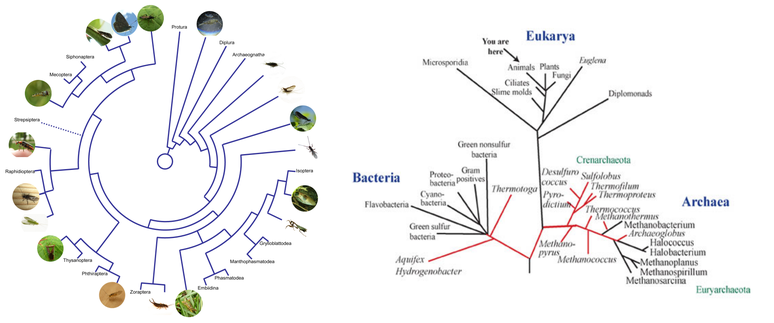
ANALYSIS OF DISTRIBUTIONS ON THE SPACE OF EVOLUTIONARY TREES

{ Yameng Ji * & Dr Michael White, School of Mathematics and Statistics. yameng.ji@ncl.ac.uk michael.white@ncl.ac.uk}

Background & Introduction

The huge amount of data produced by sequencing genomes has produced many different possible relationships between organisms, as expressed using phylogenetic trees. In order to compare the various trees, one needs to be able to describe the variability of these trees. The various plausible trees can be thought of as arising from a random walk in the collection of all possible trees of life.[1]



The project will limit the analysis to the relationships between a small number of species. For the shapes of the non-degenerate evolutionary trees from 5 to 10 species, we focused on their adjacencies and produced the transition matrices around the connections, the eigenvalues and eigenvectors of which were used in the following parts. [2]

In the simulations of real evolution data, we took the model of exponential growth with / without dying. The trees we considered were rooted with a single ancestor.

Aims

- To work out the tree space of 10 species
- By taking the dot product of the distribution from the data of a real evolution and the eigenvectors from the adjacency matrices, one can tell which growth model the data might resemble and therefore estimate the growth rate.

Materials

The number of shapes of the rooted trees is much more than the unrooted trees when generating the same number of species and every rooted tree belongs to a family of a particular shape of unrooted tree.

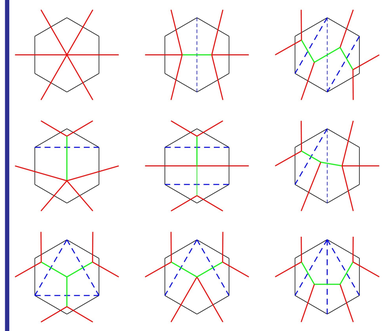
There are various ways of inserting a root into an unrooted tree so that the number of shapes of different rooted trees is much more than the unrooted ones. Given n species, the whole space of all possible collection of combinations consists of $(2n-5)!! = (2n-5)*(2n-7)*(2n-9)\cdots 3*1$ different unrooted trees. For 10 species, there are 98 possible rooted trees but only 11 geometries of unrooted evolutionary trees in the space of more than 2 million trees.

In general, the number of tree shapes grows like the square root of n factorial and so quickly becomes computationally difficult.

# species	# shapes of unrooted trees	rooted trees	size of tree space
3	1	1	1
4	1	2	3
5	1	3	15
6	2	6	105
7	2	11	945
8	4	23	10,395
9	6	46	135,135
10	11	98	2,027,025

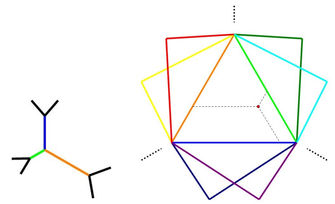
General Methods

Unrooted evolutionary trees of n species are built up from triangulating a polygon with n edges. The 'evolution' starts from a fully degenerate 'star' shape. Every step would split the species more than the previous status and create a new internal edge. Triangulation procedure finishes with a non-degenerate tree, where the polygon is divided into $(n-3)$ triangles.



A hexagon can be triangulated as above, where the 6 red edges represent the 6 species.

