CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
College of Natural Sciences

The Impact of Model Misspecification on Tree and Network Inference from Quartets

IMSI

Hector Baños
Department of Mathematics
Tuesday, August 12, 2025

Joint Work

Vu Dinn

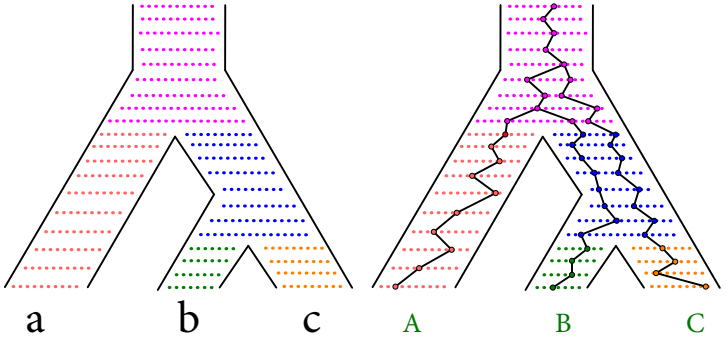
UNIVERSITY OF DELAWARE

ICERM

NSF

The Multispecies Coalescent Model

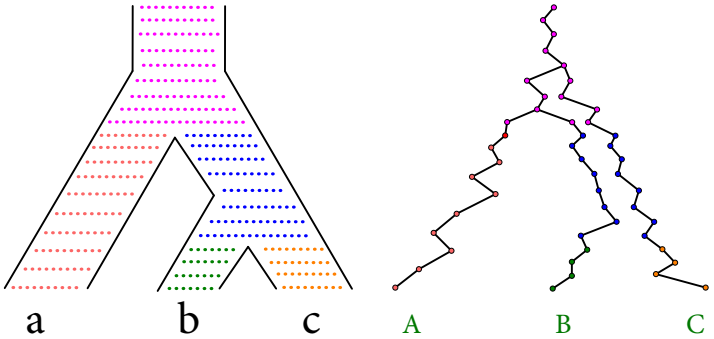
Rannala and Yang '03



The multi-species coalescent describes a stochastic model of gene tree generation accounting for gene tree incongruence.

The Multispecies Coalescent Model

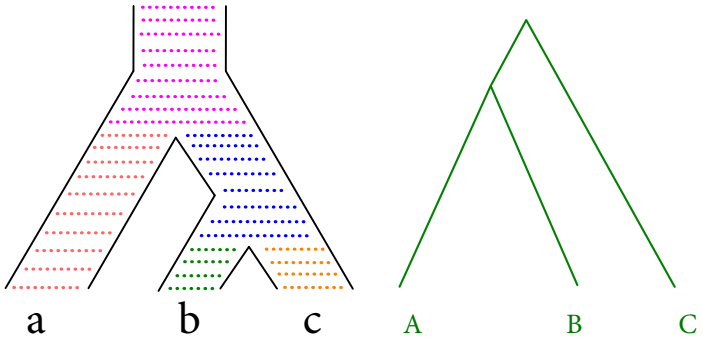
Rannala and Yang '03



The multi-species coalescent describes a stochastic model of gene tree generation accounting for gene tree incongruence.

The Multispecies Coalescent Model

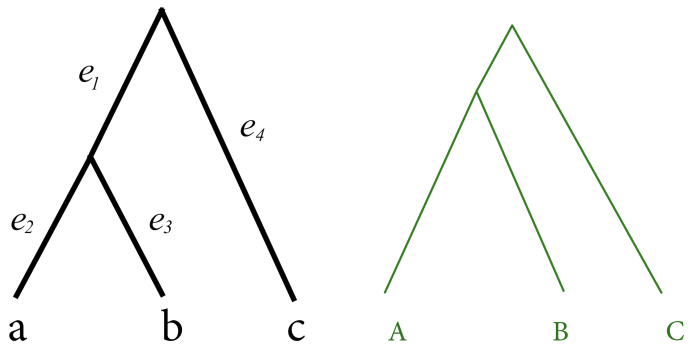
Rannala and Yang '03



The multi-species coalescent describes a stochastic model of gene tree generation accounting for gene tree incongruence.

