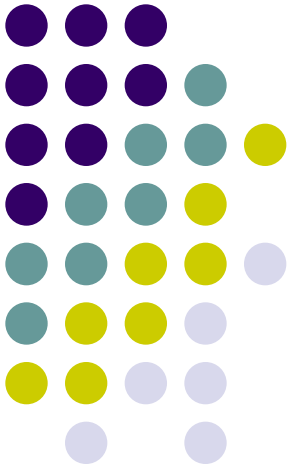
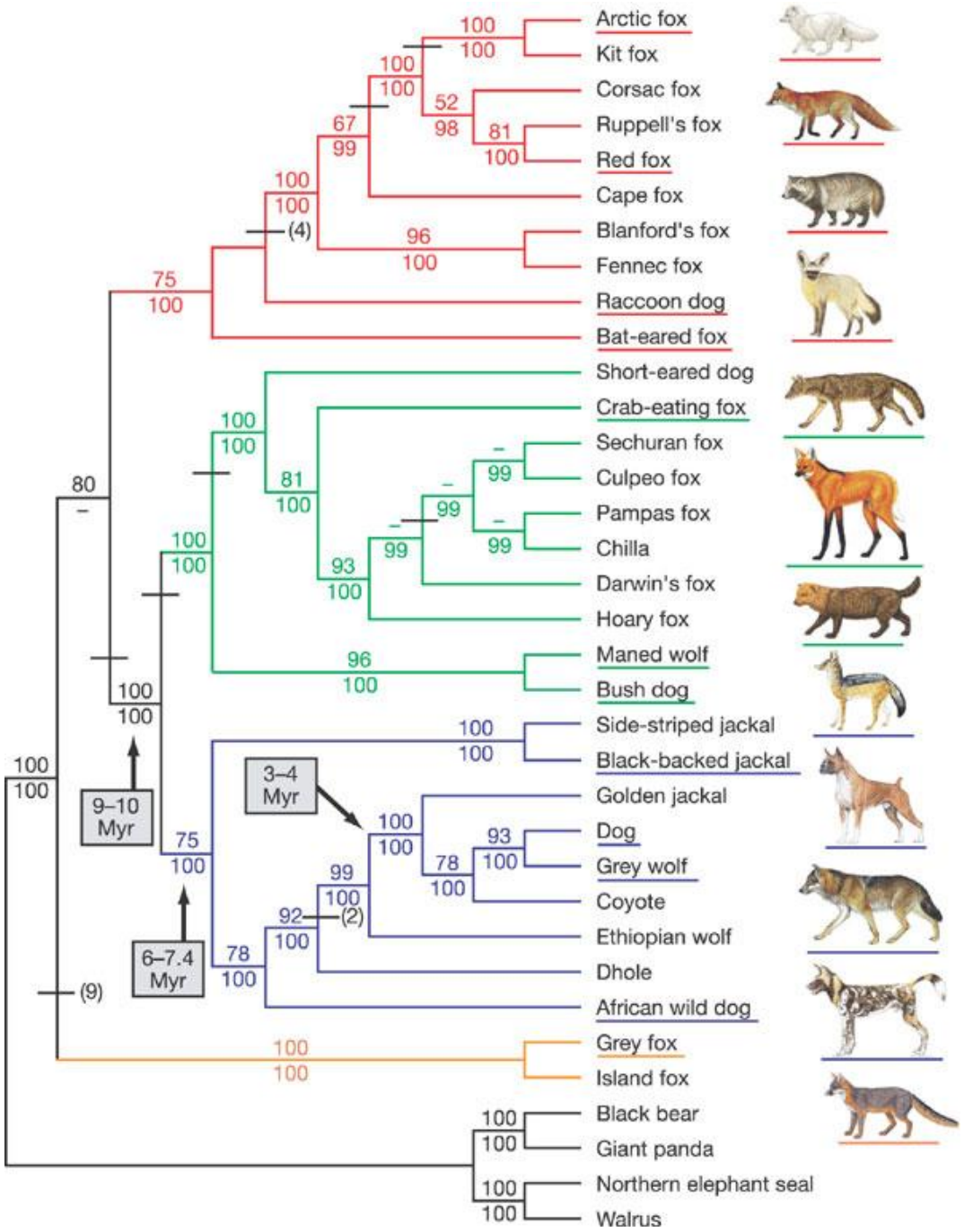
An introduction to phylogenetic networks



Steven Kelk
Department of Knowledge Engineering (DKE)
Maastricht University

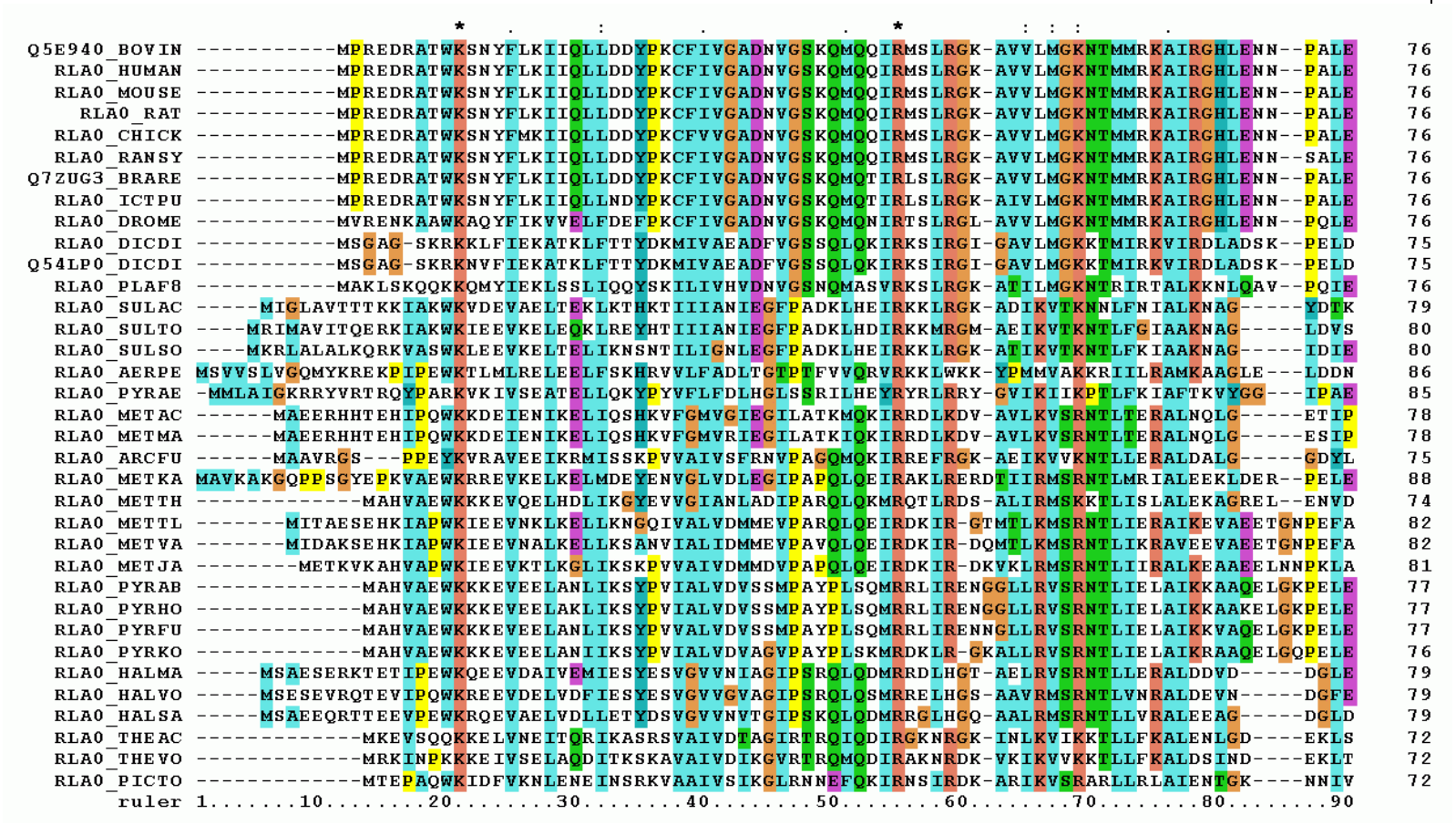
Email: steven.kelk@maastrichtuniversity.nl
Web: http://skelk.sdf-eu.org



Genome sequence, comparative analysis and haplotype structure of the domestic dog

Lindblad-Toh et al, Nature 2005

(Almost) everything begins with Multiple Sequence Alignment

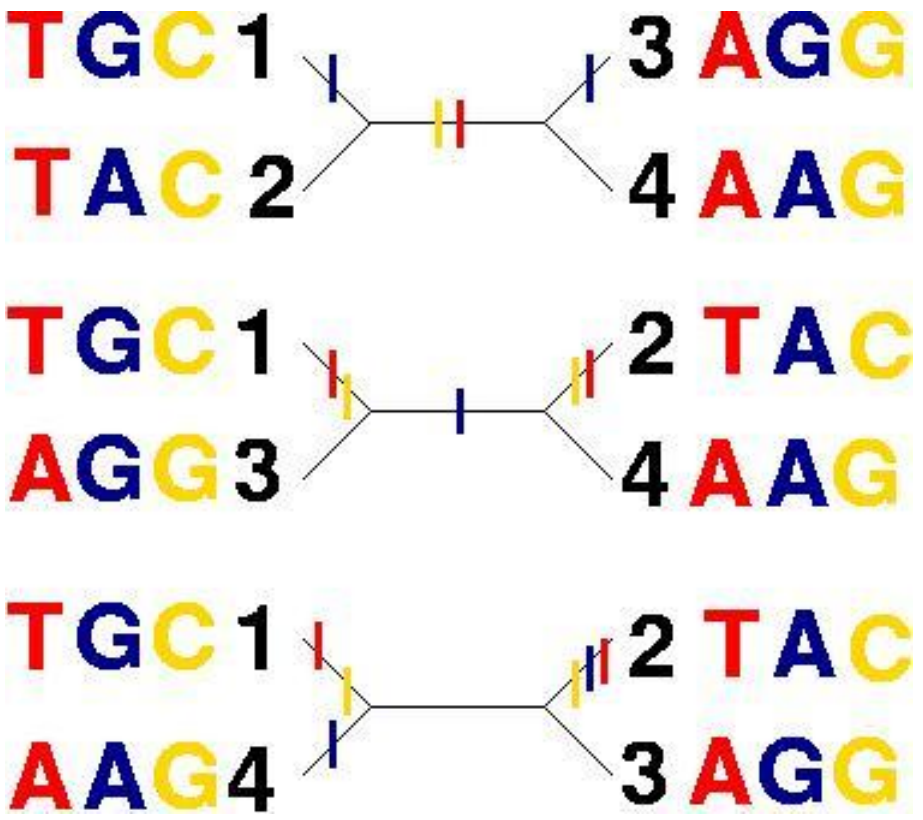


Dominant methods for building phylogenetic trees

- **Character-based methods**
    - Maximum Parsimony (MP)
    - Maximum Likelihood (ML)
- **Bayesian methods (Markov Chain Monte Carlo - MCMC)**
- **Distance-based methods**
    - Neighbour Joining
    - UPGMA
- **“Supertree” methods: glueing together smaller subtrees**

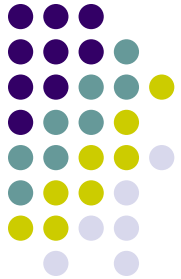
Sequence 1	T	G	C
Sequence 2	T	A	C
Sequence 3	A	G	G
Sequence 4	A	A	G

The "most parsimonious" tree solution



From: http://artedi.ebc.uu.se/course/X3-2004/Phylogeny/Exercises/mp.html

There is more to life than trees



• All these methods assume that a (single) tree is the best way to model the underlying evolution.

