Phylogenetic Networks: Inferring 3-Cycle Networks

EMBL-EBI – Goldman Group

Anthony Zhao

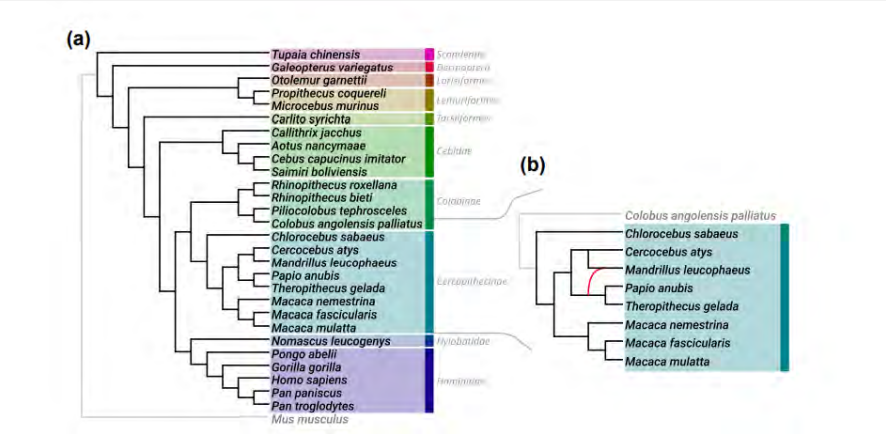
Funded by CMP, EMBL-EBI

Supervised by Samuel Martin

EMBL-EBI

Phylogenetic Networks

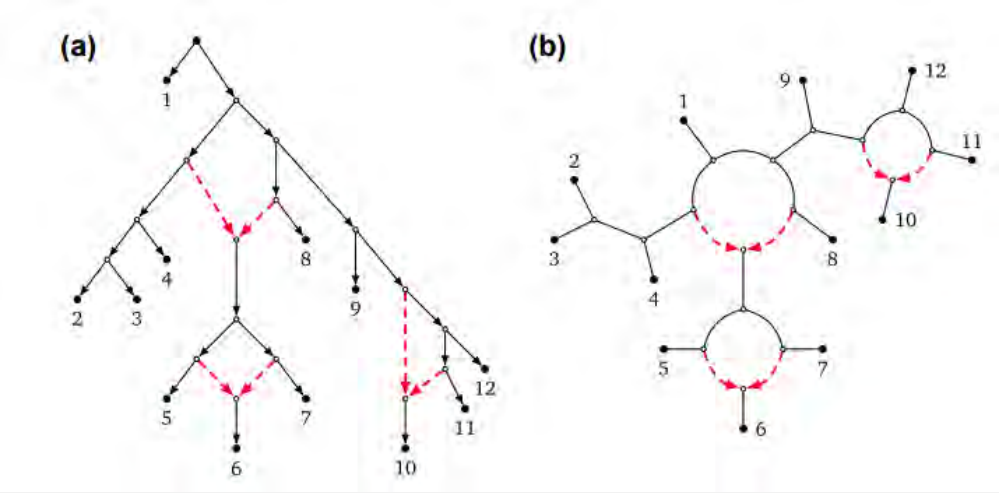
- Directed graphs that represent the evolutionary history of a group of individuals e.g a group of related species
- A common problem is to reconstruct the phylogenetic network given data on the leaves



EMBL-EBI

Semi-directed phylogenetic networks

- Reticulation edges describe events such as hybridization which may occur when different species combine to give a hybrid offspring
- Semi-directed phylogenetic networks are the unrooted graphs with only reticulation edges still directed



EMBL-EBI

Substitution based models

- We can be given the data on the individuals representing the leaf nodes in the form of a sequence of DNA
- We can align these to give a multiple sequence alignment (MSA)

