A Discussion on Phylogenetic Networks

701 Lecture 12/01/2020

Roman Kouznetsov

University of Michigan Department of Statistics

UNIVERSITY OF MICHIGAN

Table of Contents

1. (Long) Introduction
2. Methodology and Application of Phylogenetic Networks
3. Results
4. Conclusion

Terminology

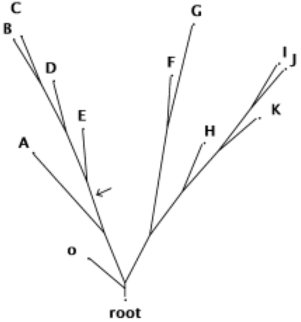
- phylogenetic tree: a tree representing evolutionary history with leaf labels and sometimes branch lengths
- phylogenetic network: a network where branches are edges and nodes are taxa (if branch lengths are available they serve as edge weights)
- split: a partition of taxa into 2 non-empty subsets of leaf nodes
- split network: a network of splits
- phenetic distance (between 2 taxa): the sum of the lengths of the edges along the shortest path between taxa

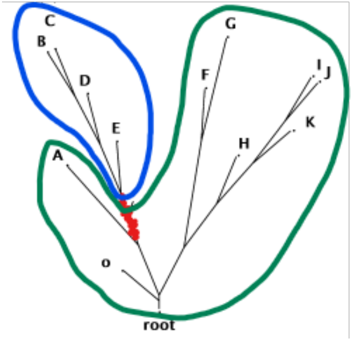
If this is confusing, that's good because this will all be flushed out with examples.

Splits

Recall that a split is just a partition of taxa into 2 non-empty subsets.

This will cause the results to have a tree-like structures that have their own unique taxa.





Split Networks

- A split network is just a network that joins a bunch of splits and their respective weights.
- If we only have one tree, we can use the tree itself for the split network; but, when we have multiple trees, the split network can store the union of their splits at once. **This is how we store a summary of multiple trees in a single structure.**

Let's examine an unlabeled structure just to get familiar with the subject.

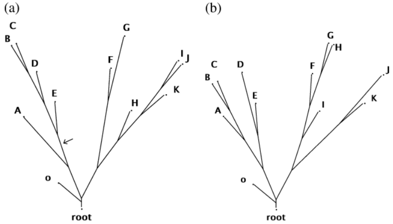
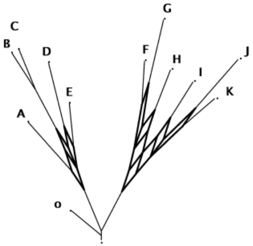


FIG. 2.—Two different trees on the taxon set $X = \{o, A, B, \ldots, K\}$.



Notation

- X: taxon split
- $\mathcal{S}$: set of splits
- $\mathcal{N}$: split network

Network Structure

The formal definition of a split network $\mathcal{N}$ if given below:

Definition

For a given taxon set X, and set of splits $\mathcal{S}$, a split network $\mathcal{N}$ is defined as a connected graph where the nodes are labelled as taxa and the edges are labeled by splits, such that:

1. removing all edges associated with split $S \in \mathcal{S}$ divides $\mathcal{N}$ into 2 connected, but separated components.
2. the edges along any shortest path in N are all associated with different splits (i.e. parallel lines in a split network correspond to the same split as we will see later)

This definition is hard to interpret, so we will explore some examples to get more familiar with split networks.

Basic Example

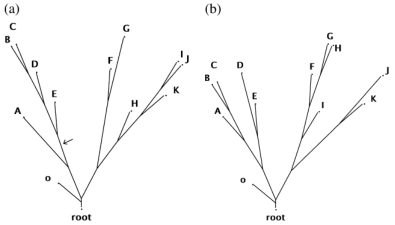
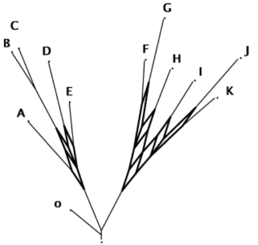
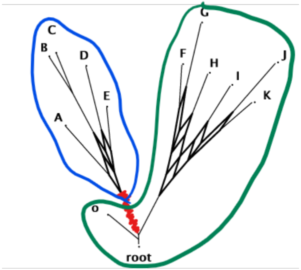


FIG. 2.—Two different trees on the taxon set $X=\{o, A, B, \ldots, K\}$.