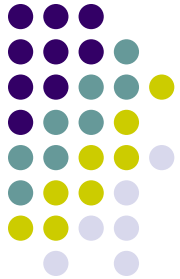
• If this is not true, then we have a problem, because there is a high risk that the output of tree-building algorithms will then be **meaningless**.

• Sometimes there are clues about this:
  • Algorithms build very badly supported trees
  • Extra knowledge about the underlying evolutionary mechanisms

• But in general it is **dangerously easy** to confuse non-treelike evolution with a **noisy tree signal**.

• Therefore **critical** to understand and model underlying mechanisms.

Why might we get weak support for a tree?



"Noisy tree"
Data *does* fit a single tree, weak support is only a consequence of "noise"

"Trees in trees"
Data consists of multiple different tree signals…but both gene and species evolution are still ultimately treelike (e.g. due to incomplete lineage sorting, gene loss, gene duplication)

"Trees in networks"
Data consists of multiple different tree signals…gene evolution is treelike, but species evolution is no longer treelike (e.g. hybridization, horizontal gene transfer)

"Reticulation"
Inherently non-treelike (reticulate) phenomena, such as meiotic, sexual recombination

Very briefly: trees in trees

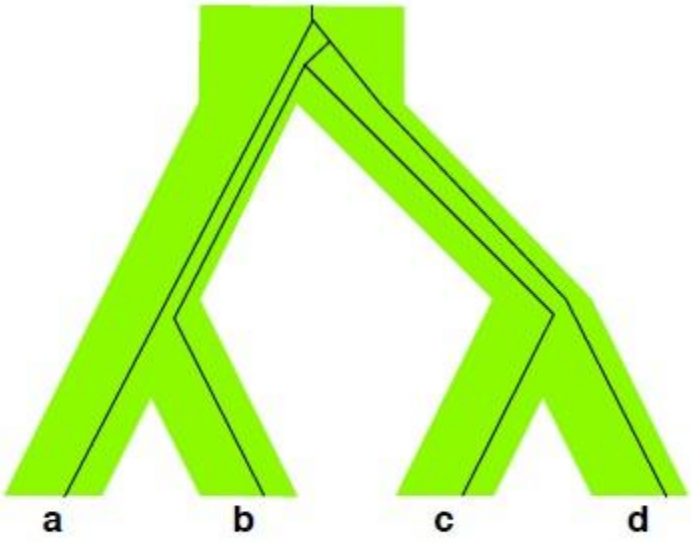
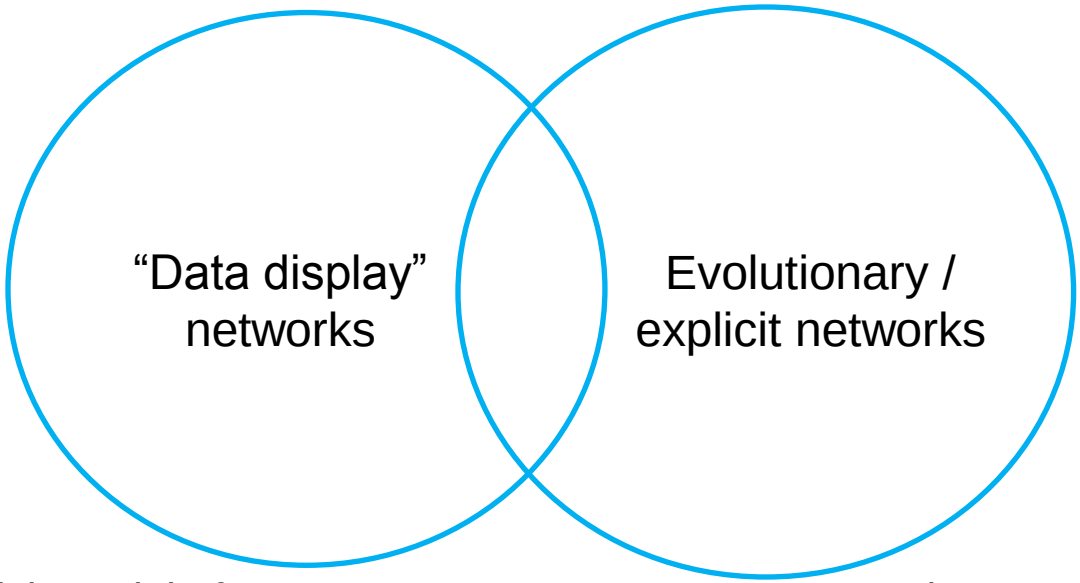


Fig. 16 A gene tree (solid lines) evolving within the branches of the species tree, where the gene tree topology is identical to that of T_2 in Fig. 1(b). The gene tree differs from the species tree due to (incomplete) lineage sorting.

From: L. Nakhleh, "Evolutionary phylogenetic networks: models and issues." In: The Problem Solving Handbook for Computational Biology and Bioinformatics, L. Heath and N. Ramakrishnan (editors). Springer, 125-158, 2010.

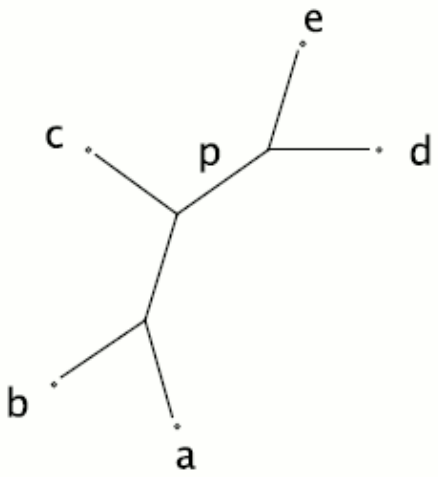
Phylogenetic networks



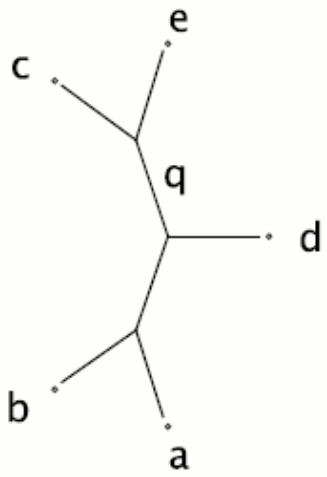
No explicit model of evolution: tries to graphically represent **where** the data is non-treelike

Tries to model the **events** that caused the data to be non-treelike

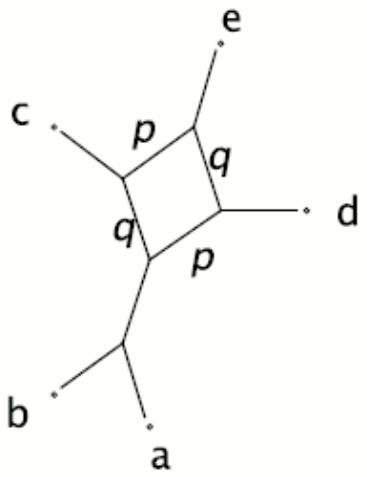
Data-display networks (1)



T_1



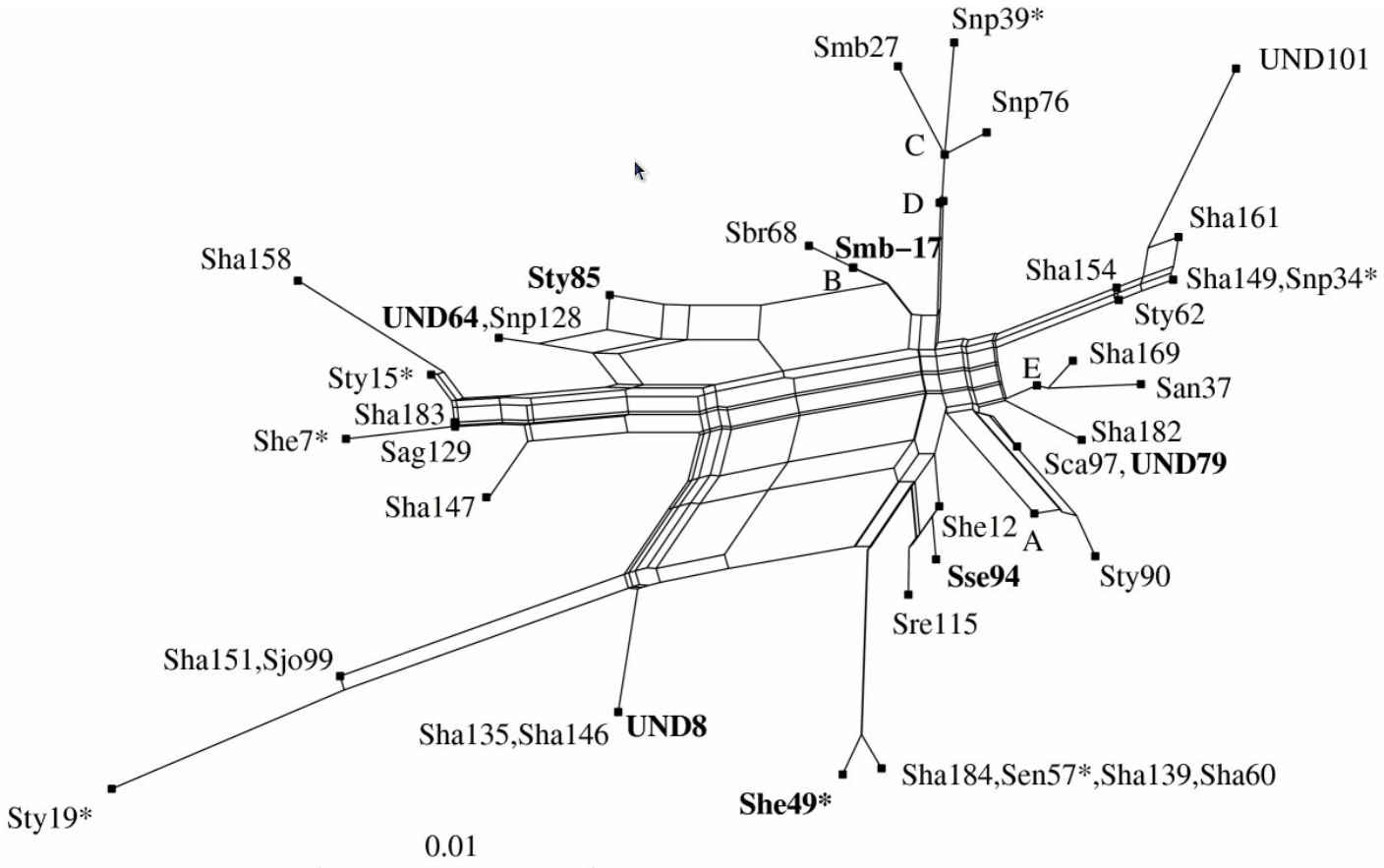
T_2



$SN(\Sigma)$

From: Daniel Huson, ISMB-Tutorial 2007: Introduction to Phylogenetic Networks

Data-display networks (2)



A phylogenetic network. The network was generated by Neighbor-Net for a sequence-based data set comprising of *Salmonella* isolates that originally appeared in [17]. A detailed network-based analysis of this data is presented in [2], where the strains indicated in bold-face are tested for the presence of recombination. Note that the network is planar (that is, it can be drawn in the plane without any crossing edges), and that parallel edges in the network represent bipartitions of the data.