Sequence A	A	G	T	T	G	A	C	T	T	C	T	C	A	G	G	T	A	T	T
Sequence B	A	G	G	T	A	A	C	T	T	C	A	G	A	T	G	A	A	A	T
Sequence C	A	G	G	T	C	A	C	–	–	G	A	C	A	G	G	C	A	T	T
Sequence D	A	G	G	T	C	A	C	–	–	G	A	C	A	G	G	C	A	–	T
Sequence E	A	G	G	T	C	A	C	T	T	G	A	G	A	–	G	C	A	–	T
Sequence F	A	G	G	T	C	A	C	T	T	G	A	C	A	G	G	C	A	T	T

- A substitution-based model is a Markov Model that assigns a transition matrix of probabilities along each edge

$$Q^{JC} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

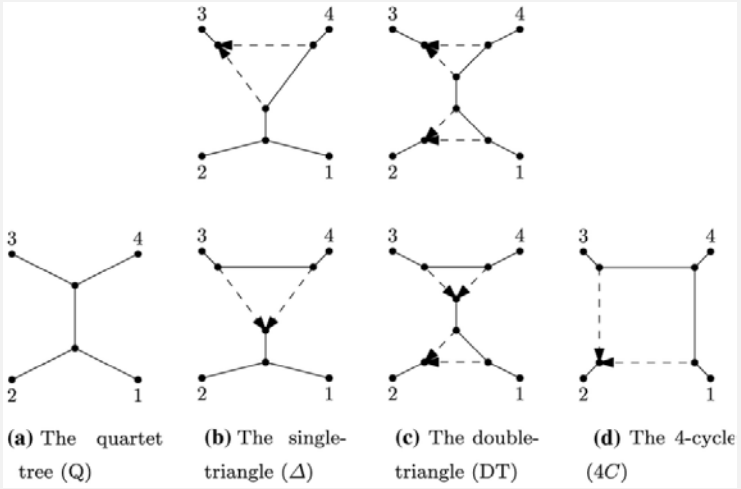
Jukes Cantor Substitution Model

- It is known that the possible 4-leaf networks determine uniquely the combined semi-directed ‘level-1’ phylogenetic network

Problem

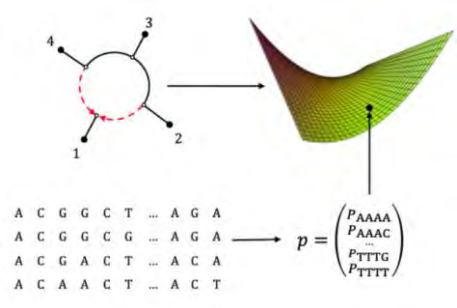
- We are given an MSA of 4 sequences of certain length
- We assume the 4 sequences are related and we want to determine a 4-leaf network to describe these relations
- We assume the Jukes-Cantor substitution model on the network.

- We want to deduce what topology the alignment comes from
- We focus on deciding whether the network is a single triangle



Algebraic Statistics approach

- By designating a root node and summing over products of probabilities we have a probability distribution for each of the 4 leaves e.g we assign probabilities of seeing each of {A, C, G, T} at each leaf
- For 4 leaves we then have 4*4*4*4 = 256 possible outcomes and the probability space of all such outcomes defines an algebraic surface called a variety
- We can find a set of polynomials that generate the variety which allows us to work with the variety easily and a point (vector) lies on the variety if all each of the generating polynomials evaluate to 0 at the point



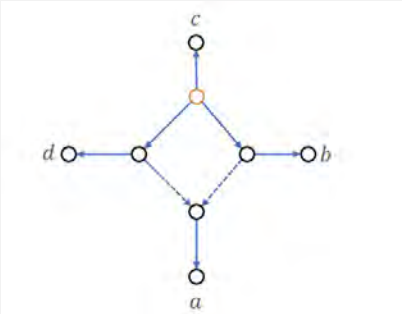
The scoring method

Input an MSA from a 4-cycle network e.g simulated

Calculate the induced leaf pattern probabilities to get a vector of probabilities

