From phylogenetic trees to phylogenetic networks: possible applications outside biology

Steven Kelk, Department of Knowledge Engineering (DKE), Maastricht University (NL).

Workshop “Capturing Phylogenetic Algorithms for Linguistics”, Leiden 2015

The Extended Chomsky Hierarchy

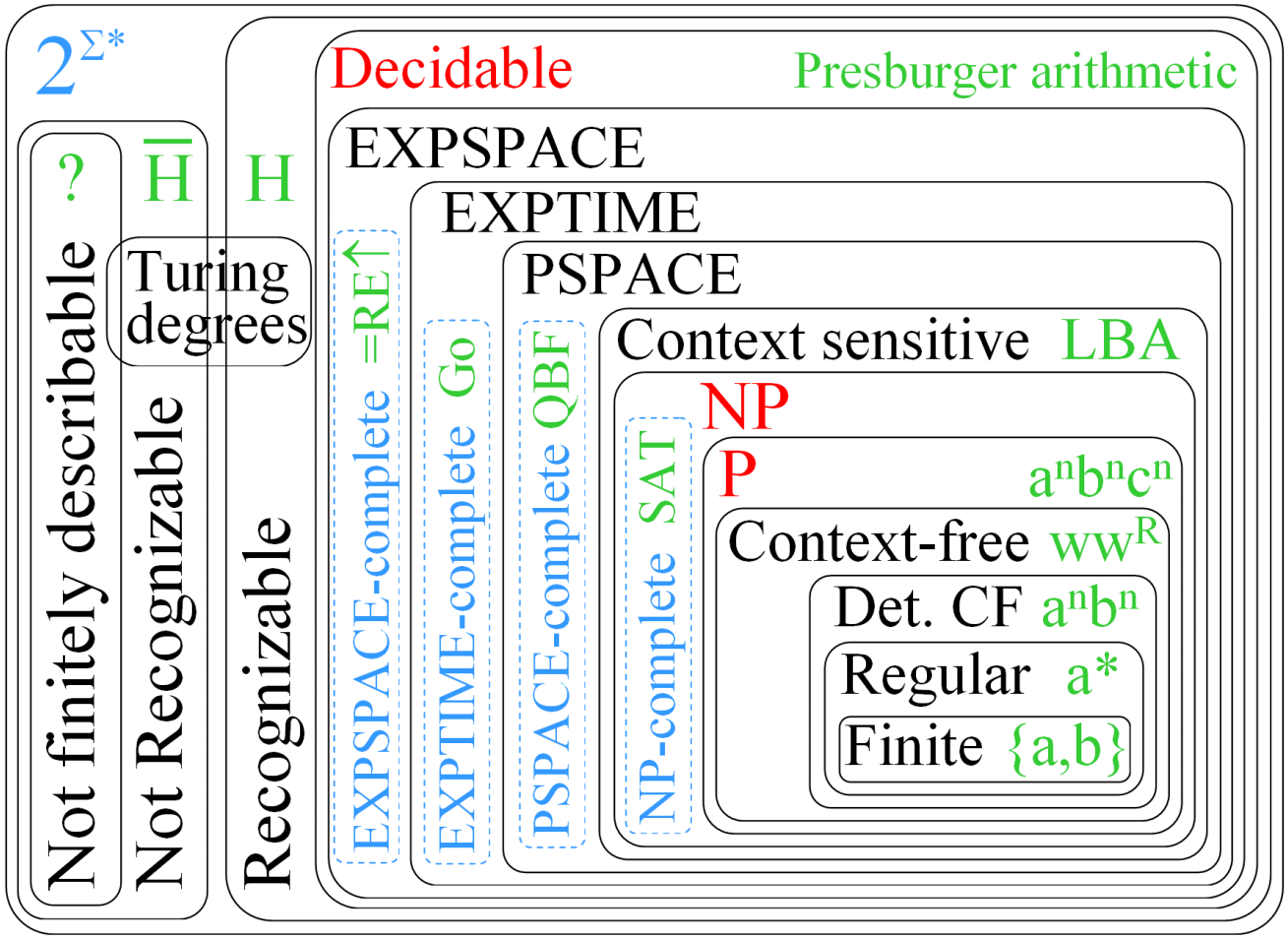
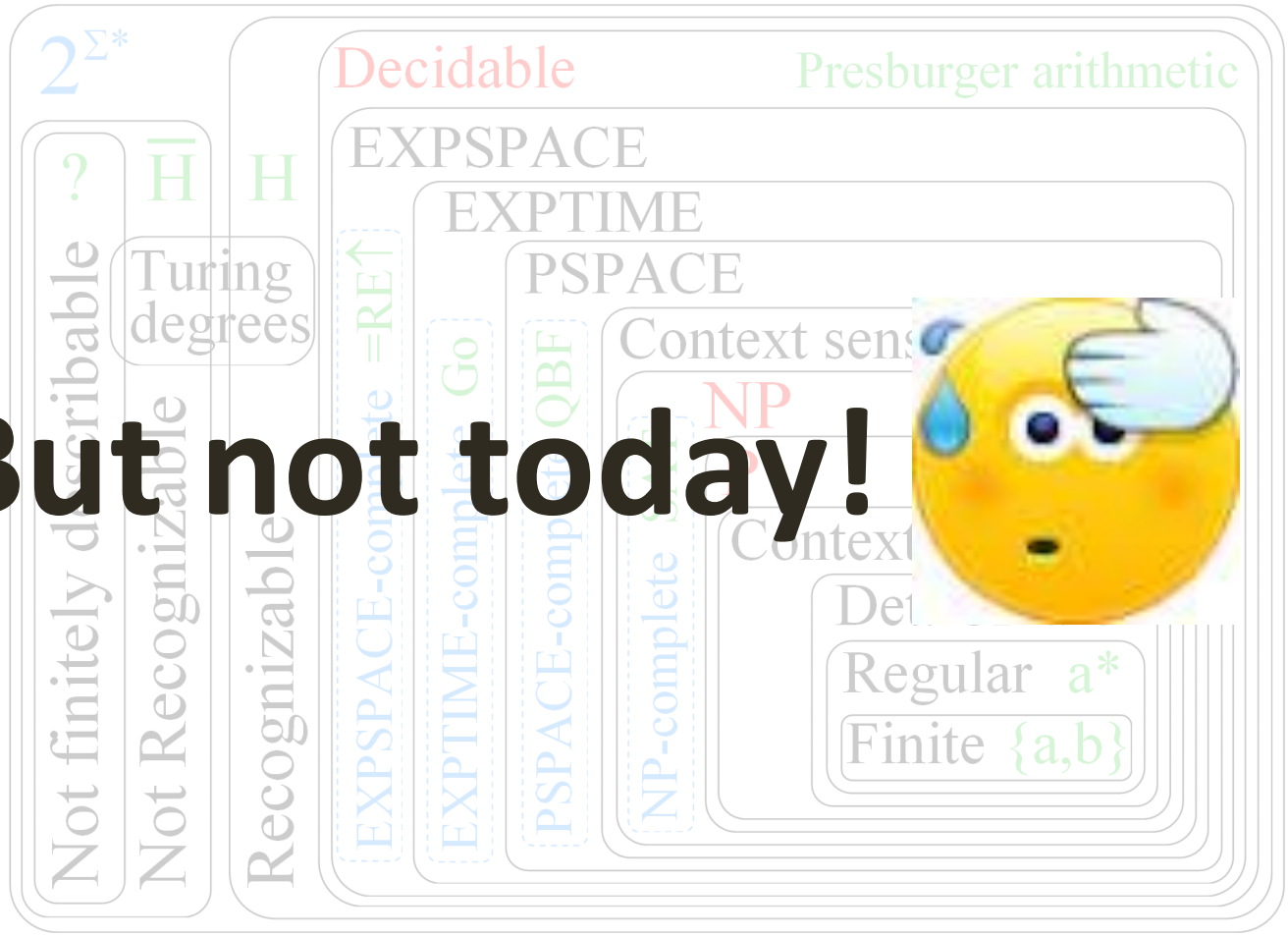