Parallel Edges Example Tree 1

Parallel edges in a network correspond to the same split. Here are a few examples of this.

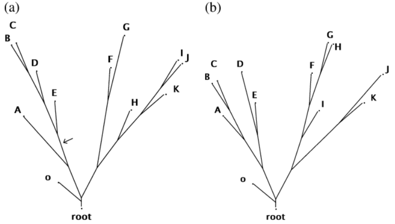
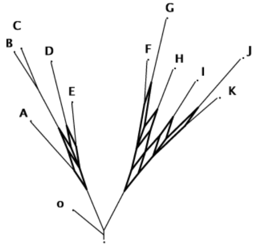
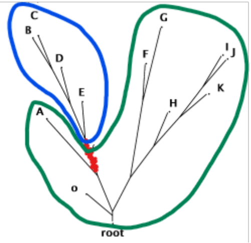
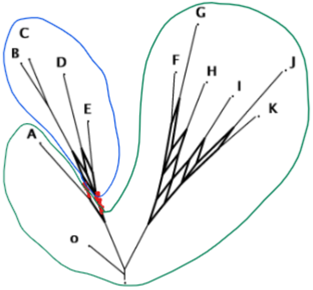


FIG. 2.—Two different trees on the taxon set $X=\{o,A,B,\ldots,K\}$.







Parallel Edges Example Tree 2

Parallel edges in a network correspond to the same split. Here are a few examples of this.

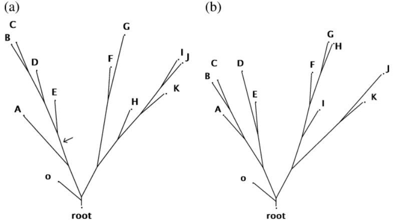
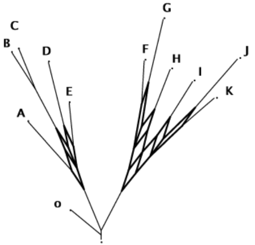
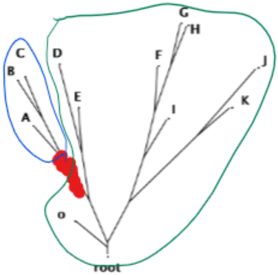
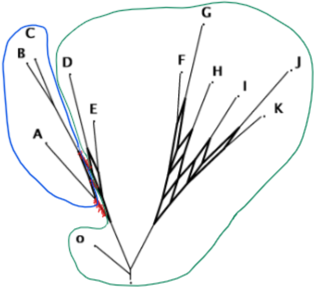


Fig. 2.—Two different trees on the taxon set $X=\{o,A,B,\ldots,K\}$.







Cumulative Representation

The main takeaway is that using additional edges in parallel, can allow for representation of a union of splits.

Red Splits = Tree 1

Purple Splits = Tree 2

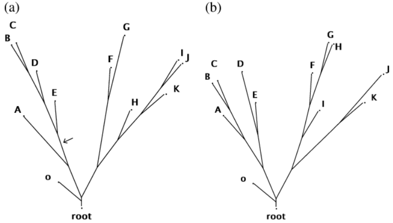
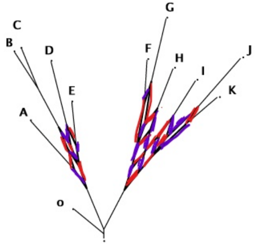


FIG. 2.—Two different trees on the taxon set $X=\{o,A,B,\ldots,K\}$.

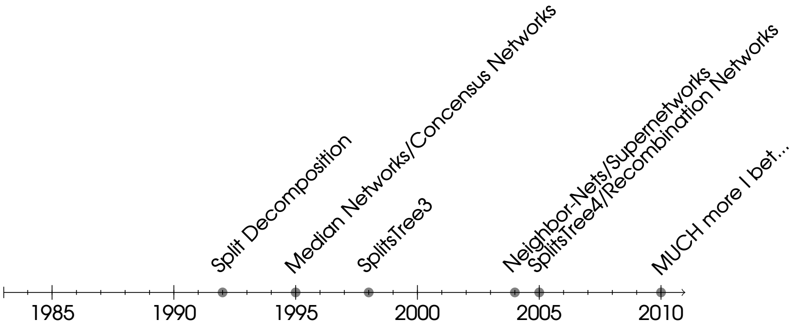


Goal of Split Networks

Q: Why would we ever want to do inference on something that clearly doesn't have a split structure?