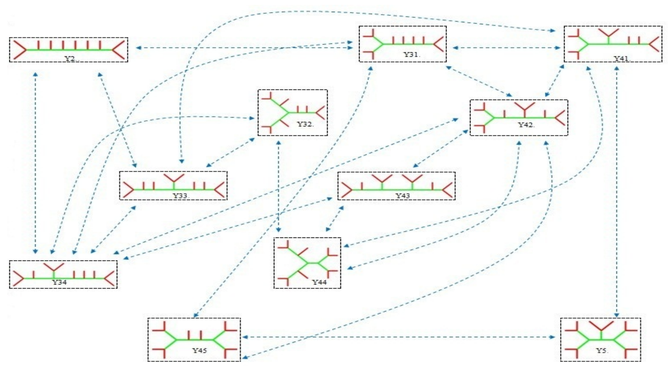
A complete tree space of n taxa consists of m triangles/tetrahedra of $(n-3)$ dimension, where m is the total number of possible unrooted trees formed by n generated species. Each edge / surface is shared by 3 triangles / tetrahedra, which indicates a neighboured degenerate tree. Each single tree, with specific order of species and internal edge lengths, can be expressed by a point inside the tree space.



The space of random walk we consider can be discrete / continuous; bounded / unbounded; cubical / triangular etc, all with different controlled parameters. Different properties of random walk will be applied in suitable cases.

Applications on 10 Generated Species

A non-degenerate evolutionary tree of n species has $(n-3)$ internal edges and $(n-2)$ internal vertices of degree 3. A random walk is generated by the following method: when the length of one internal edge shrinks to 0, the tree becomes degenerate with 6 non-zero internal edges and one internal vertex with degree 4. By 2-2-splitting the 4 edges joined to that vertex and by extending the new internal edge, the tree becomes non-degenerate again. A diagram of adjacencies (as shown below) can be encoded in a matrix of probabilities which produce the random walk, whose eigenvalues and eigenvectors will be used in estimating different models of evolution.



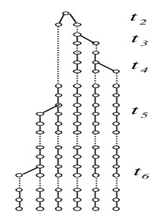
As shown above, the 11 shapes of unrooted trees can be classified into 4 groups diagonally by the number of '**Y-ends**' in each graph. It is impossible to move from one tree to another when their numbers of '**Y-ends**' are differ by 2 or 3. Every rooted tree belongs to a family of unrooted tree, which will be used in further results and analysis.

Results

In the coalescent model of evolution, the trees are built from the timeline of branching. Tree shapes are predicted by the time to coalescence. Inductive steps are taken to separate the species at the bottom. [3]

In the model of doubling, the number of species grows exponentially with a growth rate of 100% in each generation. The analysis is to select n species randomly from the final generation (generate n leaves at the bottom of a binary tree).

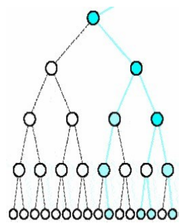
By taking the dot product of the eigenvectors from the S_{10}-invariant random walk and the theoretical distribution vectors from the two models of evolution, we obtain the two sets of angles as below:



Cosine values of angles from dot products:

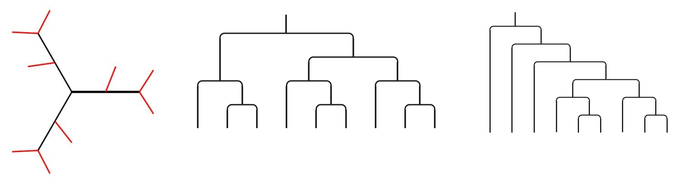
$\Longleftarrow$ Coalesecnt Model:
-0.8858, -0.3641, -0.2407, -0.1498, 0.0121, -0.1110, -0.0259, 0.0011, 0.0065, -0.0185, 0.0040

Doubling Model: $\Longrightarrow$
-0.7818, -0.5015, -0.3017, 0.1126, 0.0415, -0.1542, -0.0577, -0.0075, 0.0418, -0.0445, 0.0291



Discussion

- In the random walk, the unrooted tree could reach higher degenerate level (with less than $n-4$ internal edges). This can only happen in the discrete model. In this case, the random walk would have more freedom than those in continuous model.
- 3-fold symmetry only exists in the unrooted trees when the number of species is a multiple of 3, but does not exist in rooted binary trees. Such repeated calculations may affect the final result.



Conclusion

- Evolutionary trees can be simulated by a random walk on the tree space.
- For certain fixed species, converting rooted trees to unrooted ones will be used for further analysis.
- By working on the transition matrices and comparing probability distributions of different evolution models, the eigenvectors and eigenvalues will provide information to different geometries. These will also be applied to real data.

References

[1] Hodge, T. Cope, M. (2000) A myosin family tree.

[2] Browning, S.R. (2006) Multilocus association mapping using variable-length markov chains.

[3] Hudson RR (1991) Gene genealogies and the coalescent process.