The Multispecies Coalescent Model

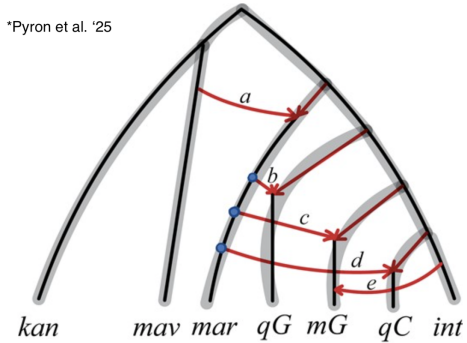
Rannala and Yang '03



The multi-species coalescent describes a stochastic model of gene tree generation accounting for gene tree incongruence.

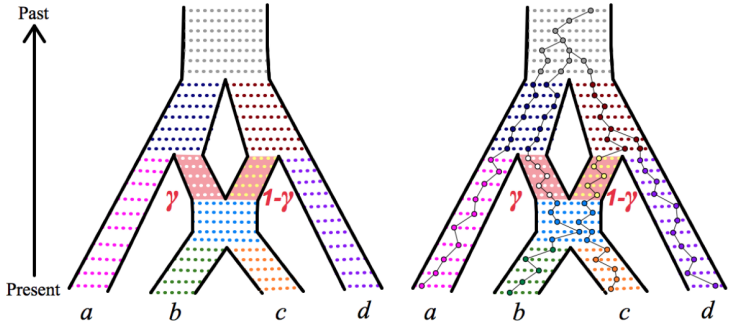
Hybridization

Hybridization occurs when two species merge genetically to create a new one.



The Network Multispecies Coalescent Model

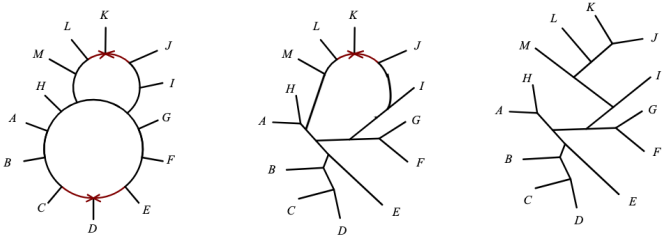
Meng & Kubatko '09 - Degnan, Yu, & Nakhleh '12



The network multi-species coalescent describes a stochastic model of gene tree generation in the presence of hybridization.

Model Misspecification

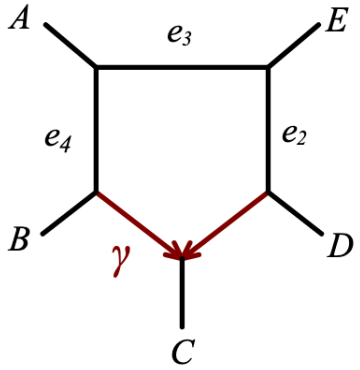
Question: How do inference methods behave when applied to data originating from a more complex topological structure than the one assumed by the method?



- Solís-Lemus, Yang & Ané. *Inconsistency of species tree methods under gene flow '16.*
- Long & Kubatko. *The effect of gene flow on coalescent-based species-tree inference '18.*
- Pang & Zhang. *Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time '22.*

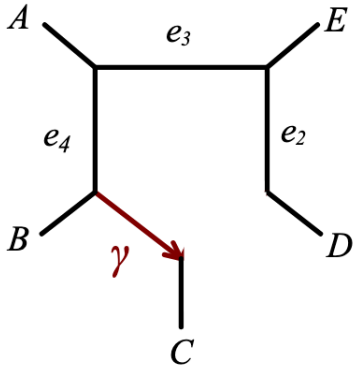
Network's Displayed Trees

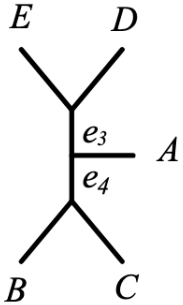
A displayed tree is obtained by removing exactly one hybrid edge from each hybridization event.