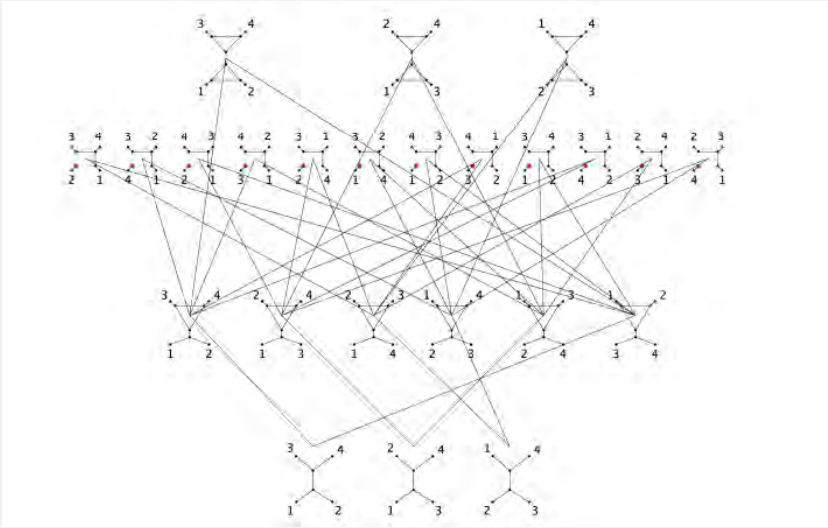
Evaluate the vector on the generating polynomials (of the same degree) of the associated varieties of the 12 distinct 4-cycle networks

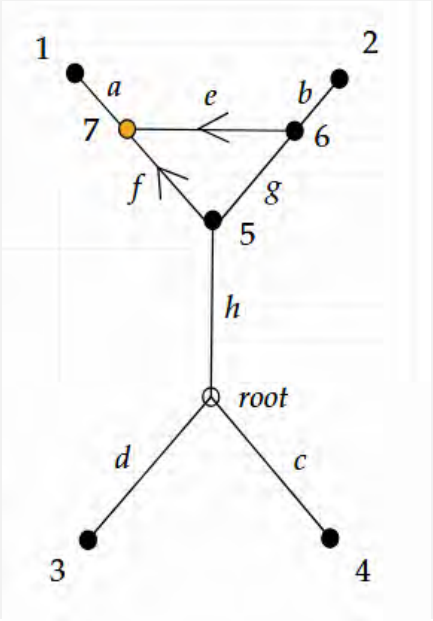
Apply the L1 norm to get 12 ‘scores’