Bryant *et al. Algorithms for Molecular Biology* 2007 **2**:8 doi:10.1186/1748-7188-2-8

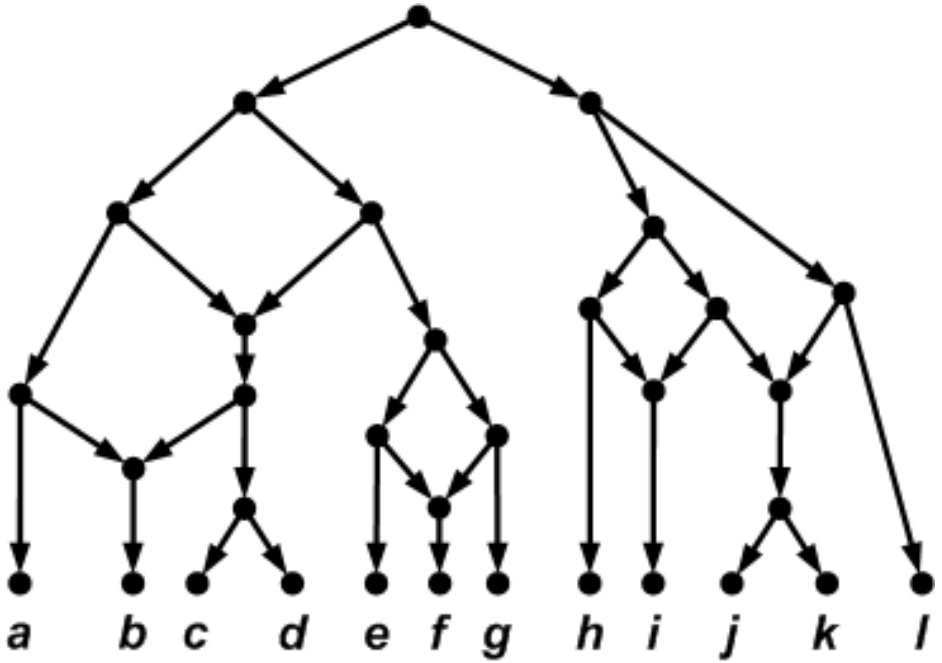
Evolutionary phylogenetic networks

- Used to explicitly model reticulate evolution:
  - Hybridization
  - Horizontal Gene Transfer (HGT)
  - Recombination

- Reticulation events have an explicit biological interpretation

- Usually rooted, with an explicit "direction" of evolution

- Underlying mathematical abstractions are often similar, despite different scale levels of interpretation



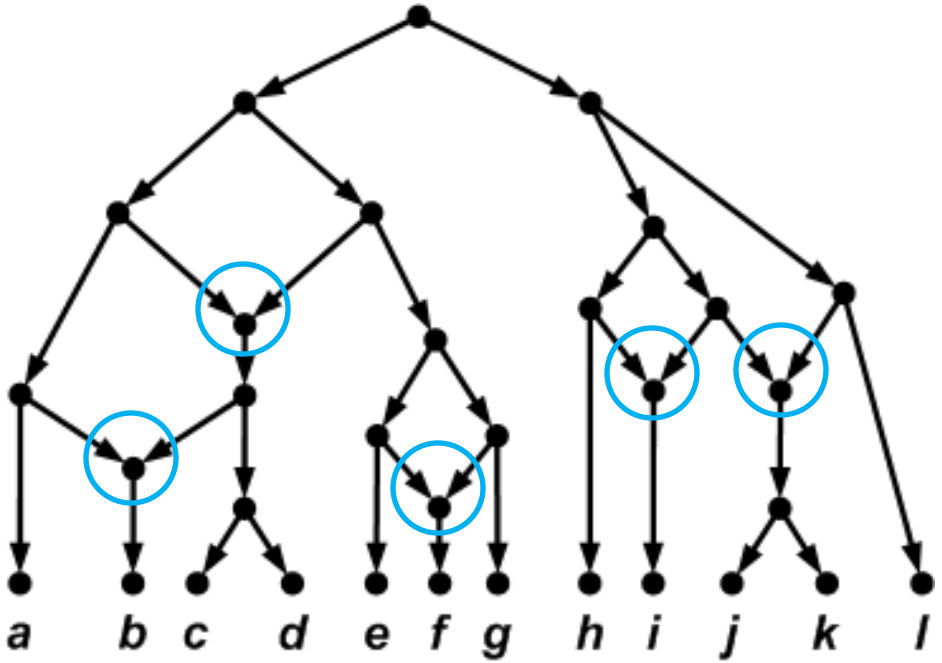
Evolutionary phylogenetic networks

- Used to explicitly model reticulate evolution:
  - Hybridization
  - Horizontal Gene Transfer (HGT)
  - Recombination

- Reticulation events have an explicit biological interpretation

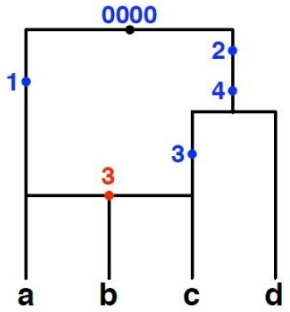
- Usually rooted, with an explicit "direction" of evolution

- Underlying mathematical abstractions are often similar, despite different scale levels of interpretation

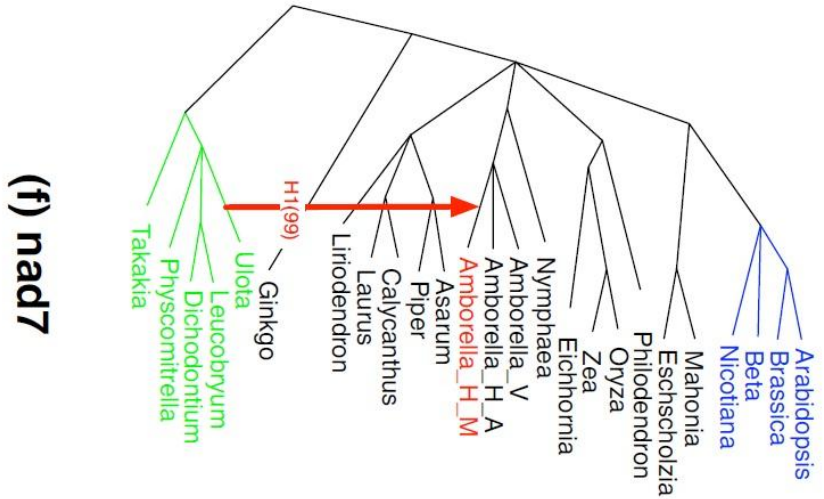


Different models and scales, but always rooted, directed acyclic graphs (DAGs)

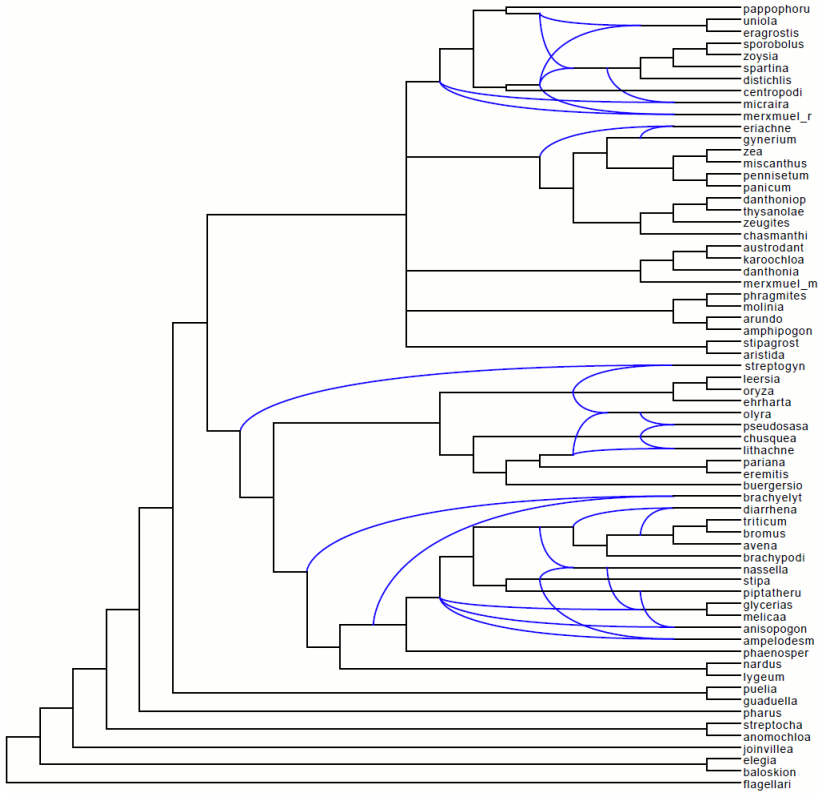
	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1



Ancestral Recombination Graph (ARG)

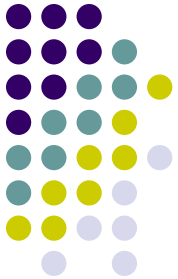


Horizontal Gene Transfer (HGT)



"Softwired cluster" network

Constructing evolutionary phylogenetic networks



- It's important to ask ourselves several questions:

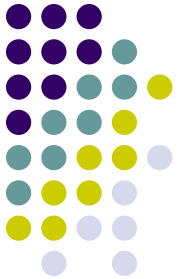
  1. MODEL: What are we trying to **model** exactly? Is it biologically realistic?

  2. OBJECTIVE: What do we consider to be an **"optimal" solution** within that model?

  3. TRACTABILITY: Is there any hope of developing **efficient algorithms** to compute optimal solutions?

- Extremely challenging to simultaneously answer these questions well!

- In the meantime: **many** different models, algorithms, packages



Several case studies

1. A “direct” method : constructing Ancestral Recombination Graphs (ARGs) by modelling **crossover events.**

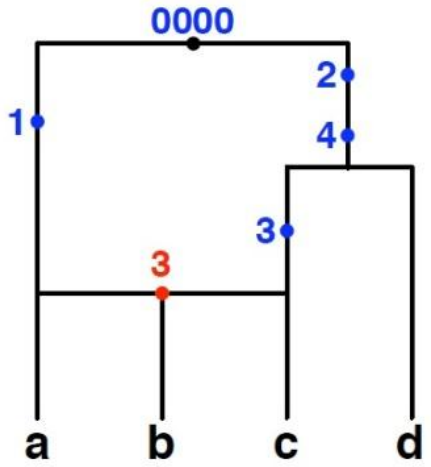
2. “The trees within” : methods which analyse phylogenetic networks based on the set of trees **contained within them.**

    a) Extensions to Maximum Parsimony (MP) and Maximum Likelihood (ML)
    b) Parsimoniously **embedding gene trees** in species networks

3. “Piecewise” methods : constructing phylogenetic networks by **merging** many smaller evolutionary hypotheses (e.g. rooted triplets, clades) into a single network.