Network's Displayed Trees

A displayed tree is obtained by removing exactly one hybrid edge from each hybridization event.

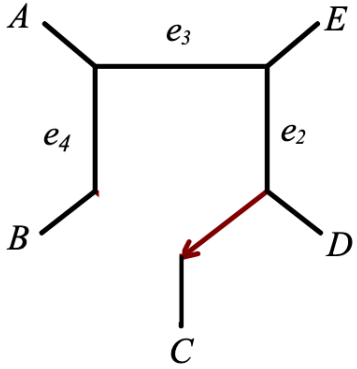


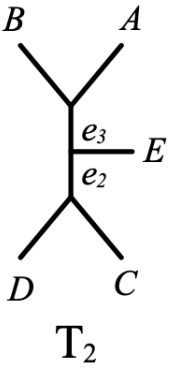


T_1

Network's Displayed Trees

A displayed tree is obtained by removing exactly one hybrid edge from each hybridization event.



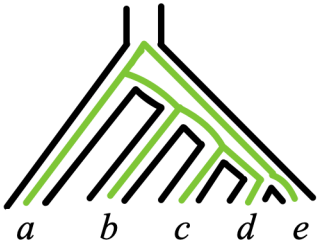


Anomalous tree

In the tree setting (i.e no hybridization),

Definition (Anomalous tree)

A species tree is said to be **anomalous** *under the MSC if the most probable unrooted gene tree does not match the unrooted species tree topology.*



Matches the unrooted species tree

$$P\left(\begin{smallmatrix} c & & b & & a \\ & d & & e & \end{smallmatrix}\right) > P\left(\begin{smallmatrix} a & & c & & d \\ & b & & & e \end{smallmatrix}\right)$$

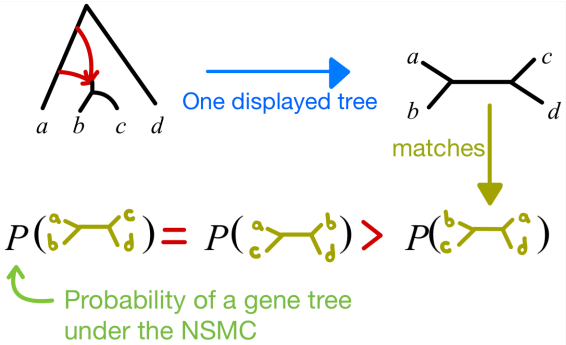
Probability of the unrooted gene tree under the NSMC

Anomalous network

Definition (Anomalous network - Ané, et al.)

Let N^+ be a rooted metric network on X. N^+ is **anomalous**† *under the NMSC if there is a gene tree whose topology is not displayed by N, that is more probable than a gene tree matching the topology of a displayed tree of N.*

For example, Solís-Lemus and Ané '16



†Differs from Pang & Zhang '22

Anomalous networks

Anomalous networks are a problem for inference methods and identifiability results.

Effects of anomalies on tree inference methods:

- Solís-Lemus, Yang, Ané. *Inconsistency of species-tree methods under gene flow* '16.

Identifiability problems for some networks:

- B. *Identifying species network features from gene tree quartets under the coalescent model* '19.

Characterizations of 4-taxon anomalous networks:

- Ané, Fogg, Allman, B., Rhodes. *Anomalous networks under the multispecies coalescent: theory and prevalence* '24.

Identifiability results which specifically require non anomalous scenarios:

- Rhodes, B., Xu, Ané. *Identifying circular orders for blobs in phylogenetic networks* '25.

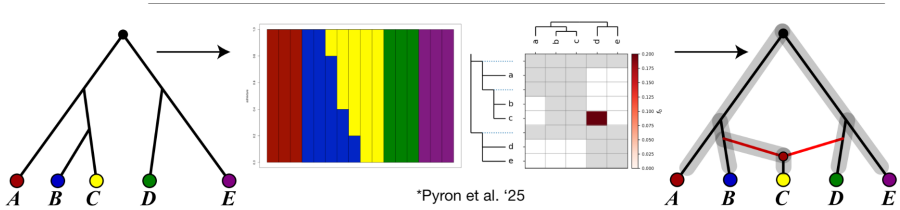
For the rest of the talk we assume NO quartet anomalous scenarios

Network Inference

Many network inference algorithms either infer "simple" networks or are not easily scalable.