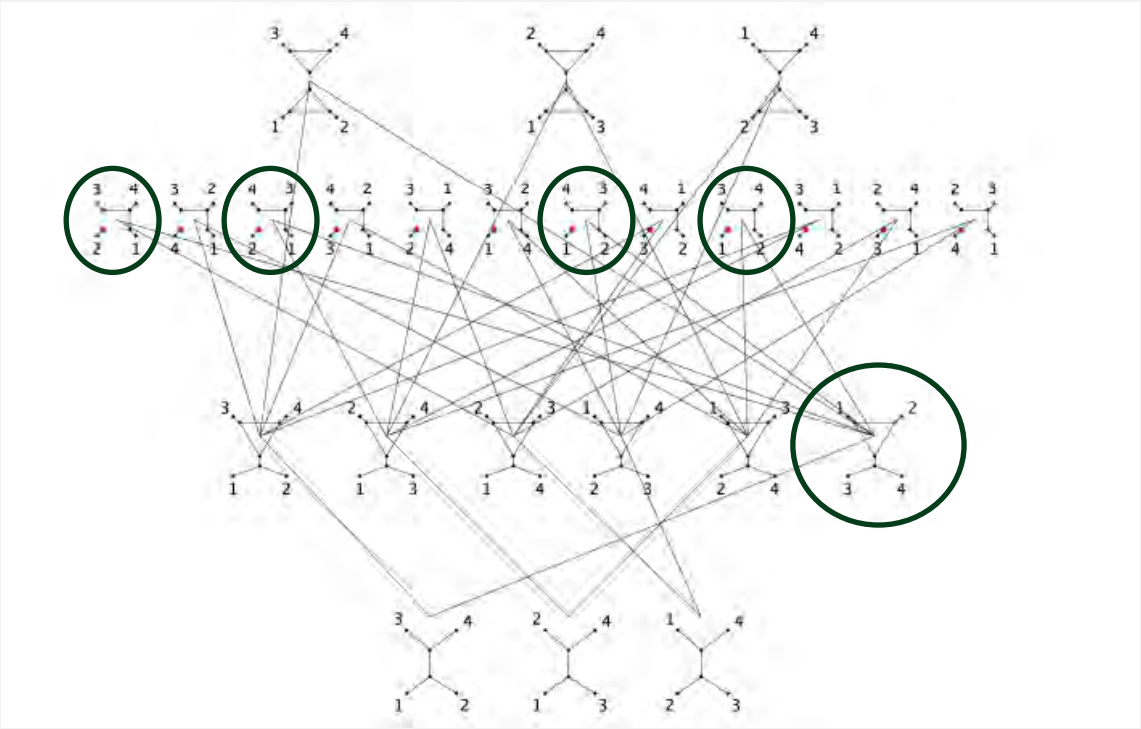
Inspiration for finding 3-cycle networks

- There is inclusion of varieties of 4 leaf networks, thus for a given 3 cycle network we would expect the score to be 0 for 4 of the 4-cycle networks (inclusion indicated by a line from bottom to the top)





Our inclusions



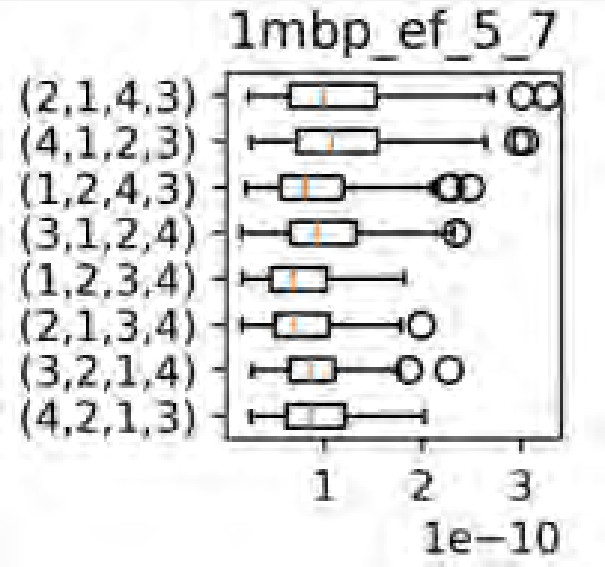
(2,1,4,3)
(1,2,4,3)
(2,1,3,4)
(1,2,3,4)

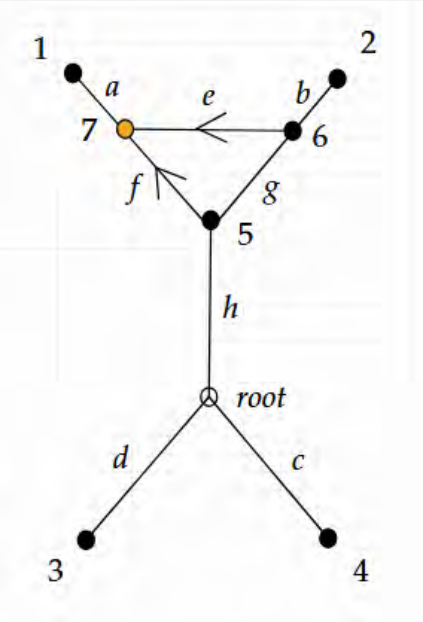
EMBL-EBI

Visualising the scores

Title describes the parameters used for the 3-cycle triangle network - the probability of which edge to cross