Case study 1: constructing Ancestral Recombination Graphs (ARGs)

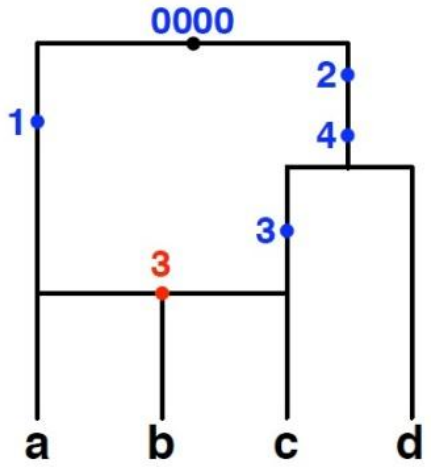
	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1



- Input is **binary character data** (i.e. strings of binary data)
- Reticulations represent **chromosomal crossover** (mostly single crossover, sometimes multiple crossover). Sometimes also gene conversion.
- Mutation model is the **"infinite sites"** model: at most one mutation per site (0 to 1, or 1 to 0).
- Goal is to construct an ARG with a **minimum number** of reticulation events.

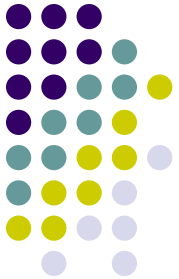
Case study 1: constructing Ancestral Recombination Graphs (ARGs)

	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1



- Input is **binary character data** (i.e. strings of binary data)
- Reticulations represent **chromosomal crossover** (mostly single crossover, sometimes multiple crossover). Sometimes also gene conversion.
- Mutation model is the **"infinite sites"** model: at most one mutation per site (0 to 1, or 1 to 0).
- Goal is to construct an ARG with a **minimum number** of reticulation events.

Case study 1: constructing Ancestral Recombination Graphs (ARGs)

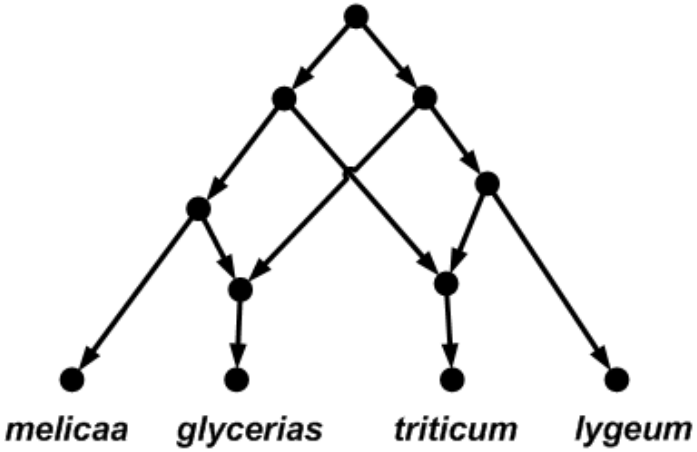


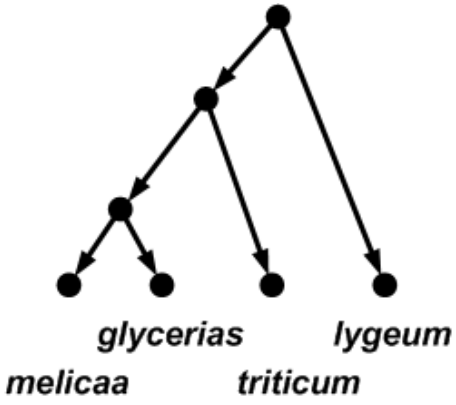
- Programs for constructing ARGs include HAPBOUND, SHRUB, BEAGLE

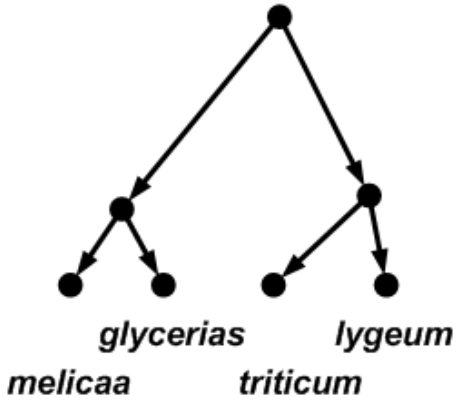
- Extensive interest and research from the theoretical computer science community (e.g. Dan Gusfield)

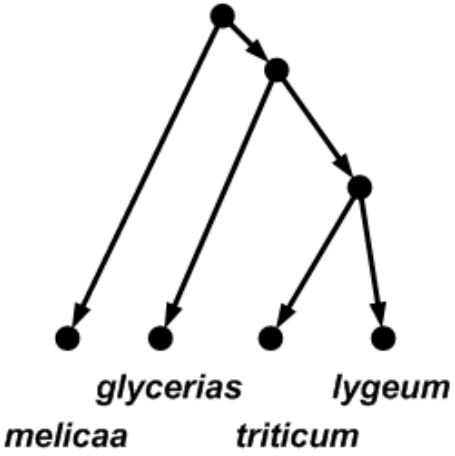
- Issues:
    - Difficult to solve (NP-hard, also difficult in practice)
    - Modelling of homoplasy (recurrent and back mutation) is in its infancy (infinite sites model excludes this)
    - Rigid biological model (crossover)
    - Software implementations still rather experimental
    - Standard phylogenetic concepts such as bootstrapping, branch-lengths etc. are not considered

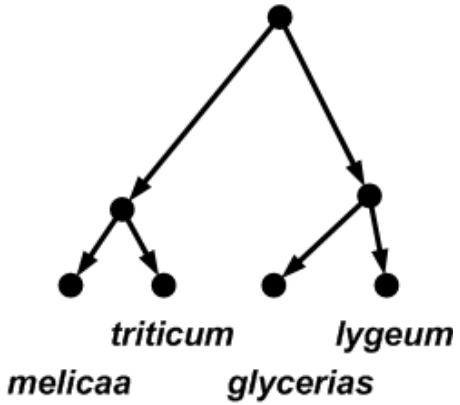
"The trees within": methods based on the set of trees inside a network



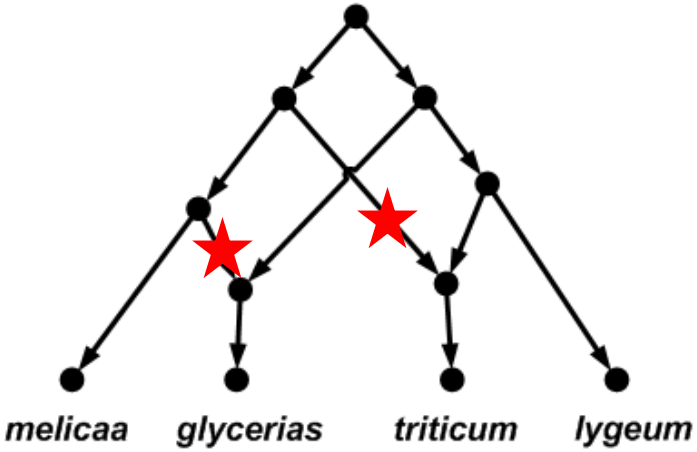


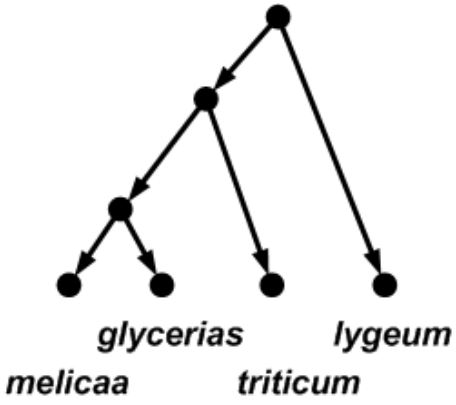


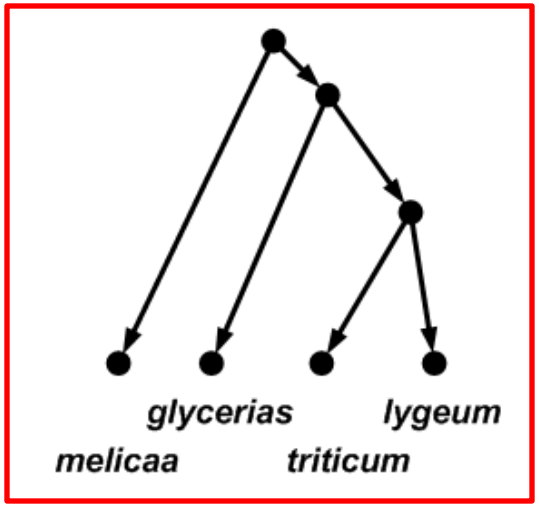


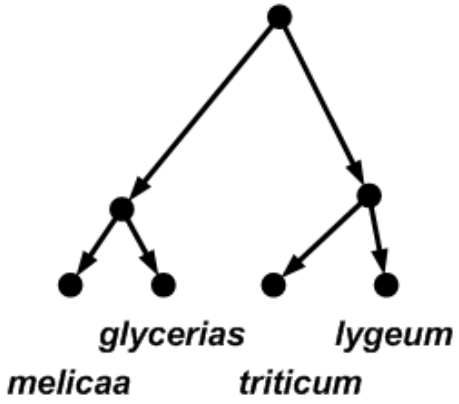


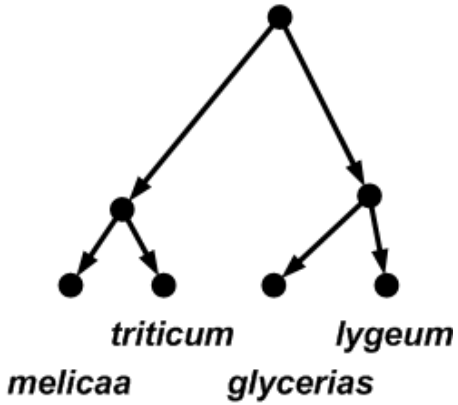
"The trees within": methods based on the set of trees inside a network



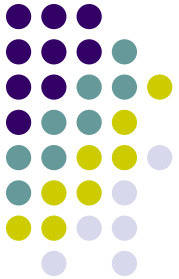








Case study 2(a): extensions to Maximum Parsimony and Maximum Likelihood



- The group of Luay Nakhleh (Rice University, USA) is very interested in this. (They are very strong on the modelling side.)