A reasonable approach avoiding either oversimplification of the network or scalability issues is to:

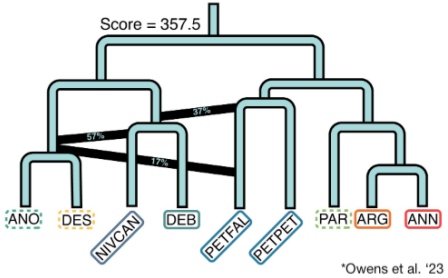
1) Infer a "displayed" tree, representing underlying tree-like relationships among species
2) then inferring hybridization events on top of it using different techniques (for e.g. `Dsuite`)



*Pyron et al. '25

The use of ASTRAL for Network Inference

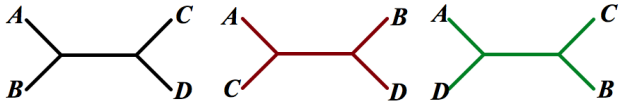
Many works have used ASTRAL for Inferring a "displayed" tree,



Owens et al., 2023; Zhou et al., 2022; Singh et al., 2022; Jensen et al., 2023; Sanderson et al., 2023; Ciezarek et al., 2024; Scherz et al., 2022; Yang et al., 2023; Lopes et al., 2023; Feng et al., 2022; Bernhardt et al., 2020; DeRaad et al., 2022; Herrig et al., 2024; Zhang et al., 2023; Zhou et al., 2023.

Quartets

A *quartet* is a binary tree on 4 taxa. There are 3 different quartet trees on 4 taxa:



Given a sample of gene trees, for any subset of four taxa, each gene tree displays exactly one quartet on those taxa.

ASTRAL is a powerful and widely used tool for species tree inference, known for its computational speed and consistency under the MSC (Mirarab, Warnow, et al. '14.).

Given a collection of gene trees $\mathcal{T}_m = \{T_1, T_2, \ldots, T_m\}$, it searches for a species tree $\mathbb{T}$ such that

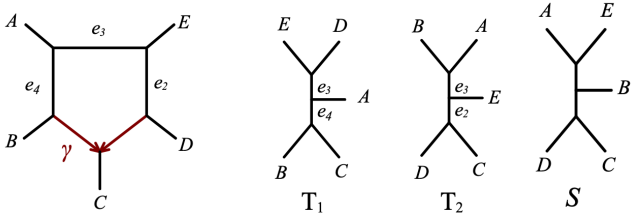
$$A(\mathbb{T}) = \frac{1}{m} \sum_{q \in Q(\mathbb{T})} w_m(q, \mathcal{T}_m) \text{ is } \textbf{maximized},$$

where $Q(\mathbb{T})$ is the set of quartet trees induced by $\mathbb{T}$ and $w_m(q, \mathcal{T}_m)$ is the number of the trees in $\mathcal{T}_m$ that induce quartet topology q.

Theorem (Allman, Degnan, Rhodes '11)

Under the MSC, the most probable quartet gene tree has the same topology as the quartet species tree.

ASTRAL under Misspecification



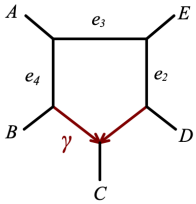
Theorem (Dinh, B.)

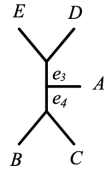
Let N be the semi-directed network on 5 taxa above and let let $y_i = \exp(-e_i)$. If

$$(1-y_2) > \frac{\gamma}{1-\gamma}(1-y_3y_4) \quad \textit{and} \quad (1-y_3) < \min\left(\frac{\gamma}{2-\gamma}(1-y_4), \frac{1-\gamma}{1+\gamma}(1-y_2)\right),$$

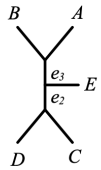
under the NMSC, the tree $\mathcal{S}$ above, which is not displayed by N, has a higher expected ASTRAL score than the displayed trees of N.