A: You don't have to perform inference on values (or trees in this case) that lie in your support. An example you may be more familiar with is saying "the expected number of children is modelled as Poisson with mean 1.86." The support for the Poisson distribution is $\mathbb{N}_0$; 1.86 isn't a part of the support, but we still understand what Poisson with mean 1.86 means (pun intended).

Timeline of Methods



Network Uniqueness

Networks have a unique set of splits.

A set of splits do NOT have a unique network structure.

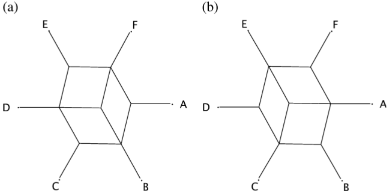


FIG. 5.—Two different representations of the same set of splits.

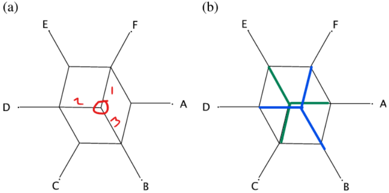


FIG. 5.—Two different representations of the same set of splits.

Representing Multiple Trees

We have established that the purpose of using split networks is to efficiently represent a structure that contains multiple trees.

Q: What techniques are used to create such networks.

A: In previous work, consensus networks were a commonly used class of split networks. The authors of this paper expand this concept to something called a confidence set on trees.

Consensus Networks

The goal of a consensus network is to simply generate a split network in the following way:

Algorithm 1: Consensus Networks

```
Result: A split network is returned.
split_count = 0 ;
ntrees = len(trees) ;
for split in splits do
    for tree in trees do
        if split in tree then
            split_count += 1
        else
    end
    if split_count/ntrees > p then
        Add the current split to the network.
    else
end
```

Confidence Sets

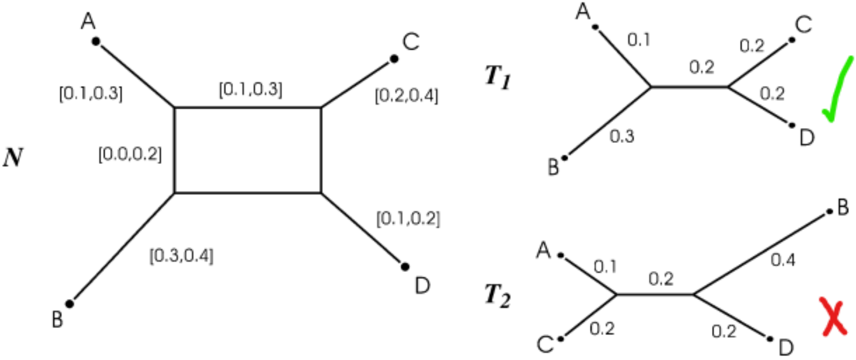
Definition

Say you have a consensus network $\mathcal{N}$ and a tree T. Network $\mathcal{N}$ comes with a complete set of intervals of weights (w) for each split. Then, a tree T is said to be within $\mathcal{N}$ if the following conditions hold:

1. Every split in T is a split in the $\mathcal{N}$.
2. For every split in T, the corresponding branch length is in the interval w_{split} given for the corresponding split in $\mathcal{N}$.
3. Every split not in T means that the corresponding w_{split} in $\mathcal{N}$ contains 0.

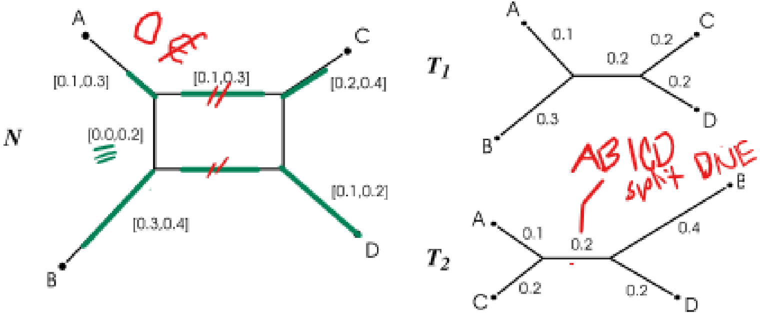
Like the previous set of rules, this is probably very confusing. Let's look at an example!

Confidence Set Example



The next slide presents more details as to why $T_1 \in \mathcal{N}$ but $T_2 \notin \mathcal{N}$.