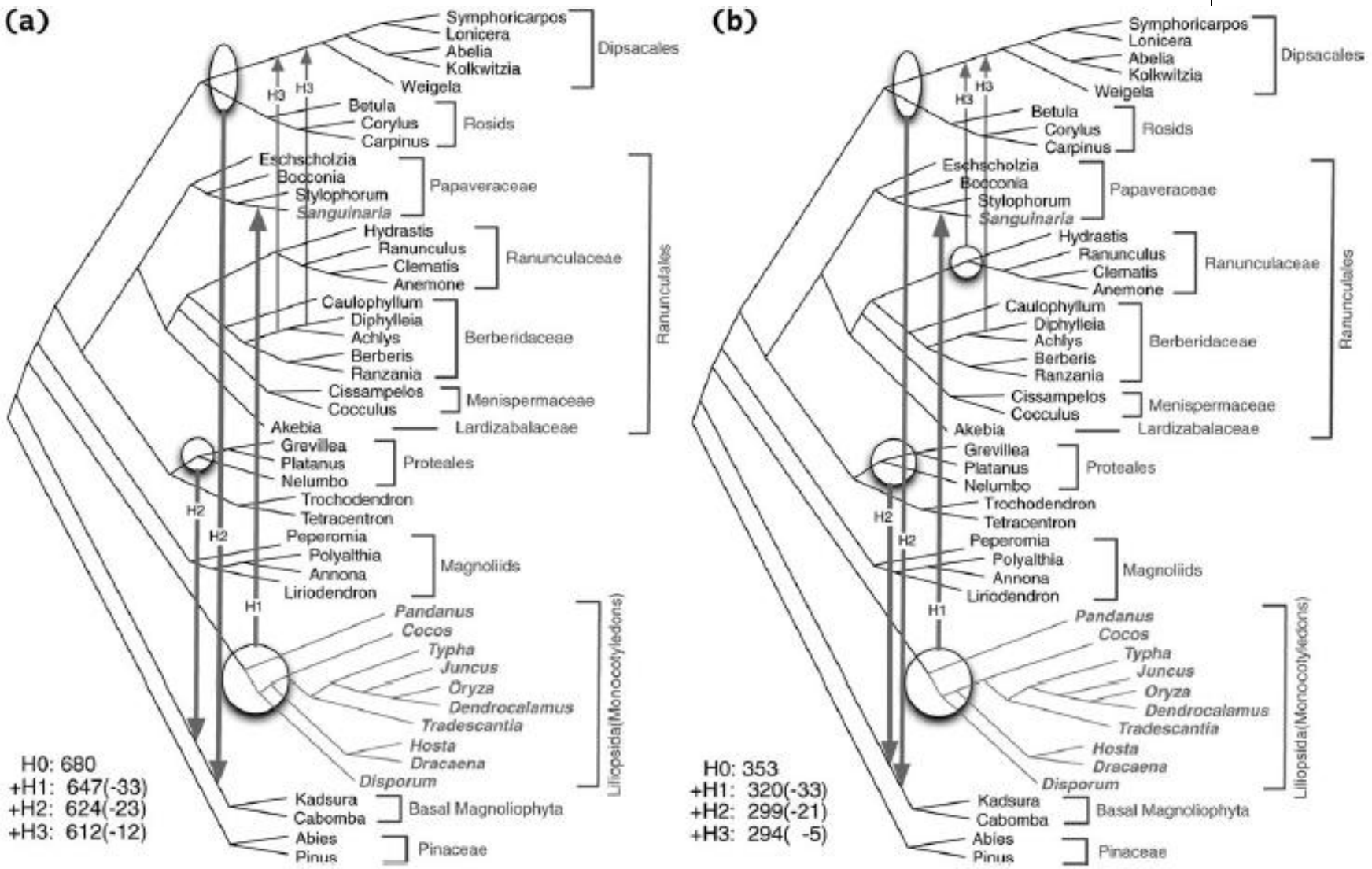
- The general idea is to define the parsimony/likelihood score of a network, as a function of the **set of trees** contained within it.

- Software: PHYLONET, NEPAL

- Issues:
  - Again, a very specific (and thus rigid) model
  - Assumed independence of characters leads to problems
  - More reticulations = better score, so when do we stop adding reticulations?
  - Even “small” variant (e.g. here is a network, compute the best parsimony score for it) is algorithmically challenging
  - Algorithms for the “big” variant (i.e. find me the best network) are still very basic

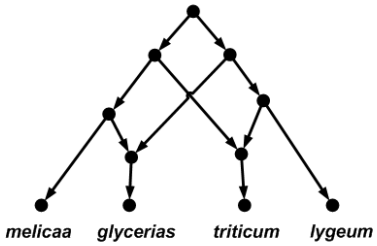
From: Jin, G., Nakhleh, L., Snir, S., Tuller, T.: *Inferring phylogenetic networks by the maximum parsimony criterion: A case study.* Molecular Biology and Evolution 24(1), 324–337 (2007).

MP analysis based on the ribosomal protein gene *rps11* of a group of 47 flowering plants, which was analysed by Bergthorsson et al (2003)

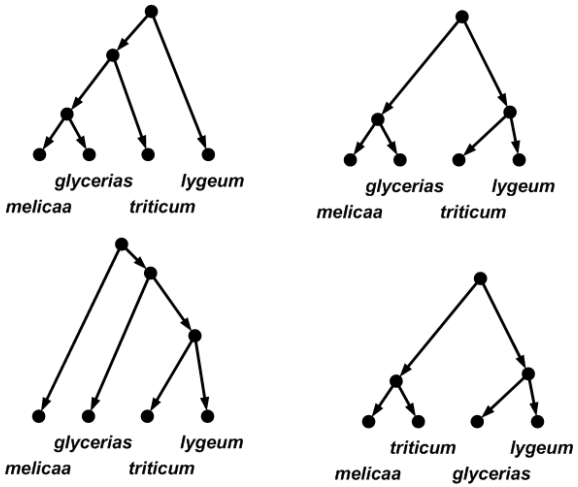


Case study 2(b): combining multiple gene trees into a single species network

- Recall this example:



Species network



Four gene trees contained in the species network

- Input: a set of gene trees

- Output: a species network that **contains all the input gene trees** and which has a **minimum number** of reticulations

From: **Fast computation of minimum hybridization networks,** Benjamin Albrecht, Celine Scornavacca, Alberto Cenci and Daniel H. Huson, to appear in Bioinformatics (2011).

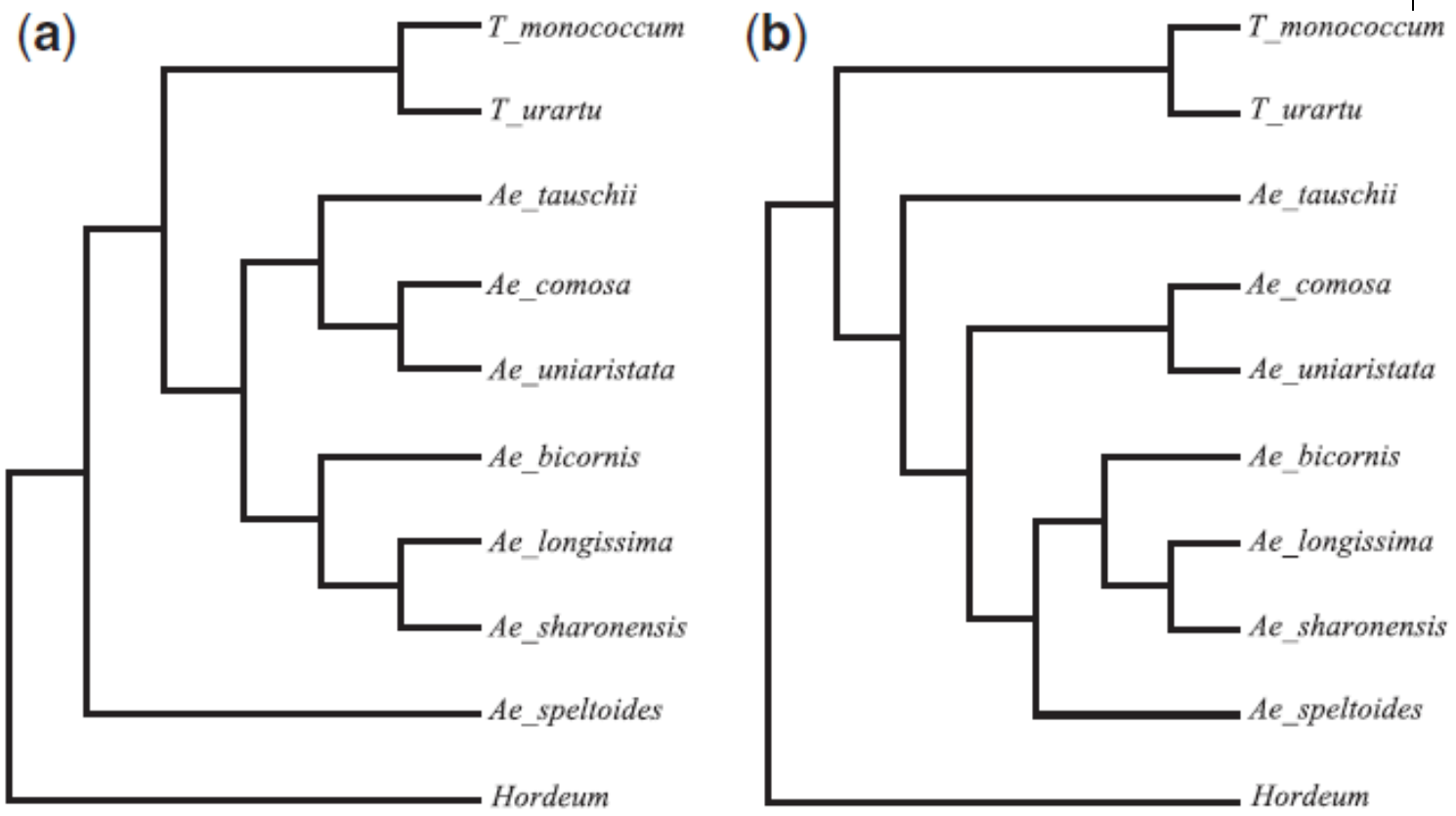


Fig. 3. The two consensus trees computed from 100 bootstrap replicates for the matK (**a**) and PinA (**b**) datasets.

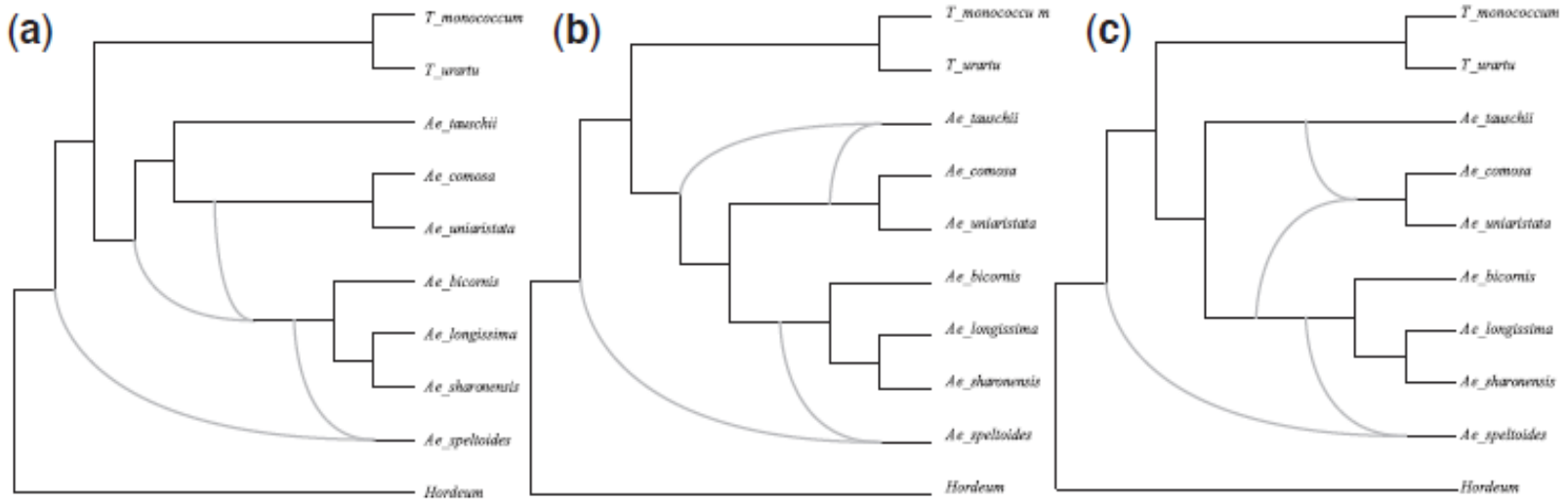
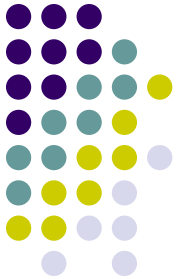


Fig. 4. The three hybridization networks obtained by the described algorithm for the matK and PinA consensus trees of Figure 3.