ASTRAL under Misspecification

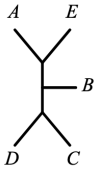




T_1



T_2



S

Quartet in T_1	Quartet in T_2
AB|**DE**	**AB**|**DE**
$AD\|BC$ $BC\|DE$	**AB**|**CD** **BE**|**CD**
AE|**BC** $AC\|DE$	$AB\|CE$ **AE**|**CD**

$$A(T_1) = P(\boldsymbol{AB}|\boldsymbol{DE}) + P(AD|BC) + P(BC|DE) + P(\boldsymbol{AE}|\boldsymbol{BC}) + P(AC|DE)$$
$$A(T_2) = P(\boldsymbol{AB}|\boldsymbol{DE}) + P(\boldsymbol{AB}|\boldsymbol{CD}) + P(\boldsymbol{BE}|\boldsymbol{CD}) + P(AB|CE) + P(\boldsymbol{AE}|\boldsymbol{CD})$$
$$A(\mathcal{S}) = P(AE|BD) + P(\boldsymbol{AB}|\boldsymbol{CD}) + P(\boldsymbol{BE}|\boldsymbol{CD}) + P(\boldsymbol{AE}|\boldsymbol{BC}) + P(\boldsymbol{AE}|\boldsymbol{CD})$$

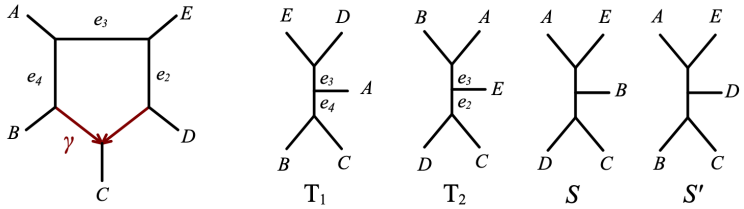
ASTRAL under Misspecification

Corollary (Dinh, B.)

Let N be the semi-directed network on 5 taxa below and let let $y_i = \exp(-e_i)$. If

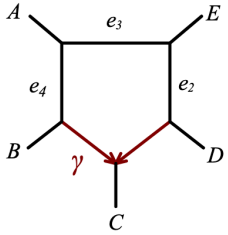
$$(1-y_4) > \frac{\gamma}{1-\gamma}(1-y_3y_2) \quad \textit{and} \quad (1-y_3) < \min\left(\frac{\gamma}{2-\gamma}(1-y_4), \frac{1-\gamma}{1+\gamma}(1-y_2)\right),$$

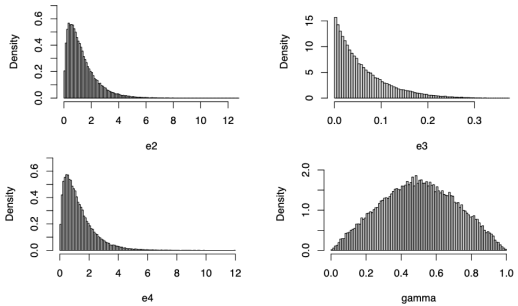
under the NMSC, the tree $\mathcal{S}'$ below, which is not displayed by N, has a higher expected ASTRAL score than the displayed trees of N.



Simulations

We sampled 10^6 sets of parameters from the network below.



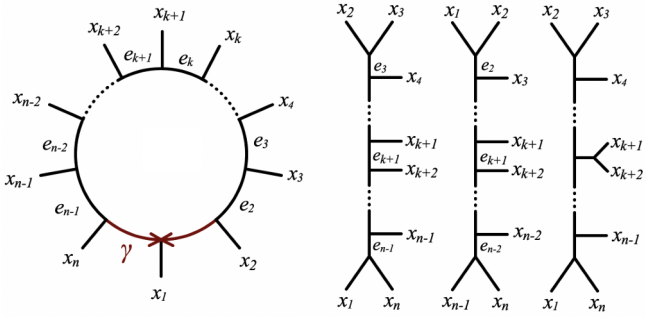


Range for γ	(0,1)	(0.2,0.8)	(0.4,0.6)	(0,0.1)∪(0.9,1)
Proportion of parameters in Θ	0.06	0.08	0.10	0.01