Confidence Set Example (cont.)



- T_1 is drawn on the network $\mathcal{N}$ in green.
- T_2 is not in $\mathcal{N}$ because Rule 3 is broken. In clearer terms, the split $AB|CD$ does NOT exist on T_2. But, the split $AB|CD$ weight interval in $\mathcal{N}$ doesn't contain 0. This suggests that every tree in $\mathcal{N}$ must have the split $AB|CD$. Therefore, $T_2 \notin \mathcal{N}$. QED

Usefulness of Confidence Networks

Definition

"A network with intervals assigned to edges is an 'X% confidence' network if, for different random samples, it has an X% probability of containing the 'true' tree."

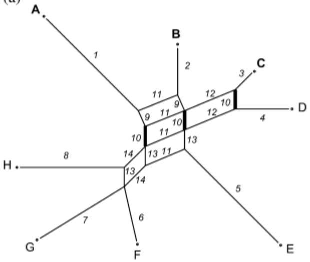
Discussion Question: Is the reason the word true is in quotes in this definition because we are using a nonparametric structure where the ground truth (the actual true tree) is frequently unknown?

Intuitive Representations: Geometry

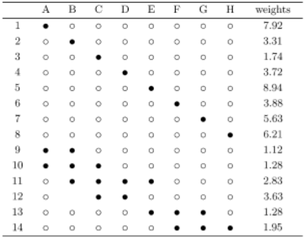
"Suppose that the splits are indexed from 1 to m. A tree can then be coded as a point in m-dimensional space: the ith coordinate is the length corresponding to the i^{th} split, or 0 if that split is not present in the tree (Holmes 2005). The split network then corresponds to a box in m-dimensional space: the range of values in the i^{th} dimension is given by the interval for the i^{th} split. A tree is contained in the network if the corresponding point is contained in the box."

Intuitive Representations: Adjacency Matrix

(a)



(b)



	A	B	C	D	E	F	G	H	weights
1	●	○	○	○	○	○	○	○	7.92
2	○	●	○	○	○	○	○	○	3.31
3	○	○	●	○	○	○	○	○	1.74
4	○	○	○	●	○	○	○	○	3.72
5	○	○	○	○	●	○	○	○	8.94
6	○	○	○	○	○	●	○	○	3.88
7	○	○	○	○	○	○	●	○	5.63
8	○	○	○	○	○	○	○	●	6.21
9	●	●	○	○	○	○	○	○	1.12
10	●	●	●	○	○	○	○	○	1.28
11	○	●	●	●	●	○	○	○	2.83
12	○		●	●	○	○	○	○	3.63
13	○	○	○	○	●	●	●	○	1.28
14	○	○	○	○	○	●	●	●	1.95

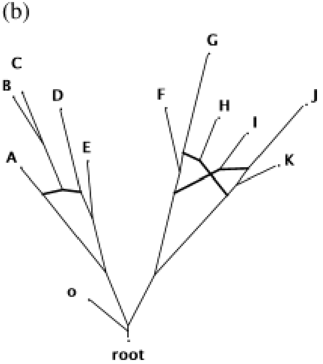
a) Split Network

- Nodes labelled as lettered taxa.
- Edges labelled as numbered branches.

b) Adjacency Matrix

- Edges 1-7 only have one entry in the matrix as they all connect to intermediate nodes and edges.
- All other edges correspond to splits that can separate the network into 2 sets of taxa with each set having a cardinality greater than 1.

Reticulate Networks



- Reticulate Network $\approx$ Recombination Tree

Summary: Split v. Reticulate Networks

Split Networks

- unrooted
- nodes $\neq$ ancestors
- a structure that effectively incorporates a bunch of trees simply by storing a set of splits with corresponding weights

Reticulate Networks

- rooted
- nodes $=$ ancestors
- effectively modelling on ARGs
- like split networks, isn't a tree structure but recombination events are the constraint on multiple trees (i.e. an ARG will only take as many trees as necessary to recreate some inferred recombination structure)

Recap: Goal of Split Network Inference

Goal: Create an inference method on trees that do not tie us down to a single tree.

Attempted Model: Use a known ML model on trees.

Obstacle: These models make restrictions that do not allow for incompatible signals with or between trees.

Solution: Use split networks that can represent these incompatibilities in a single object.

You thought we were done with terminology?