Case study 2(b): combining multiple gene trees into a single species network

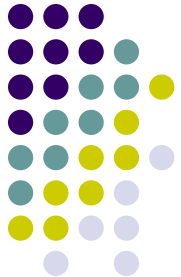


- There has been a huge amount of research from the theoretical computer science community for the case when the input consists of **exactly two binary gene trees**

- The result is a lot of very nice math, and increasingly fast algorithms (such as HYBRIDNET and an algorithm in DENDROSCOPE 4)

- Issues:
  - No software exists for **two non-binary** trees (multifurcations = soft polytomies, important for modelling uncertainty)
  - No software exists to reliably compute optimal solutions for **three or more trees**, even when binary
  - Issues with time-consistency
  - Multiple solutions? Branch lengths? Bootstrapping?
  - Rooting problems

Case study 3: "Piecewise" methods: combining triplets into a single species network



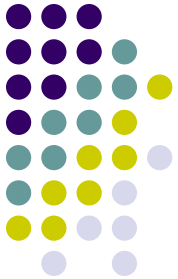
- **Rooted triplets:** phylogenetic trees with only 3 leaves

- The idea is that it might be easier to build lots of very small trees (rooted triplets) and to merge them into a single network, then to try and construct the network in one go

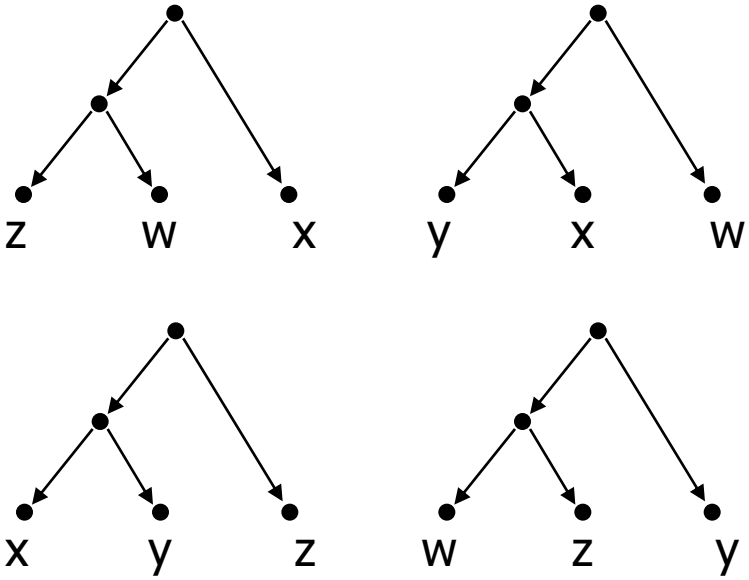
- Rooted triplets can be inferred directly/ad-hoc or extracted from gene trees

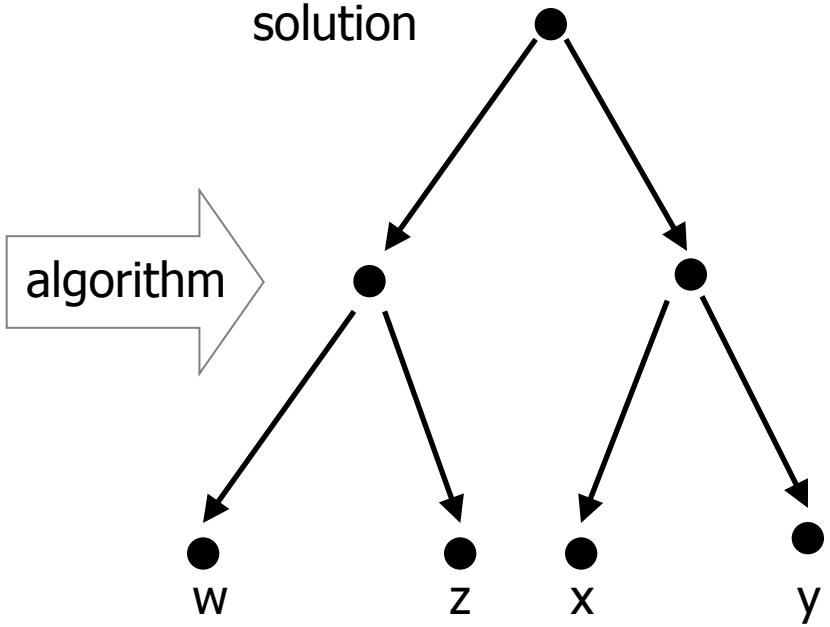
- Idea is similar to trees i.e. combine them into a single network such that the number of reticulations is **minimised**

Case study 3: "Piecewise" methods: combining triplets into a single species network

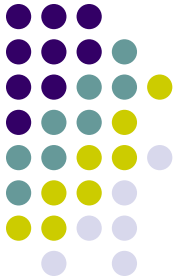


- For example. Suppose I want to reconstruct a plausible evolution for the species set {w,x,y,z}.
- I am given a set of rooted triplets zw|x, yx|w, xy|z, wz|y. (Note zw|x = wz|x.)

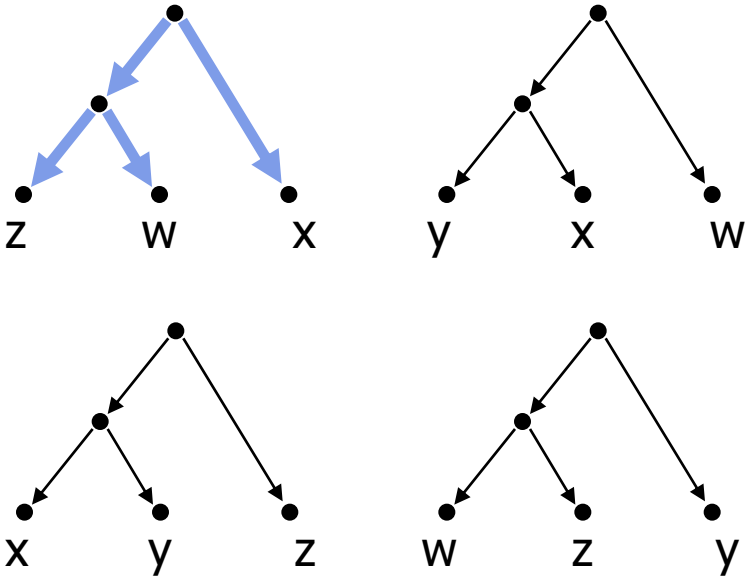


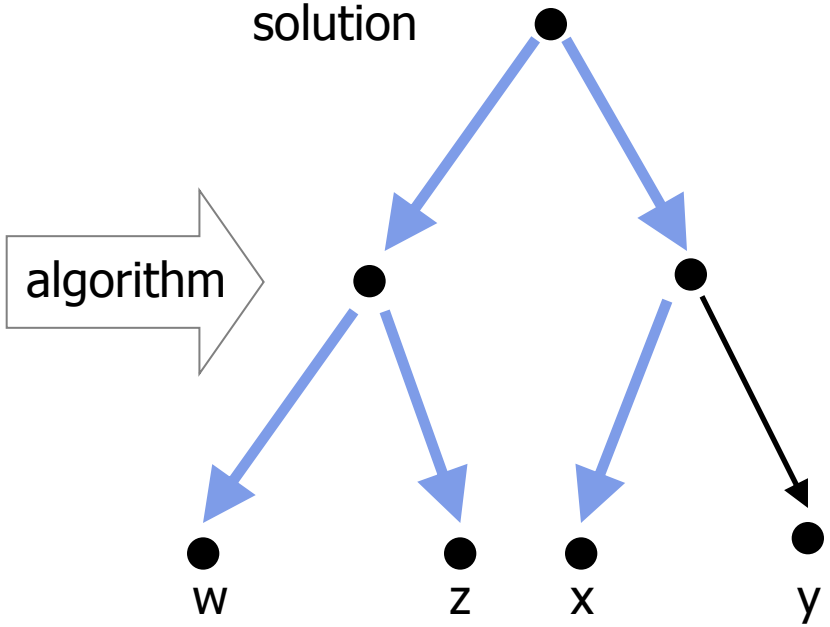


Case study 3: "Piecewise" methods: combining triplets into a single species network

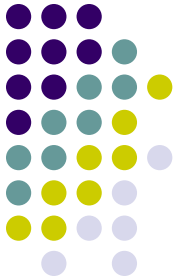


- For example. Suppose I want to reconstruct a plausible evolution for the species set {w,x,y,z}.
- I am given a set of rooted triplets zw|x, yx|w, xy|z, wz|y. (Note zw|x = wz|x.)

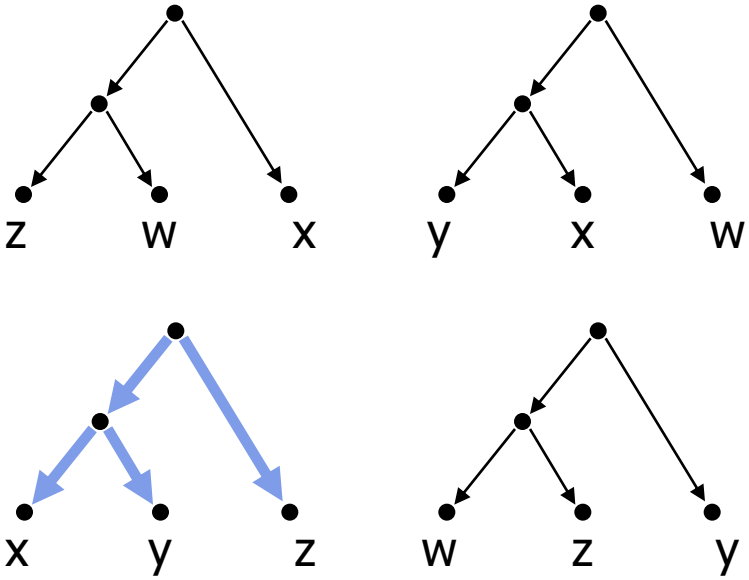


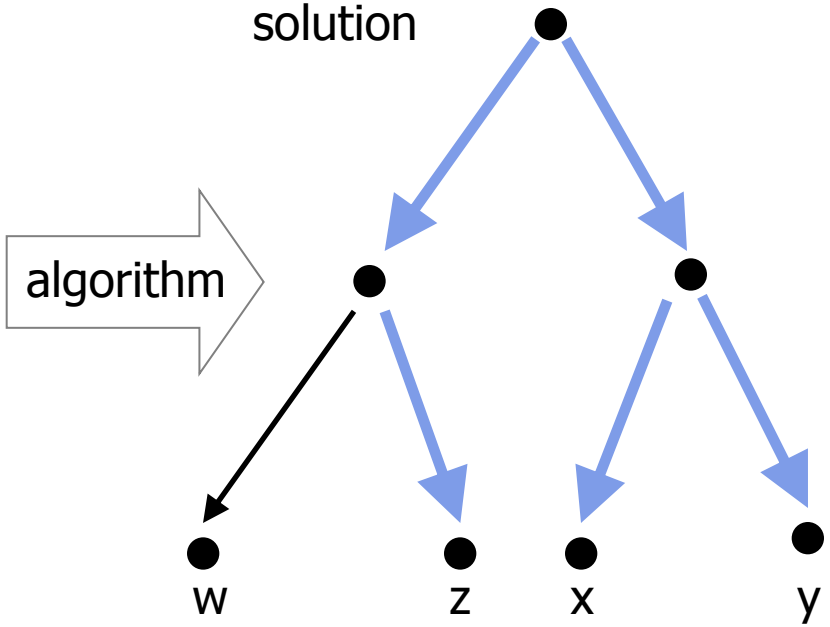


Case study 3: "Piecewise" methods: combining triplets into a single species network



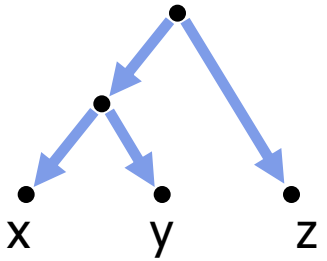
- For example. Suppose I want to reconstruct a plausible evolution for the species set {w,x,y,z}.
- I am given a set of rooted triplets zw|x, yx|w, xy|z, wz|y. (Note zw|x = wz|x.)