The 4-leaf quartets

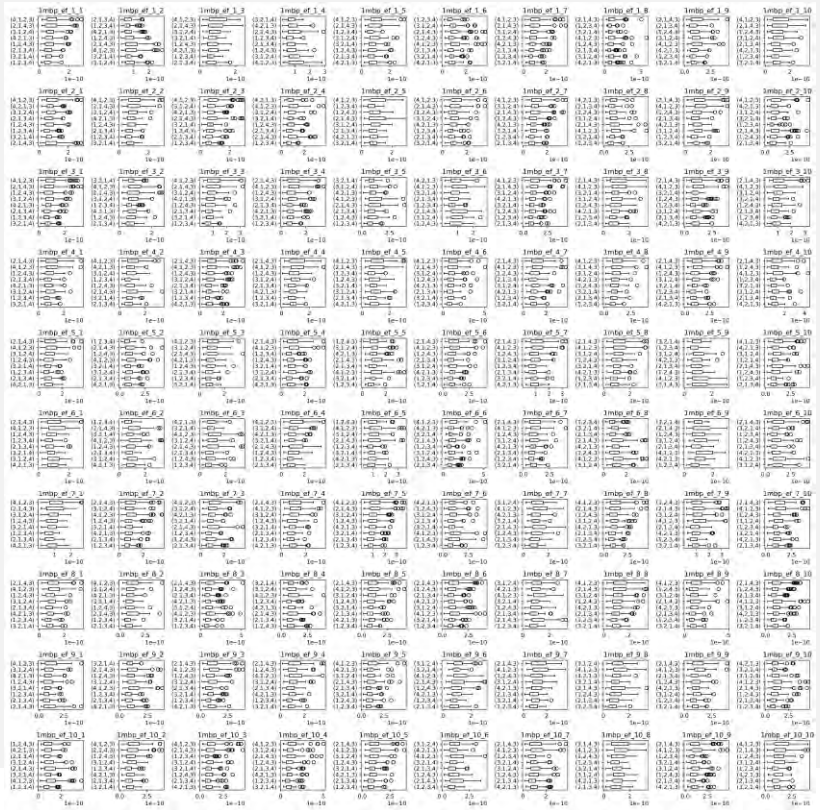


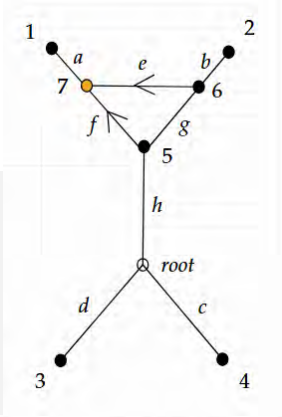


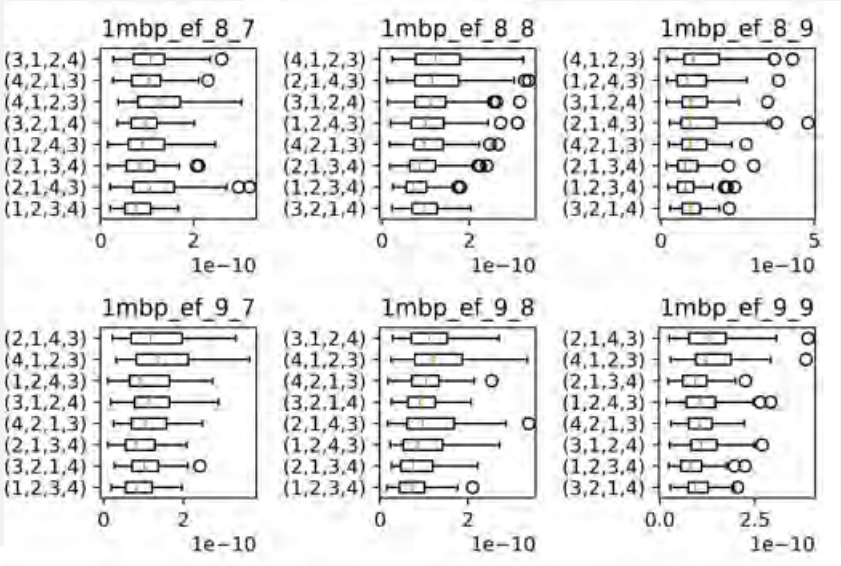
Box plots of scores, e_f

f

e

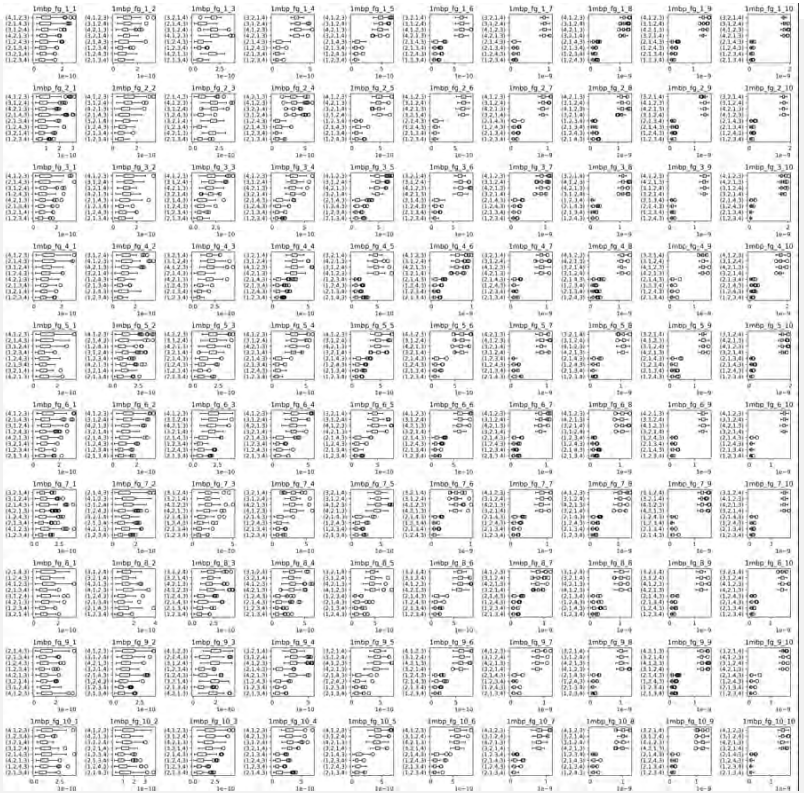


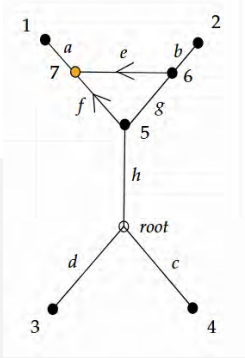