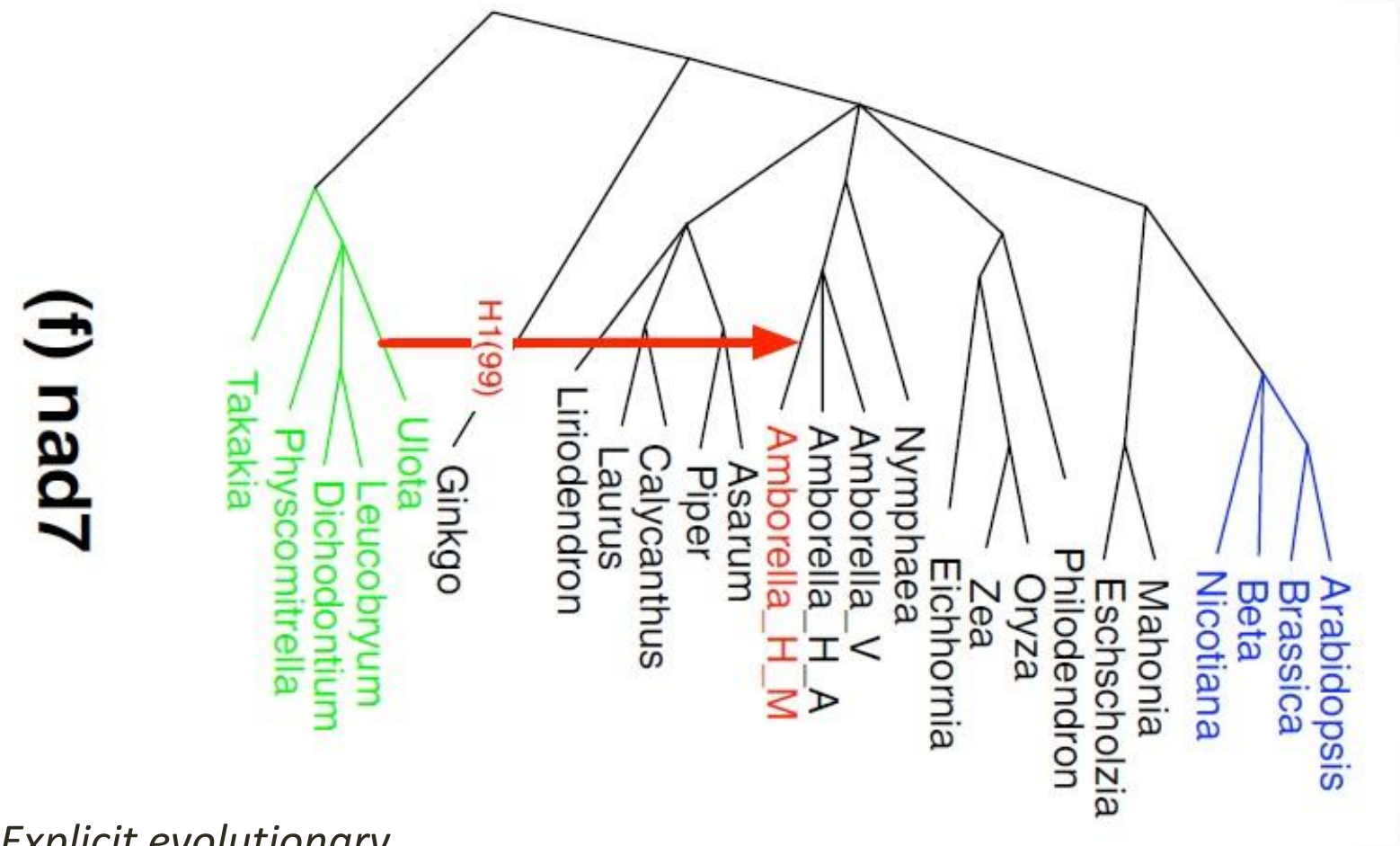
Figure from: http://www.cs.virginia.edu/~robins/cs3102/slides/Chomsky_large.gif

The Extended Chomsky Hierarchy



But not today!