This is Behavior can be Generalized

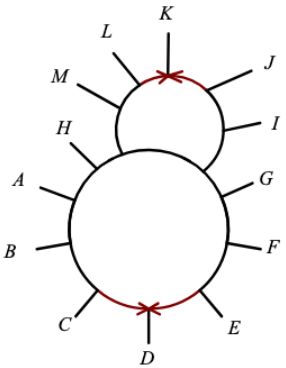
For sets of parameters satisfying the inequalities, we simulated gene trees using `PhyloCoalSimulations` (Ané, Fogg, Allman '24).
In all these, `ASTRAL` did not recover a displayed tree.

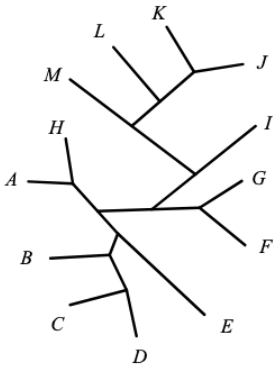
We generalized the results for bigger networks:



This is Behavior can be Generalized

We showed this behavior occurs in more complex networks





Not displayed!

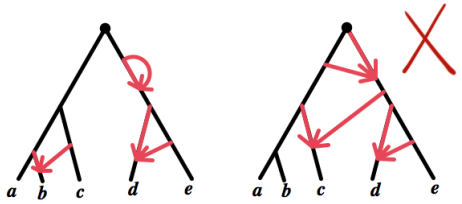
Why things behave like this?

This is a problem of misspecification on quartet-based methodologies not ASTRAL itself. For example:

SNaQ is a quartet-based method for network inference under the NMSC. It uses a pseudo-likelihood framework. One key assumption is that the network is **level-1**

Definition

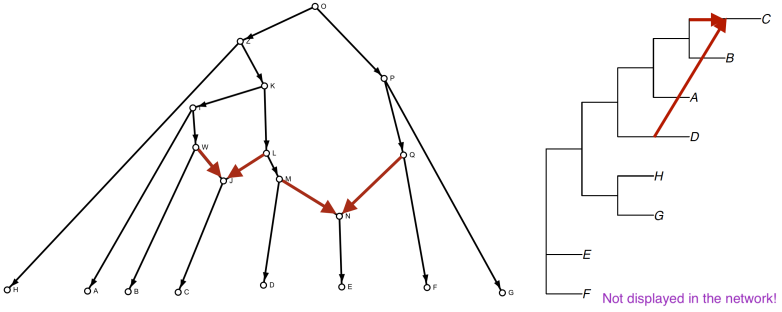
A network $\mathcal{N}$ is **level-1** if no pair of cycles in $\mathcal{N}$ share an edge.



level-1 Not level-1

SNaQ

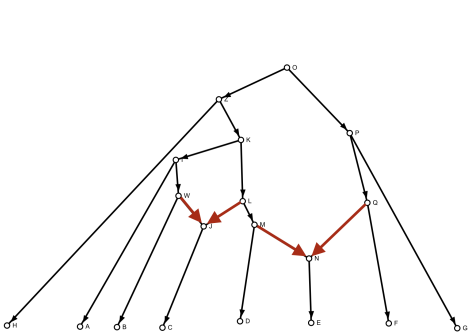
We have preliminary results showing that same issues can occur for `SNaQ` when inferring networks.

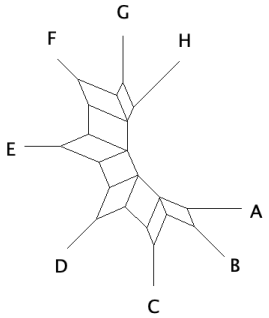


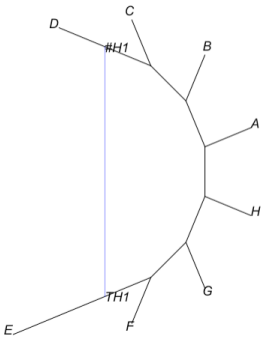
Along the Solís-Lemus Lab, we are exploring the behavior of SNaQ via a simulation study (In progress).