- sampling error: random error stemming from sample size selection (almost always stemming from using too few sites)
- systematic error: mistakes in the assumptions or method of a model
- long-branch attraction: long branches are so long and unique that they get modeled as one when in reality they are separate lineages

Important Takeaway: Systematic error is a shortcoming of a model that can lead results that claim observed data came from the wrong tree or not even any tree at all.

Solutions to Errors

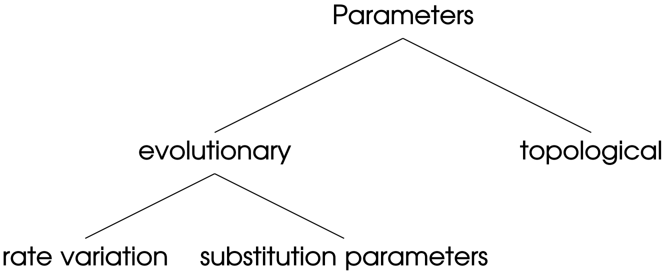
- sampling error: nonparametric bootstrap sampling or sampling from the posterior (**these are already implemented**)
- systematic error: model under a split network structure (**this is a huge motivation for this paper**)

Split Network Inference Algorithm

Algorithm 2: Inital Strategy for Using Split Networks in Phylogenetic Inference

```
Result: Using split networks in phylogenetic inference.
while trying to find network do
    Construct a split network using the best available model and
      method. ;
    if network is tree-like then
        Return the current network and use it.
    else
        Assess what problems exist in the model assumptions and if
          ambiguities are explainable.
    end
end
```

Using Split Networks for Inference



Parameters

evolutionary

topological

rate variation

substitution parameters

SplitsTree4

This paper is incredibly informative, but it really serves as an academic advertisement for the software the authors developed: SplitsTree4. Features of SplitsTree4:

- Supports plugins and complete control of the program from command line.
- Implements median networks, split decomposition, and consensus networks. (more on this next slide)
- Implements recombination and hybridization network.
- Implements ML estimation of distances from amino acid and nucleotide sequences under standard evolutionary models.
- Has a graphical interface that allows for interactive exploration of data. (mainly menu/drag and drop eperiences)

SplitsTree4 Methods

- median networks (parsimony): method that constructs maximum likelihood split networks directly from character data
- split decomposition: method that constructs split networks from inferred distance matrices
- consensus networks: method that constructs split networks from sets of trees

Note: Which split network inference method we decide to use can highly be decided on which of the sources of data is available and/or most reliable.

Applied Experiments

The authors demonstrate the capabilities of SplitsTree4 with 3 different experiments:

- Examining Split Decomposition in Heterogeneous Evolution Settings
- Using ML Based Split Decomposition Techniques to Determine Monophylogeny
- Cocktail of Network Methods on the Dusky Dolphins

Heterogeneous Evolution

Problem Statement: Many models say that different regions (clades) of a tree evolve at the same rate. This can result in systematic error if this is not the case.

Solution: Use split decomposition (an inferential technique on split trees).

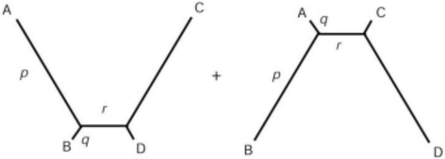
Why does this work????: We get to use an object that stores multiple trees that can have different branch lengths. Obviously, the object itself is not directly a tree, but a combination of trees that together in tandem form some sort of reasonable heterogeneous evolutionary structure.

Discussion Question: What prevents a method that assigns a prior to branch lengths at different clades? I thought that sampling from this kind of posterior would also generate heterogeneous evolution, but I'm probably wrong.

Heterogeneous Evolution (cont.)

- Maximum Parsimony
- PhyML
- Split Decomposition

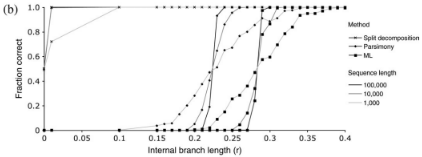
A tree was generated where the first half of sites were evolved with the left set of branch and the second half with the right set with p=0.75, q=0.05 being fixed and r varying from 0 to 0.4.



Heterogeneous Evolution Results

A tree is said to be correct, if it accurately detects the $AB \mid CD$ split with at most one alternative split.

Main Idea: As r increases, the split has higher weight associated with it, so all methods improve as r increases. Which one does best?



Heterogeneous Evolution Results (cont.)