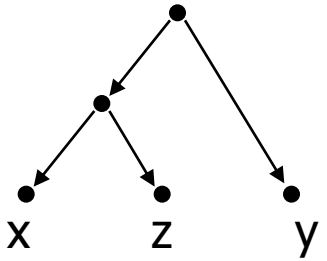


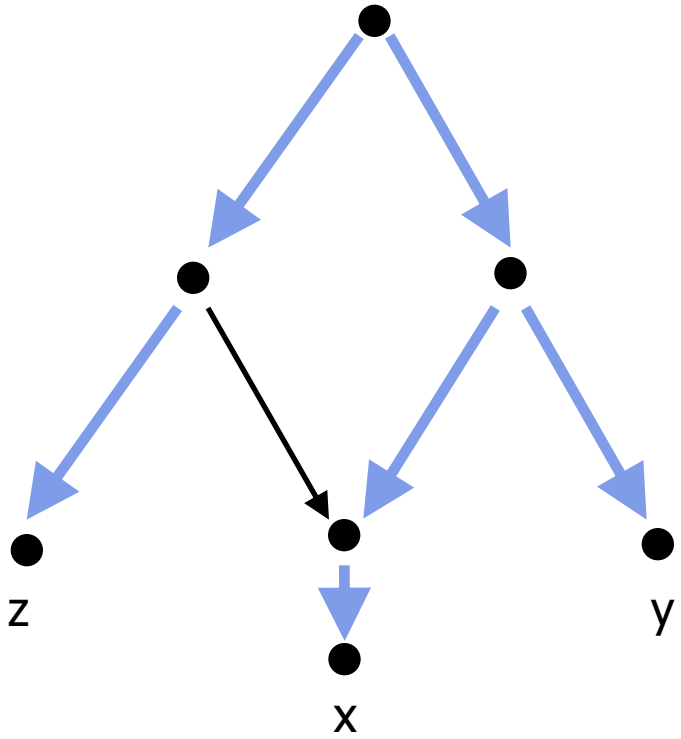


Case study 3: "Piecewise" methods: combining triplets into a single species network

- For example, suppose the input is {xy|z, xz|y}.

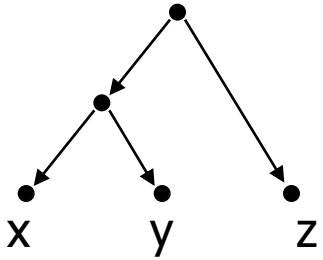


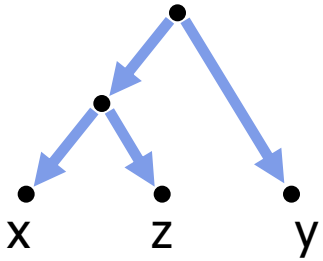


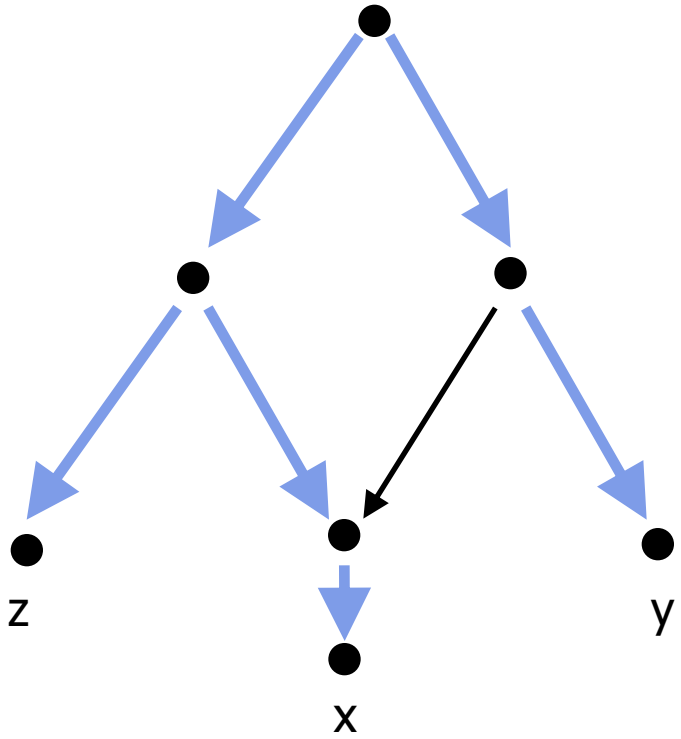


Case study 3: "Piecewise" methods: combining triplets into a single species network

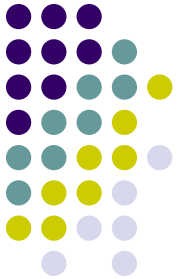
- For example, suppose the input is {xy|z, xz|y}.







Case study 3: "Piecewise" methods: combining triplets into a single species network



- There are several programs for building networks from rooted triplets (LEVEL2, LEV1ATHAN, SIMPLISTIC)

- In theory the advantage for the user (above trees) is that it is not necessary to first construct entire gene trees; the user can instead choose to specify only high-quality fragments of them as input.

- Also possible to construct the rooted triplets from heterogeneous sources (because abstraction is "value free").

- Issues:
  - How do we generate good rooted triplets in the first place?
  - Input-side demands to ensure tractability are too restrictive
  - Small amount of noise can inflate the number of reticulations
  - Multiple solutions? Branch lengths? Bootstrapping?
  - **Lack of memory: topology is not preserved**

Case study 3: "Piecewise" methods: combining triplets into a single species network

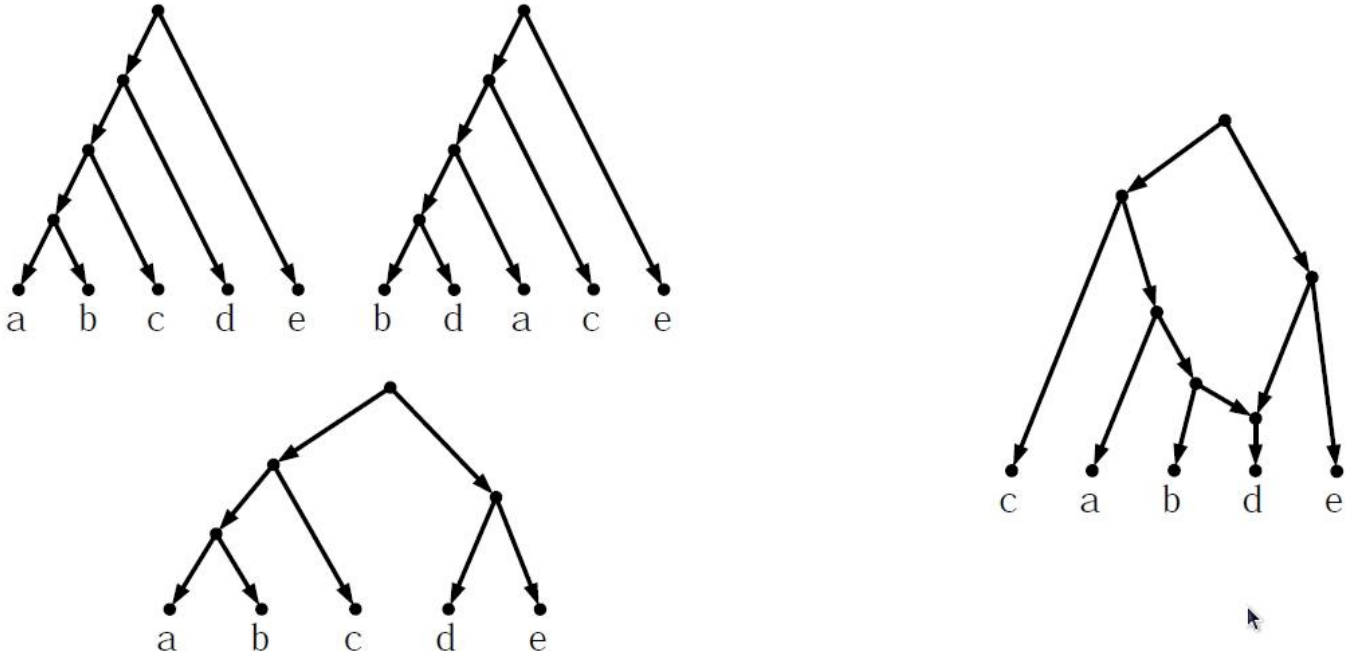


Figure 5. The level-1 network on the right with a single reticulation represents the union of the clusters (and triplets) obtained from the three trees on the left. However, any network that displays all three trees will have at least two reticulations and have level at least two.

