The Mosaic Model of Nucleotide Substitution

Michael Woodhams
Michael Charleston
Jeremy Sumner

School of IT, University of Sydney

Common assumptions in likelihood calculations:

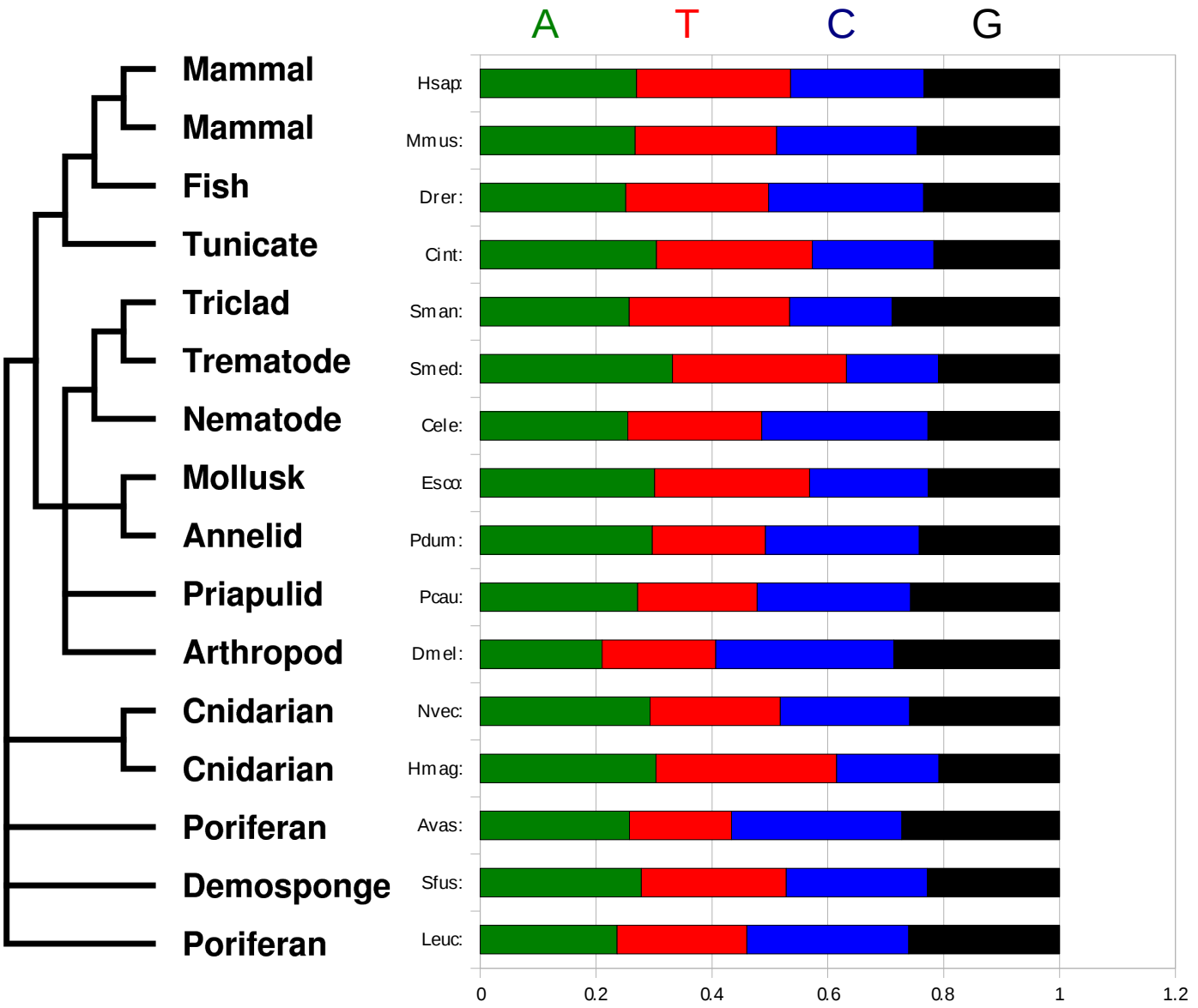
Evolution follows a Markov process on a tree