Result:

- As the number of sites increases, each method gets to a correct proportion of 1 faster.
- As r increases, all the methods actually converge on the correct tree (hence why all eventually reach fraction correct of 1). This is by design of the experiment.
- Only split decomposition achieves a super high accuracy when r is low.
- This result shows how the improper model assumptions can lead to trees that don't yield correct results despite having r large enough to detect a split.
- Split networks have a ridiculous advantage because they can choose 2 trees. For example, imagine if split decomposition learned the combination of maximum parsimony and PhyML.

Animal Phylogeny Results

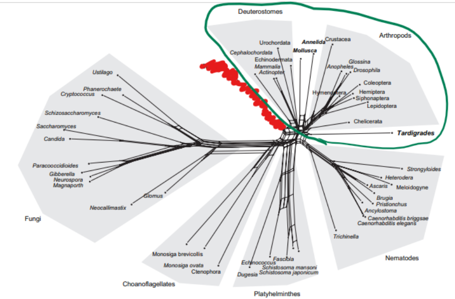
So... believe me... I tried to make sense of the actual organisms and jargon, and I just do not understand. But, the experiment is still easily summarized!

The idea is the authors are taking 2 competing answers of grouping: coelometes and ecdysozoa. Their findings suggest that evidence suggests the latter has more data going for it and systematic error is the reason there is much debate on the subject.

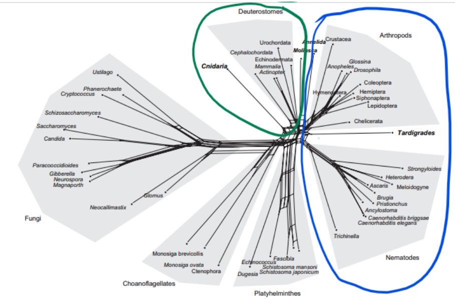
Source of Systematic Error: Including a taxon not previously included in research suggests that the former outcome is likely to stem from long-branch attraction.

Animal Phylogeny

Coelomates

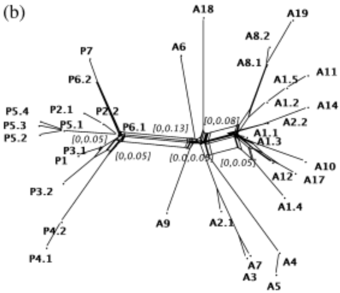


Ecdysozoa

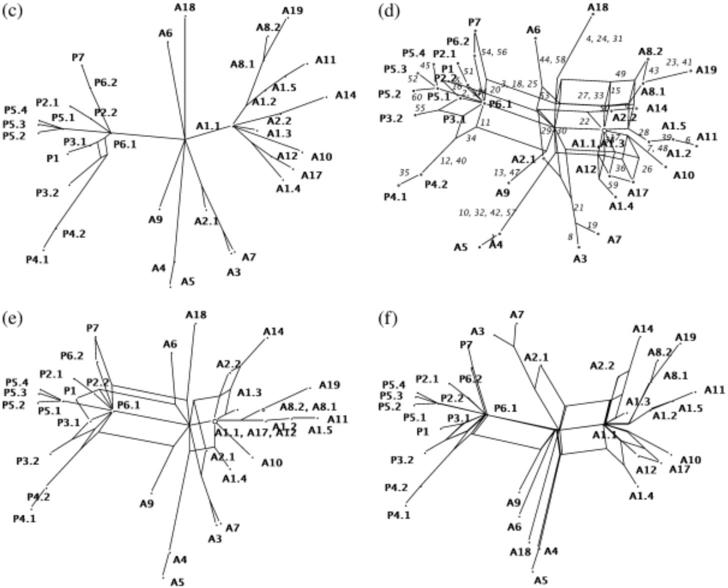


Evolution of Dolphins

This experiment investigates the phylgeography of dusky dolphins using data from 60 variable positions for 36 different haplotypes seen in 124 individuals sampled from Peru, Argentina, and South Africa. Here is an example of a 95% confidence tree for the "true" phylogeny of dolphins.



Evolution of Dolphins Visuals



Conclusion

- Phylogenetic networks have been neglected (up until around 2005) because of a lack of statistical framework and convenient software. (Authors words, not mine.)
- SplitsTree4 was at the time the tool that addressed this concern.
- The performance of split decomposition - one of the more common phylogenetic network inference techniques - yields promising results in application.