NANUQ and NANUQ+

NANUQ and NANUQ+ are quartet-based level-1 network inference methods (Allman, B., Rhodes, Wicke).







We suspect quartet-based methods such as SVDquartets and PhyNEST have similar behavior (joint work with Dinh, Allman, and Rhodes).

Can we avoid this issues?

We believe that this behavior can be extended.
That is, for $m > k$, there are parameters such that inferring a level-k network from data that came from a level-m network is problematic.

How could we overcome this?

- Displayed tree inference method

Along Pyron and some colleagues we have a 'proof of concept' method
Systematic Biology, Volume 74, Issue 1, January 2025, Pages 124–140
(Not easily scalable, theoretical limitations)

Quartet Concordance Factors

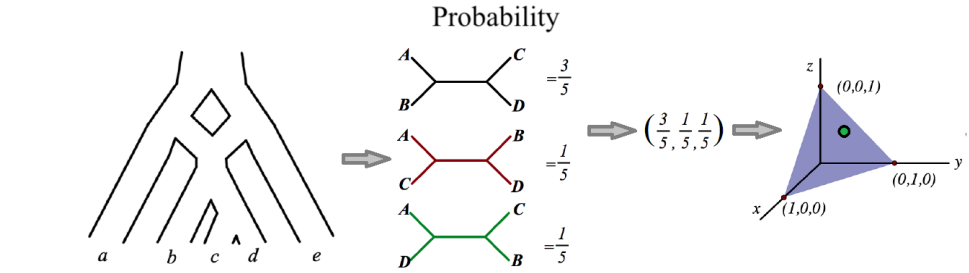
Following the approach of C. Solís-Lemus and C. Ané 2016

Definition

Let $\mathcal{N}$ be a species network. Then

$$CF_{abcd} = (CF_{AB|CD}, CF_{AC|BD}, CF_{AD|BC})$$

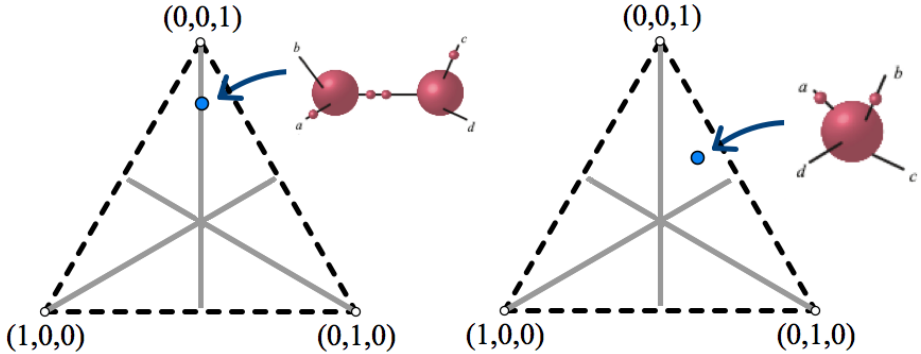
is the triplet of probabilities of gene trees quartets under the NMSC.



Cut and Non-cut Concordance Factors

The concordance factor CF_{abcd} is:

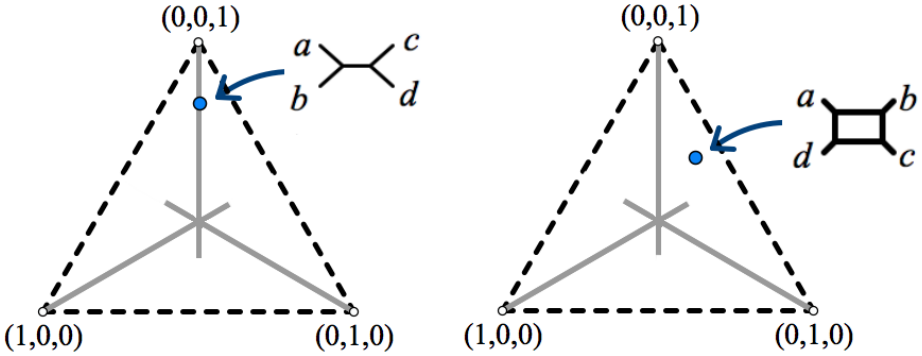
- a **cut** CF if two of its entries are equal, in addition the third is distinct, or
- a **non-cut** CF if it is not cut.