Sites are independent

SRH: stationary, reversible, homogeneous

HGT, hybridization, lineage sorting, convergent mutations, CpG mutations,
Constraints on 3rd position changed by 1st or 2nd, changing base frequencies
Changing transition/transversion ratio

Base Frequency in a Deep Phylogeny



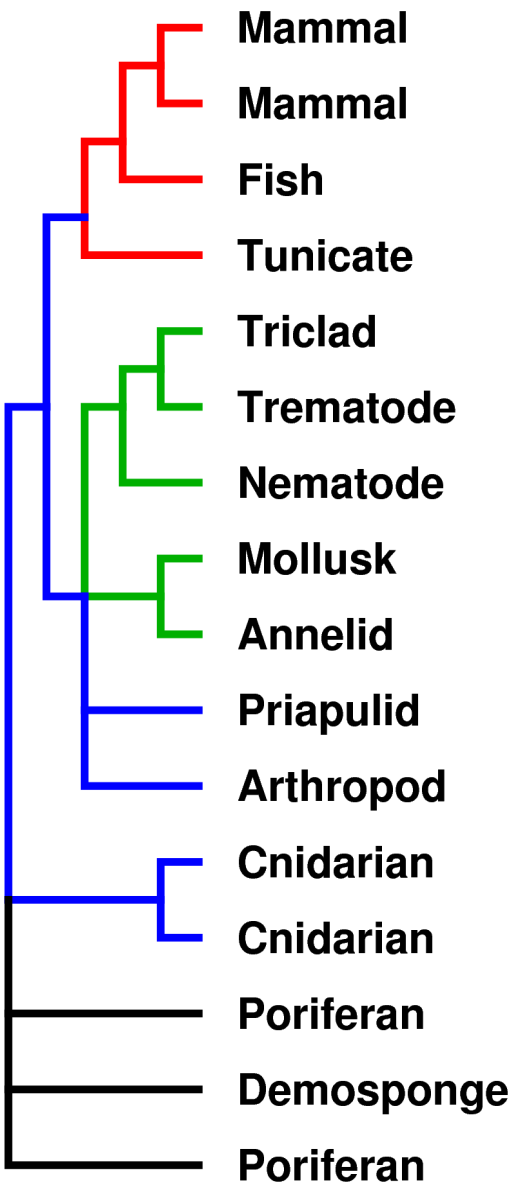
Non-stationary, hence non-time-reversible.

(1245 bases from HSP70C, data and tree Rokas et al Science 2005)

Non-homogeneous models

Different models apply on different parts of the tree.

The Mosaic model contains a number of submodels, each potentially having its own tunable parameters.



We start by assigning models to taxa.

Not knowing the true tree, we can't be sure the models will end up adjacent to each other.

Model assignment could be by base frequency, transition/transversion ratio to near neighbours, and by distance.