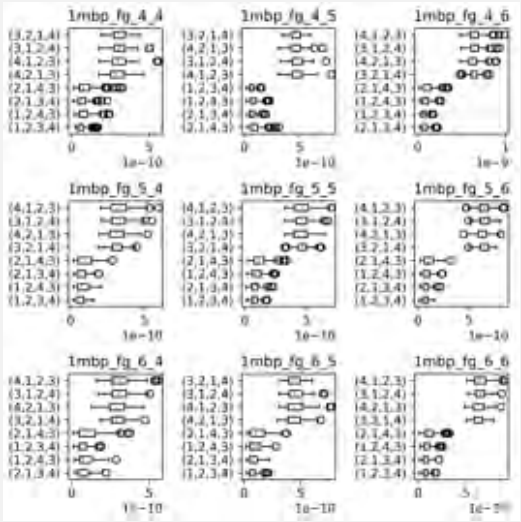


EMBL-EBI

F_g

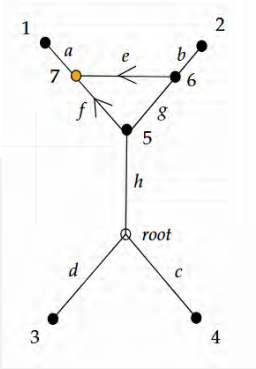


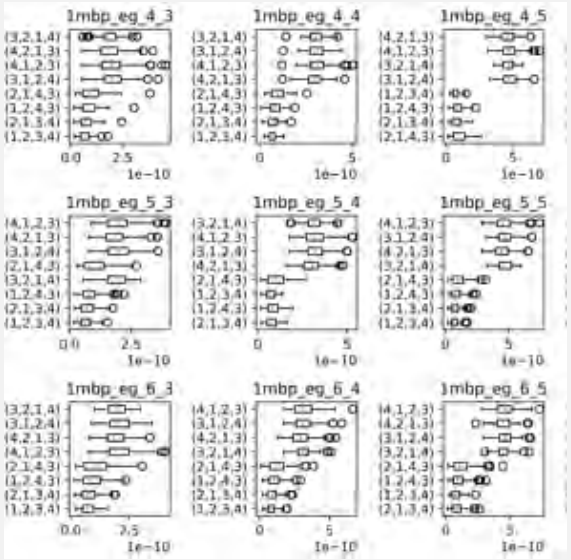

1
2
a
e
b
7
6
f
g
5
h
root
d
c
3
4


1mbp_fg_4_4
1mbp_fg_4_5
1mbp_fg_4_6
1mbp_fg_5_4
1mbp_fg_5_5