Figure from: http://www.cs.virginia.edu/~robins/cs3102/slides/Chomsky_large.gif



*Explicit evolutionary hypothesis combining vertical + **horizontal** events*

Figure from: Bootstrap-based Support of HGT Inferred by Maximum Parsimony, Park et al. BMC Evolutionary Biology 2010, 10:131

Roadmap of talk

1. Biological motivation for phylogenetic networks
2. Models and algorithms – where are there opportunities for generalization to non-biological fields?
3. Looking forwards…

Roadmap of talk

1. Biological motivation for phylogenetic networks
2. Models and algorithms – where are there opportunities for generalization to non-biological fields?
3. Looking forwards…

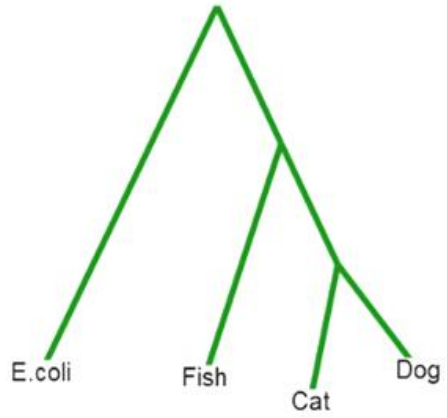
Computational complexity cannot be ignored!

Biological motivation for phylogenetic networks

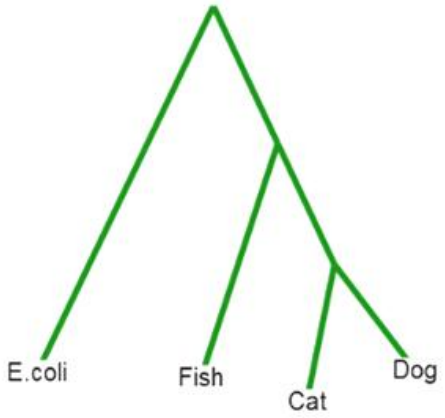
Gene trees, species trees

- The “classical” assumption:

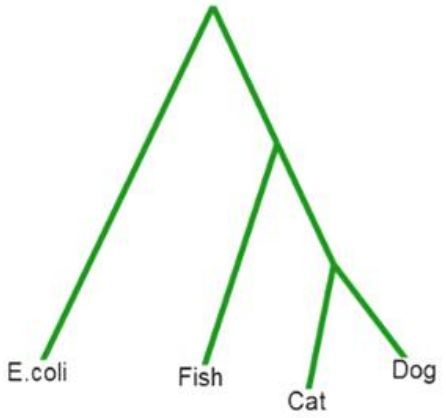
Gene 1 Gene 2 Gene 3



Gene Tree 1



Gene Tree 2

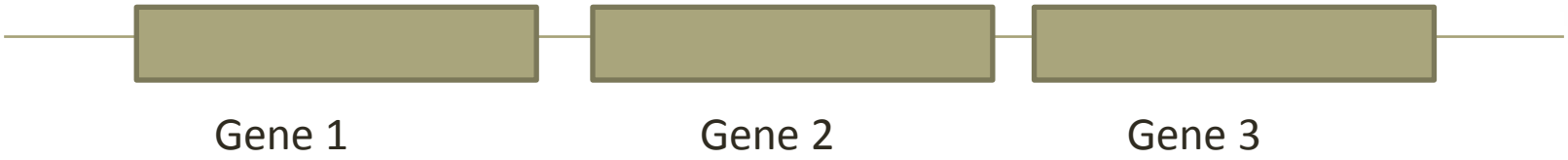


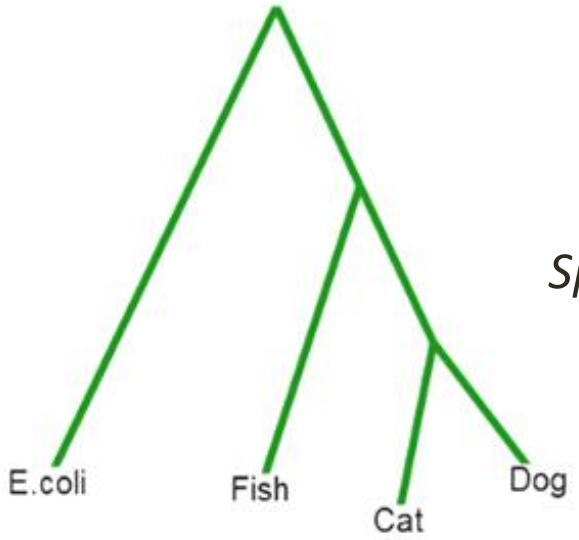
Gene Tree 3

Different genes, same tree

Gene trees, species trees

- The “classical” assumption:

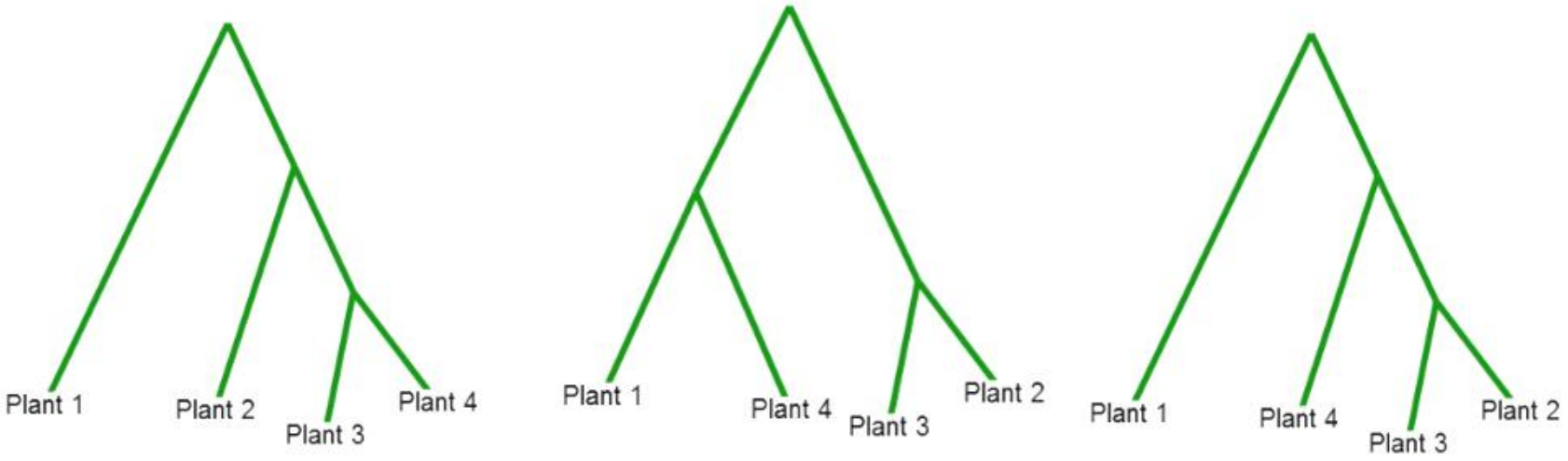




Species tree (same as the gene trees)

Gene trees, species trees

- But, as more data becomes available, we often see...



Three distinct gene trees!

So what is the species tree? Can we even say there is a species "tree"?