Mammal

Mammal

Fish

Tunicate

Triclad

Trematode

Nematode

Mollusk

Annelid

Priapulid

Arthropod

Cnidarian

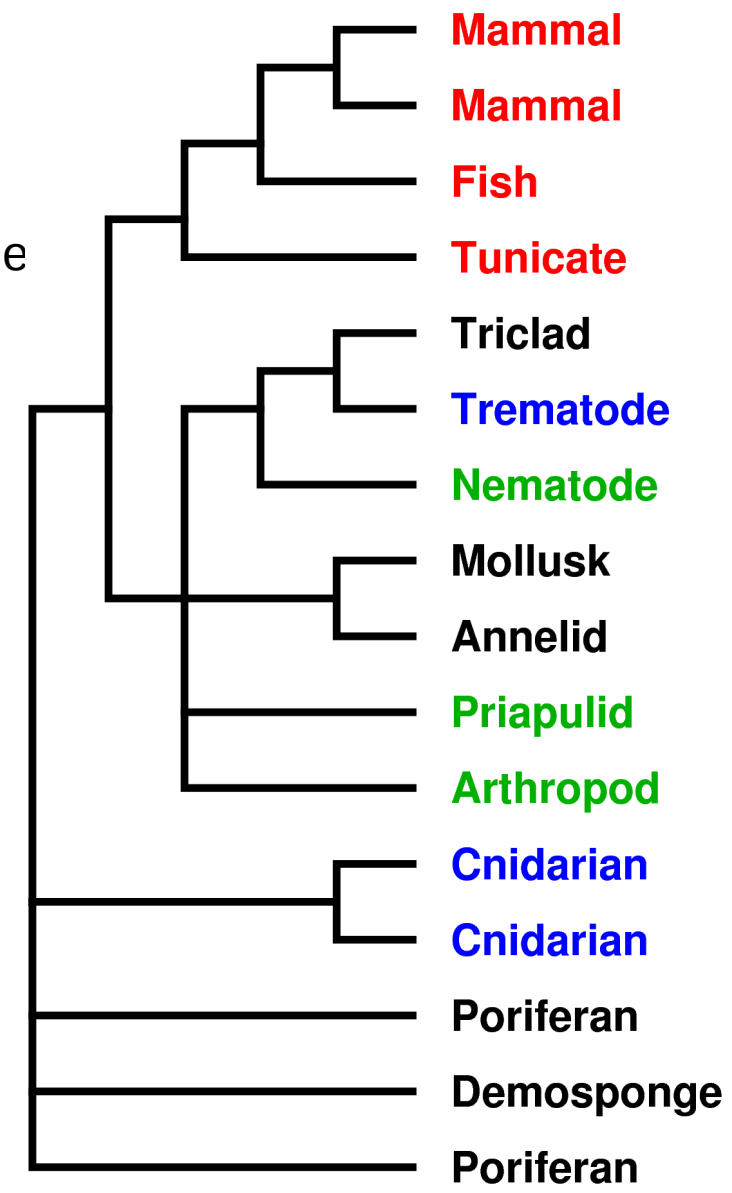
Cnidarian

Poriferan

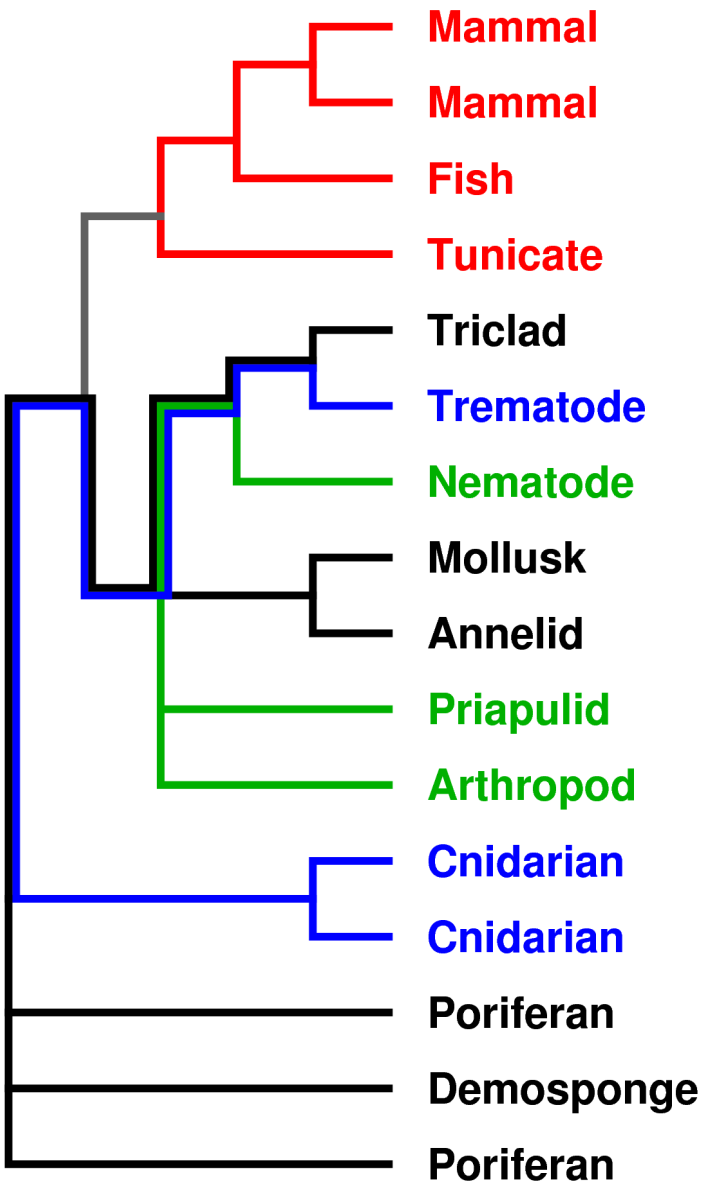
Demosponge

Poriferan

Then we apply the particular topology we are calculating the likelihood on



And we ‘colour’ the branches with the applicable models



(Rule 1: A model applies to a branch if that branch lies on the path between two taxa with that model
Rule 2: If a branch has no models from rule 1, apply a default model to it.)

How to Find a Maximum Likelihood Tree

(standard version)

Start with a reasonable guess at the best tree topology (e.g. NJ tree)
Loop:
    Perform some tree topology change (e.g. NNI, SPR)
    Optimize branch lengths to maximize likelihood
    Compare this likelihood with previous likelihood
    Accept or undo the topology change depending
        on likelihood comparison and search algorithm
Until heuristic says we've done enough
Output best tree found.

How to Find a Maximum Likelihood Tree

(Mosaic version 1)

Start with a reasonable guess at the best tree topology (e.g. NJ tree)
Loop:
    Perform some tree topology change (e.g. NNI, SPR)
    Assign models to each edge
    Optimize branch lengths and **model edge weights** to maximize likelihood
    Compare this likelihood with previous likelihood
    Accept or undo the topology change depending
        on likelihood comparison and search algorithm
Until heuristic says we've done enough
Output best tree found.

How to Find a Maximum Likelihood Tree

(Mosaic version 2)

Start with a reasonable guess at the best tree topology (e.g. NJ tree)
Loop:
    Perform some tree topology change (e.g. NNI, SPR)
    Assign models to each edge
    Optimize branch lengths and **model edge weights** to maximize likelihood
    Penalize likelihood for number of edge weight parameters
    Compare this likelihood with previous likelihood
    Accept or undo the topology change depending
        on likelihood comparison and search algorithm
Until heuristic says we've done enough
Output best tree found.

Implementation Details

The model is being implemented within the PAL* library
+ Open source
+ Object oriented
+ Very flexible (because OO)
+ All the groundwork already done (e.g. tree search)
+ Efficient likelihood calculation algorithm
- No longer in active development
- Essentially undocumented
- Complex code
- Often assumes models are time reversible.