1mbp_fg_5_6
1mbp_fg_6_4
1mbp_fg_6_5
1mbp_fg_6_6

EMBL-EBI

E_g

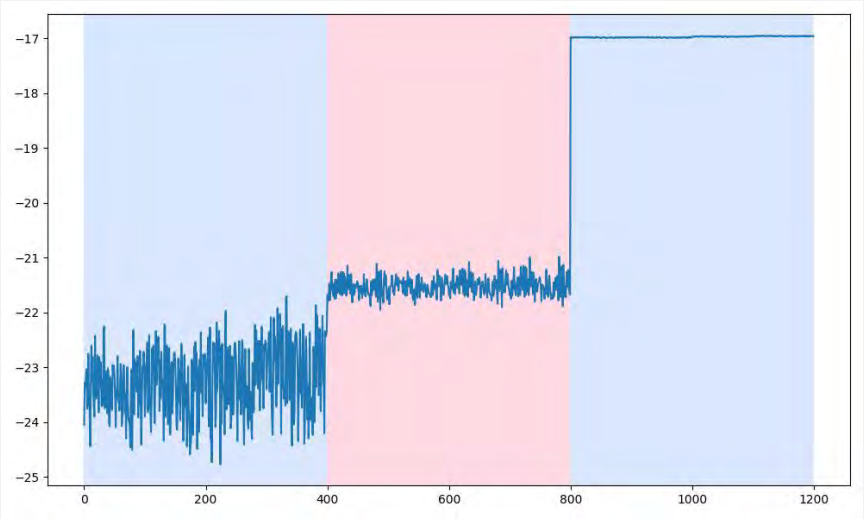

1
2
a
e
b
7
6
f
g
5
h
root
d
c
3
4


1mbp_eg_4_3
1mbp_eg_4_4
1mbp_eg_4_5
1mbp_eg_5_3
1mbp_eg_5_4
1mbp_eg_5_5
1mbp_eg_6_3
1mbp_eg_6_4
1mbp_eg_6_5
1e-10