Cut and Non-cut Concordance Factors

The concordance factor CF_{abcd} is:

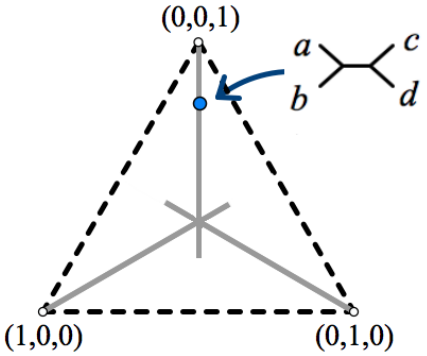
- a **cut** *CF* if two of its entries are equal, in addition the third is distinct, or
- a **non-cut** *CF* if it is not cut.

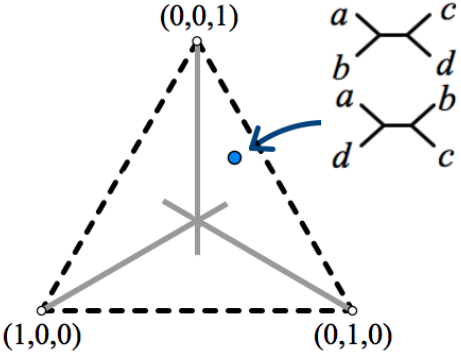


Cut and Non-cut Concordance Factors

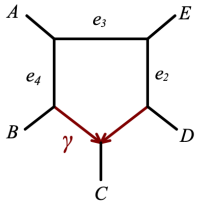
The concordance factor CF_{abcd} is:

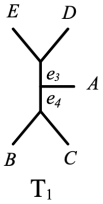
- a **cut** *CF* if two of its entries are equal, in addition the third is distinct, or
- a **non-cut** *CF* if it is not cut.



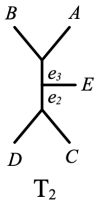


Cut and Non-cut Concordance Factors

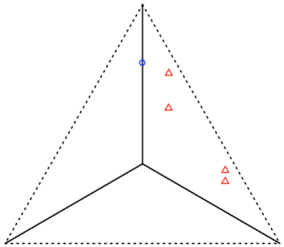




T_1



T_2



$$A(T_1) = P(\boldsymbol{AB}|\boldsymbol{DE}) + P(AD|BC) + P(BC|DE) + P(\boldsymbol{AE}|\boldsymbol{BC}) + P(AC|DE)$$
$$A(T_2) = P(\boldsymbol{AB}|\boldsymbol{DE}) + P(\boldsymbol{AB}|\boldsymbol{CD}) + P(\boldsymbol{BE}|\boldsymbol{CD}) + P(AB|CE) + P(\boldsymbol{AE}|\boldsymbol{CD})$$
$$A(\mathcal{S}) = P(AE|BD) + P(\boldsymbol{AB}|\boldsymbol{CD}) + P(\boldsymbol{BE}|\boldsymbol{CD}) + P(\boldsymbol{AE}|\boldsymbol{BC}) + P(\boldsymbol{AE}|\boldsymbol{CD})$$

Future work - Join us!

CSUSB

Vu an I believe that by weighting the gene tree quartets that arise from a **cut** CFs will lead to better displayed tree estimation.

Things to figure out:

- Optimal weight
- Identify the family of networks for which the algorithm is consistent
- Implementation

This is an open invitation

Thank you!

Questions?

Hector Banos
hector.banos@csusb.edu

CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
College of Natural Sciences

ICERM