* Phylogenetic Analysis Library, http://www.cebl.auckland.ac.nz/pal-project/

Implementation Problems

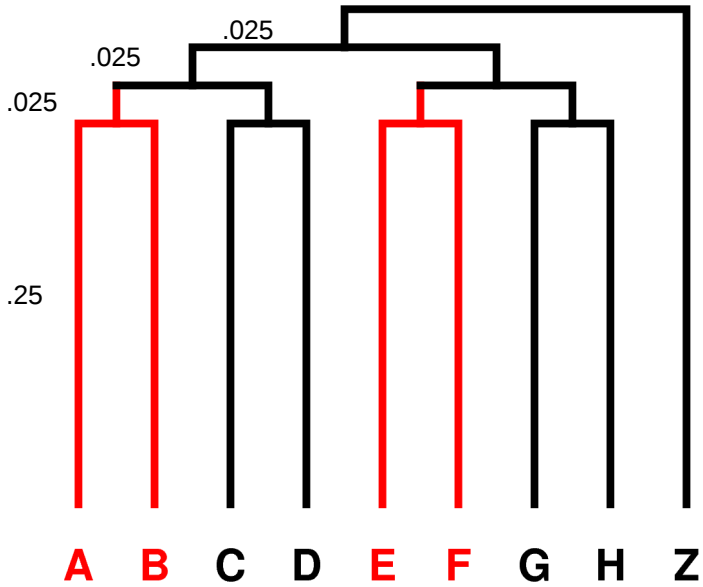
Undocumented code, unmarked obsolete methods...

Non-time-reversibility: Final step in calculating a likelihood is multiplying by prior probability on base frequencies. This changes depending on where you are in the tree, or even which end of an edge you evaluate at.

Branches now have more parameters than just length.

Rates across sites yet to be implemented.

Monte Carlo Testing

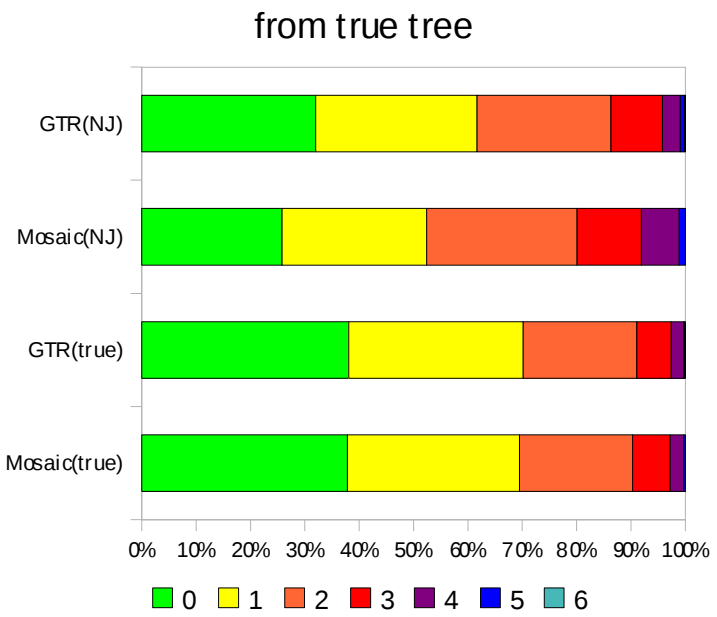


Root base freq (.35,.35,.15,.15)
Black = HKY85 model, Tr/Tv = 1, base freq = (.35,.35,.15,.15)
Red = HKY85 model, Tr/Tv = 10, base freq = (.15, .15, .35, .35)
Sequence length 300 bases. 1000 alignments.
MC data produced by FILO program
http://www.it.usyd.edu.au/~mcharles/software/filo/filo.php

Test model comparison:

Mosaic model: HKY85 on taxa C,D,G,H,Z (1+3 parameters)
HKY85 on taxa A,B,E,F (1+3 parameters)
GTR model. (5+3 parameters)

Robinson Foulds Distance



AIC model comparison on true tree:
Min -14.65 (favours GTR)
5% 7.67
10% 13.28
25% 21.05
50% 39.33
75% 38.17
90% 46.04
95% 50.82
Max 76.10 (favours Mosaic)

320 297 245 95 33 8 1
258 266 277 118 70 11 0
381 321 209 63 23 2 1
378 317 208 69 25 3 0

Acknowledgments

Peter Jarvis (for the name)
ARC (for the money)