EMBL-EBI

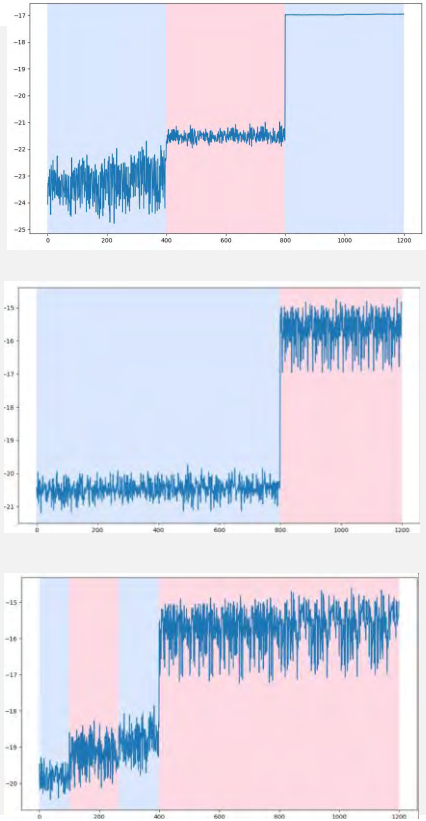
A different plot

- We can obtain a different plot that offers some more visual context
- We fix a given MSA
- Of the 12 4-cycles, suppose we simulate 100 scores for each e.g by bootstrapping the DNA alignment
- We sort the means then plot the log scores in 100 score blocks in the sorted order



Detecting the 3-cycle

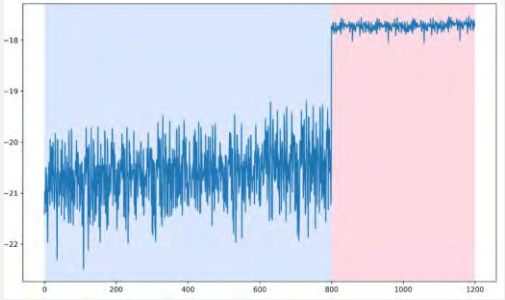
- We only have a single score for each of the 12 4-cycle topologies since we are given a single alignment, so we use bootstrap methods to estimate other simulated alignments
- With good parameters we can see there is distinct differences between the plots for different topologies
- There are changepoint detection algorithms to determine where ‘jumps’ occur
- A simpler statistic is the proportion of the score/variance seen for the 4 lowest scoring topologies



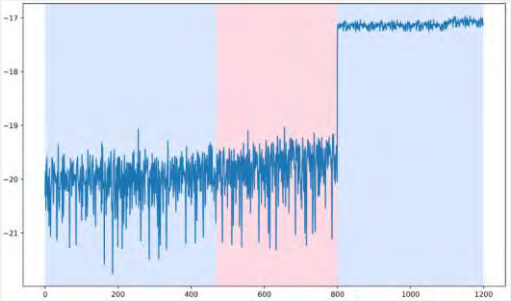
Problems

- There is a large parameter space and behaviour differs a lot across this space
- For certain parameters very difficult to find a distinguishing feature since values/plots from a 3-cycle topology may look very similar to those from a 4-leaf tree

3-cycle network



4 Leaf Tree

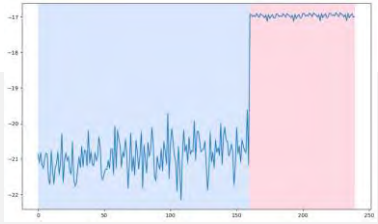


EMBL-EBI

Further analysis ideas

- Apply perturbations to the alignment e.g systematically resampling the output on leaf 1, 2 if leaf 1 and 2 share the same value
- Bootstrap the result as before to get a different plot and scores with possibly different breakpoints
- Hope to see consistent and different (across different topologies) that can help with identification
- However, it is not clear if these correspond to anything physical and it is difficult to have intuition for what to expect

Base plot (3-cycle)



Perturbed plots