example of going from a tree to a matrix



$$\begin{array}{c|ccccc} & A & B & C & D & E \\ A & * & \delta & \beta & \alpha & \alpha \\ B & \delta & * & \beta & \alpha & \alpha \\ C & \beta & \beta & * & \alpha & \alpha \\ D & \alpha & \alpha & \alpha & * & \gamma \\ E & \alpha & \alpha & \alpha & \gamma & * \end{array}$$

Example of going from a tree to a matrix



$$\begin{array}{c|ccccc} & A & B & C & D & E \\ A & * & \delta & \beta & \alpha & \alpha \\ B & \delta & * & \beta & \alpha & \alpha \\ C & \beta & \beta & * & \alpha & \alpha \\ D & \alpha & \alpha & \alpha & * & \gamma \\ E & \alpha & \alpha & \alpha & \gamma & * \end{array}$$

Note for a tree of n taxa, the corresponding rate matrix will have $n-1$ parameters.

Using rate matrices borne from trees

We build a 20 taxa tree whose leaves are amino acids.

Using rate matrices borne from trees

We build a 20 taxa tree whose leaves are amino acids.

From this tree, we build a rate matrix using the aforementioned method: it will have 19 free parameters.

Using rate matrices borne from trees

We build a 20 taxa tree whose leaves are amino acids.

From this tree, we build a rate matrix using the aforementioned method: it will have 19 free parameters.

Fit this matrix to an empirical amino acid substitution rate matrix: measure goodness of fit.

What types of trees we tested

What types of trees we tested

- aaRS trees: trees whose first split is aaRS class

What types of trees we tested

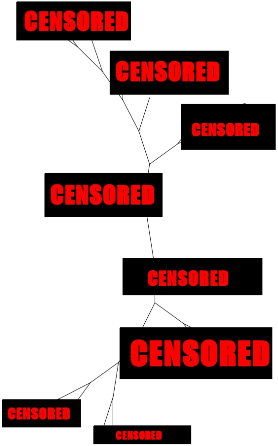
- aaRS trees: trees whose first split is aaRS class
- random trees: randomly generated 20 taxa trees

What types of trees we tested

- aaRS trees: trees whose first split is aaRS class
- random trees: randomly generated 20 taxa trees
- ten-ten trees: randomly generated 20 taxa trees with the constraint of the first split having ten taxa on each side

A special tree to test

A special tree to test



What runs we did

What runs we did

For each type of tree (aaRS, random, ten-ten), generate $n = 100,000$ of them.

What runs we did

For each type of tree (aaRS, random, ten-ten), generate $n = 100,000$ of them.

For each tree, generate the rate matrix and fit it to an empirical amino acid substitution model.

What runs we did

For each type of tree (aaRS, random, ten-ten), generate $n = 100,000$ of them.

For each tree, generate the rate matrix and fit it to an empirical amino acid substitution model.

Record goodness of fit score.

What runs we did

For each type of tree (aaRS, random, ten-ten), generate $n = 100,000$ of them.

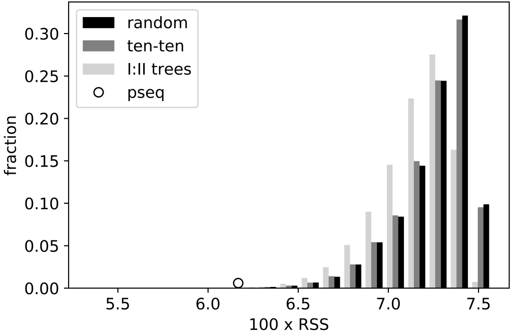
For each tree, generate the rate matrix and fit it to an empirical amino acid substitution model.

Record goodness of fit score.

Output for each type of tree is 100,000 goodness of fit scores.

UNIVERSITY *of* TASMANIA

Results!



(Using the LG empirical amino acid substitution matrix Le and Gascuel (2008).)

Wrapping up

Wrapping up

- Random trees fit the same as ten-ten trees so it would appear that tree shape is having not impact

Wrapping up

- Random trees fit the same as ten-ten trees so it would appear that tree shape is having not impact
- aaRS trees fit better than random trees

Wrapping up

- Random trees fit the same as ten-ten trees so it would appear that tree shape is having not impact
- aaRS trees fit better than random trees
- The pseq tree fit really quite well

Wrapping up

- Random trees fit the same as ten-ten trees so it would appear that tree shape is having not impact
- aaRS trees fit better than random trees
- The pseq tree fit really quite well
- The results support the hypothesis that aaRS class had an impact on the rates of change of amino acids

Thanks for listening!

References I

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure*, pages 345–352.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320.