One tree to rule them all…?

- There is nothing wrong with the idea of tree-like evolution.
- What *is* wrong with the classical view of evolution, is that there is a *single* tree that can simultaneously explain everything, in all cases.
- The reality is more complex. There are often *multiple* conflicting (*"incongruent"*) tree signals involved.
- There are actually many different evolutionary phenomena that can cause multiple conflicting tree signals to arise.

Table 2.1. Causes of reticulation in phylogenetic analyses

Estimation errors

(i) incorrect data
- inadequate data-collection protocol
- poor laboratory / museum / herbarium technique
- lack of quality control after data collection
- misadventure

(ii) inappropriate sampling
- distant outgroup
- rapid evolutionary rates
- short internal branches

(iii) model mis-specification
- wrong assessment of primary homology
- wrong substitution model
- different optimality criteria

Biological conflict

(iv) analogy
- parallelism
- convergence
- reversal

(v) homology
- hybridization
- introgression
- recombination
- horizontal gene transfer
- genome fusion
- deep coalescence
- duplication–loss

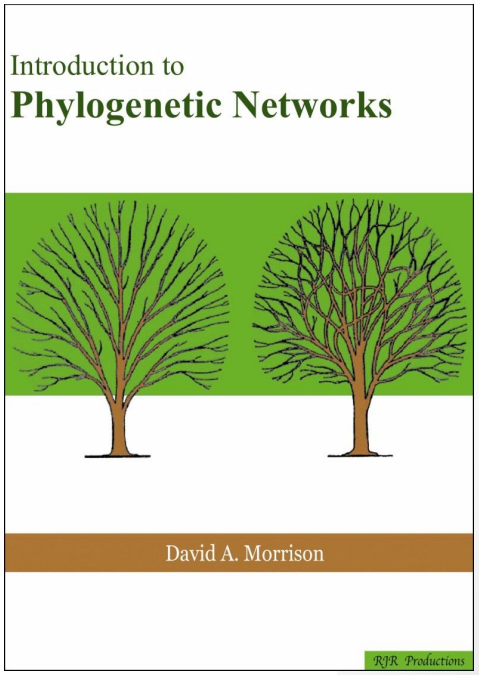


Table 2.1. Causes of reticulation in phylogenetic analyses

Estimation errors

(i) incorrect data
- inadequate data-collection protocol
- poor laboratory / museum / herbarium technique
- lack of quality control after data collection
- misadventure

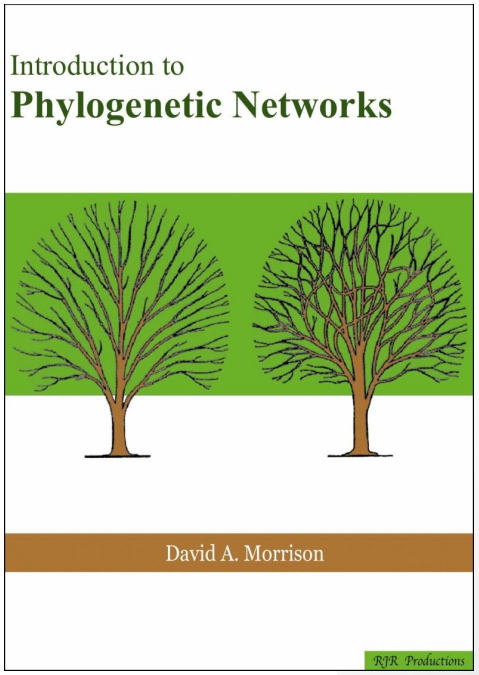
(ii) inappropriate sampling
- distant outgroup
- rapid evolutionary rates
- short internal branches

(iii) model mis-specification
- wrong assessment of primary homology
- wrong substitution model
- different optimality criteria