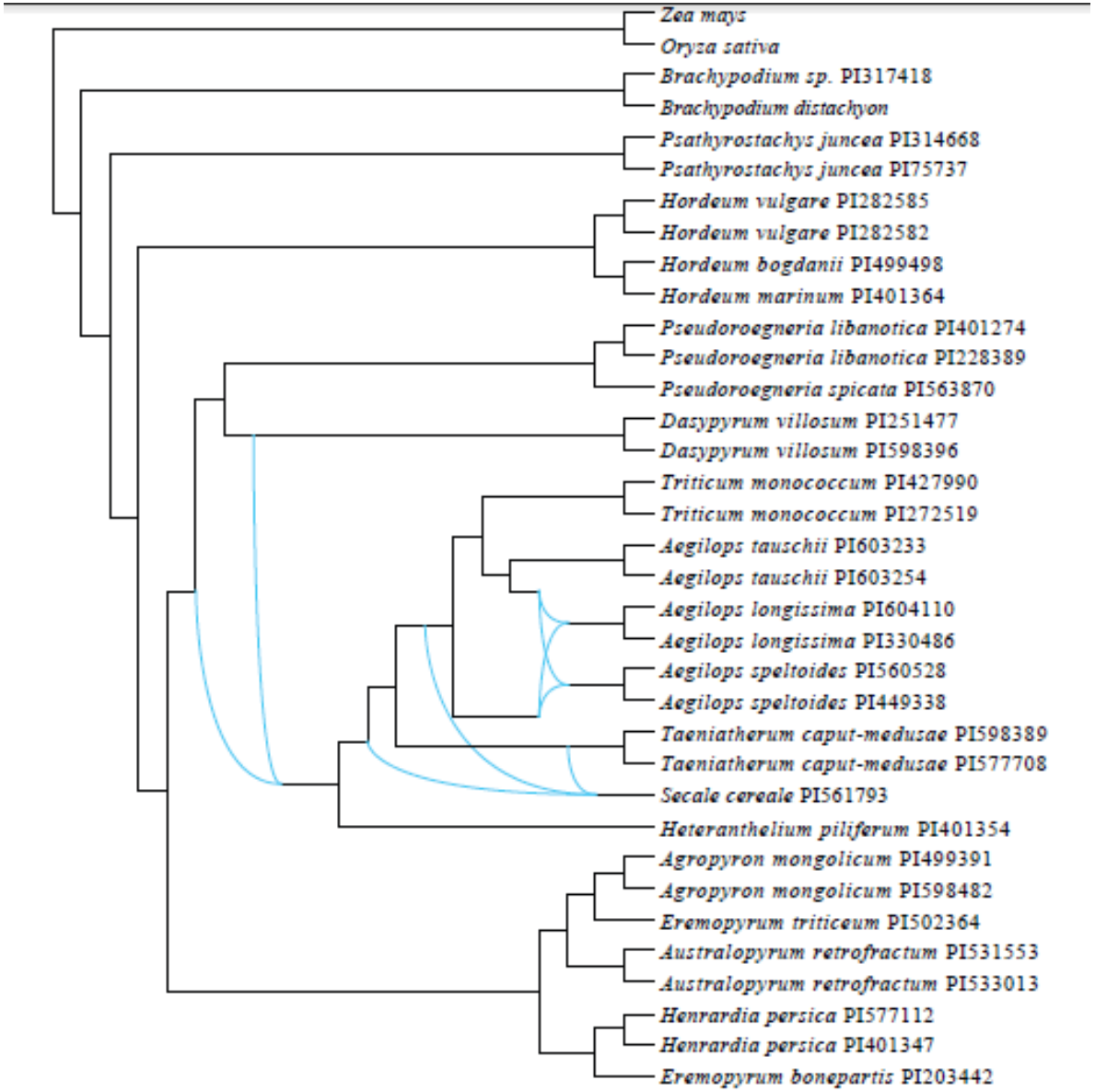
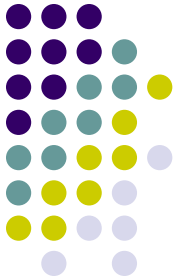


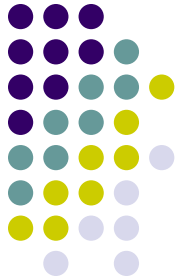
Figure 3 Multigenic network of Triticeae. Network obtained from the 27 individual gene trees modified with PhySIC_IST [56] using a correction threshold of 0.9 (see details in Methods).

From: **Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae)**, Escobar et al, BMC Evolutionary Biology 2011, 11:181



So…how far have we come? What do we still have to do?

So…how far have we come? What do we still have to do?



Summary of problems

• Optimal solutions are in general very difficult to compute. As a consequence it can be difficult to know whether solutions hypothesise (far) too much reticulation

• Algorithms for combining multiple gene trees into a single species network are too fixated by the **case of two binary gene trees**

• Piecewise methods often **forget the topology** of the gene trees they came from, and can be difficult to generate accurately

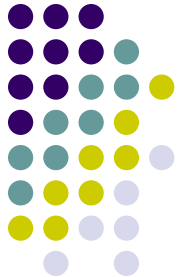
• Extensions to Maximum Parsimony and Maximum Likelihood are promising (and the people working on this really understand the biology) but are still in their **infancy**, advocate one specific model, and are computationally very hard

• Algorithms in general do not generate multiple optimal solutions and have no network equivalent of **common concepts** such as bootstrapping, branch-lengths etc.

• Mathematical abstractions sometimes too general, sometimes too rigid. Simultaneous expertise in combinatorial optimization, bioinformatics and biological modelling required (very rare!)

• Difficult to distinguish phenomena that distort tree signals

Ideas for the future (1/5)



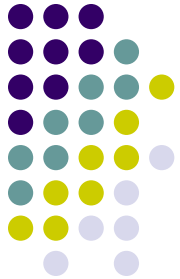
- **Remember the context...**

  - **“**“Everyone” seems to build phylogenetic trees, but “nobody” uses software for (evolutionary) phylogenetic networks. What’s going wrong?

  - Remember that the concept of “phylogenetic network” covers a very wide array of disparate evolutionary phenomena, many of which are still poorly understood.

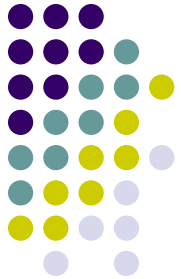
  - Is it realistic, then, to expect that there is **one model/software package to rule them all?** Perhaps it can and should remain a specialised phenomenon, adapted ad-hoc on a case-by-case basis?

Ideas for the future (2/5)



- **"But where are the bootstrap values?"**
  - Biologists always ask this ☺
  - Phylogenetic tree construction is so standardized that certain concepts (such as bootstrapping) are seen as essential.
  - It's therefore important to develop (standardized?) equivalents for phylogenetic network construction.
  - There is some reason for optimism here, since the question "how confident are you that this is the right solution?" can at least partially be answered in a model-neutral way.
  - My colleague Leo van Iersel (CWI, Amsterdam) has received a VENI grant to research this.

Ideas for the future (3/5)



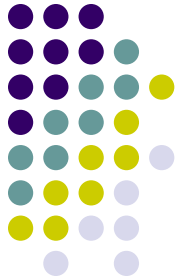
- **Choose the correct level of mathematical ambition**

  - Mathematicians and computer scientists like computationally hard problems that become intractable as the input size tends to infinity, because solving them well = publications.

  - But biologists are sometimes (often?) interested only in networks with a very small amount of reticulation.

  - So in some cases it might be better to spend less time on developing super-efficient, narrowly applicable software packages, and more time on less-efficient but “broader” software packages.

Ideas for the future (4/5)



- **Better co-ordination between computer scientists and biologists**

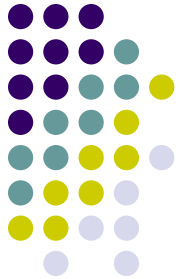
  - Scientists working on the algorithmic efficiency side of phylogenetic networks rarely have more than a superficial understanding of the biological model. More contact with biologists needed!

  - *"The future of phylogenetic networks"* – modelling workshop at Lorentz Center in Leiden, NL, under review

  - In the meantime: can we make a strength from this limitation?

  - Even when biological models differ, the underlying mathematical abstractions often have many common features (because of common DAG backbone). Perhaps a "standardised algorithmic layer" of routines can and should be developed?

Ideas for the future (5/5)



- **Towards a common modelling and hypothesis-testing language**

  - On the modelling side, it would be fascinating to try and develop a standardized modelling language for formally describing different phylogenetic network models

  - On the user-side we might also imagine a hypothesis-testing language e.g. *"Tell me how many phylogenetic networks contain at most 3 reticulation events and in which both species X and Y occur below a reticulation."*

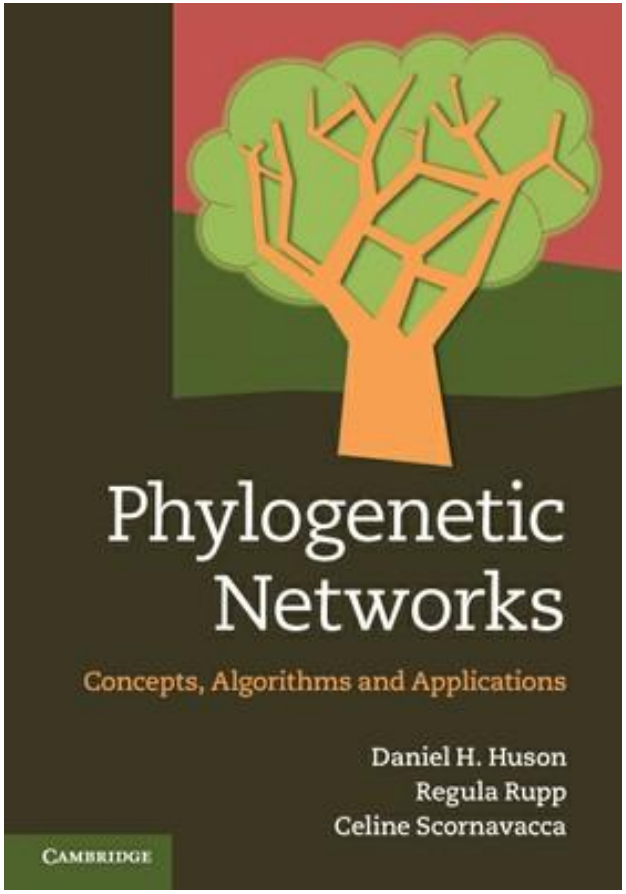
  - This might actually be plausible if biologists are mainly interested in solutions with a small amount of reticulation, because then there is more computational time available to spend on answering such advanced queries.

Finally…further reading

• Luay Nakhleh, *"Evolutionary phylogenetic networks: models and issues."* In: The Problem Solving Handbook for Computational Biology and Bioinformatics, L. Heath and N. Ramakrishnan (editors). Springer, 125-158, 2010.

• Daniel Huson, Regula Rupp and Celine Scornavacca, *“Phylogenetic Networks”*, Cambridge University Press.

• David Morrison, *“An introduction to phylogenetic networks”,* Dystenium LLC, New York, to appear in 2012.

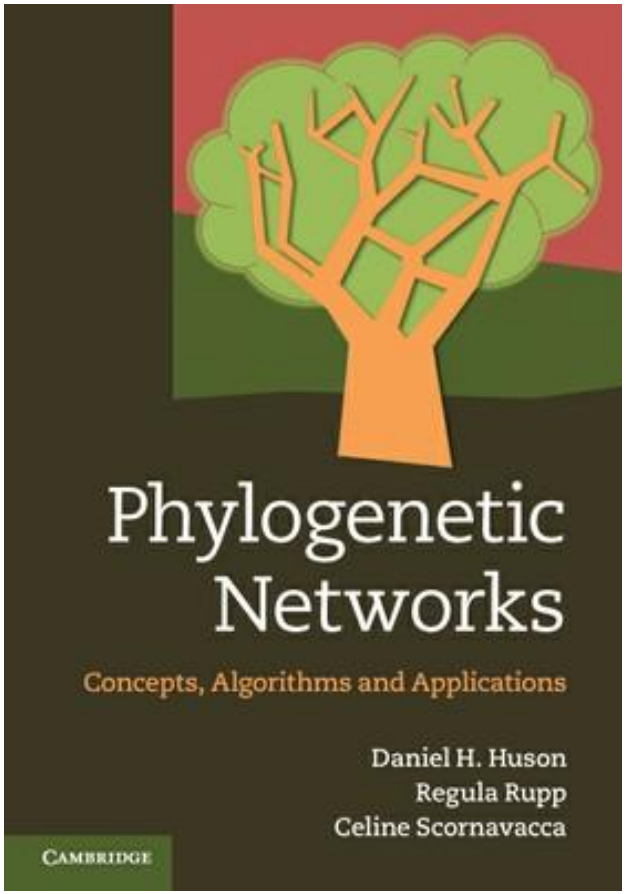


Finally…further reading

• Luay Nakhleh, *"Evolutionary phylogenetic networks: models and issues."* In: The Problem Solving Handbook for Computational Biology and Bioinformatics, L. Heath and N. Ramakrishnan (editors). Springer, 125-158, 2010.

• Daniel Huson, Regula Rupp and Celine Scornavacca, *“Phylogenetic Networks”*, Cambridge University Press.

• David Morrison, *“An introduction to phylogenetic networks”,* Dystenium LLC, New York, to appear in 2012.



Thanks for listening