Biological conflict

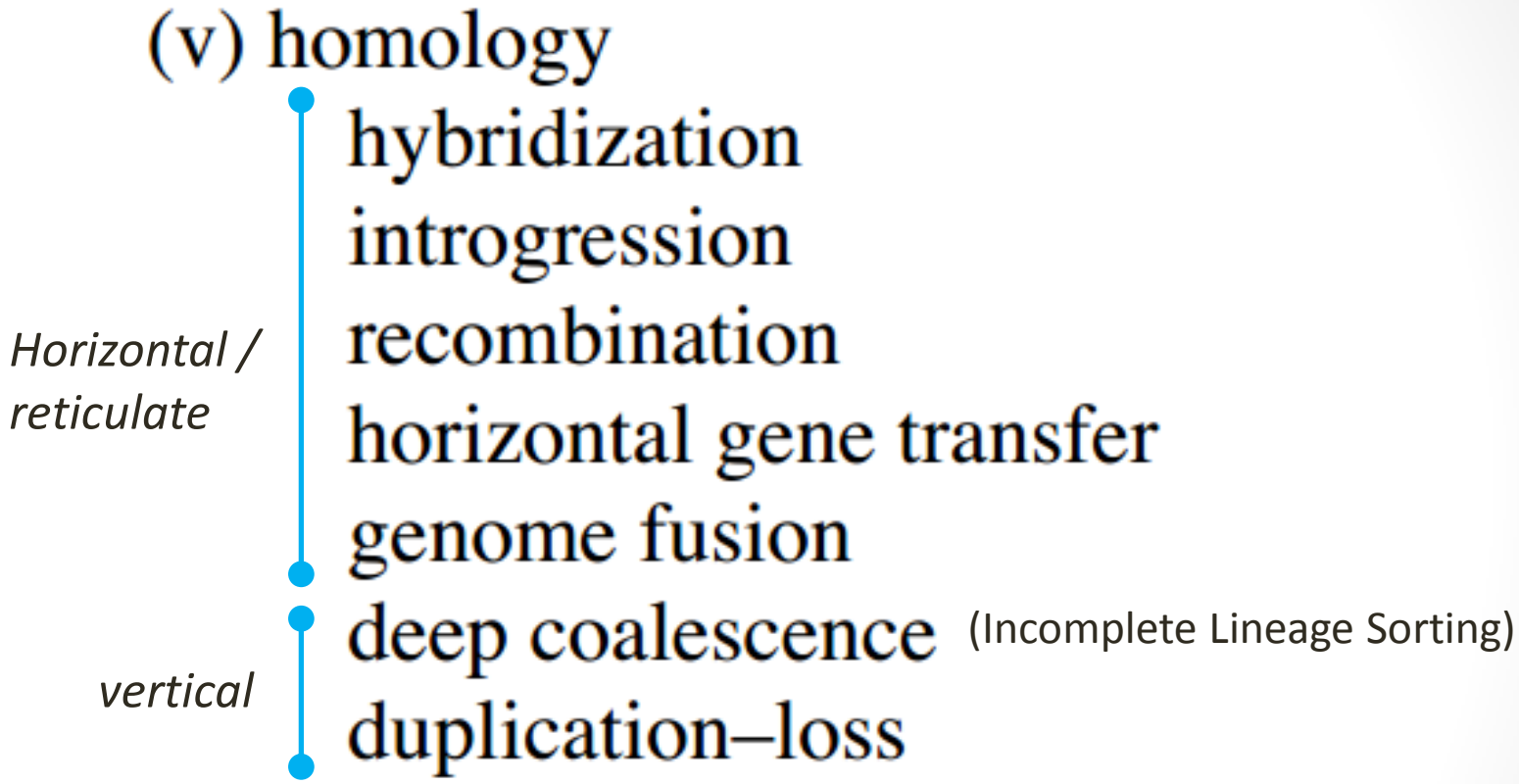
(iv) analogy
- parallelism
- convergence
- reversal

(v) homology
- hybridization
- introgression
- recombination
- horizontal gene transfer
- genome fusion
- deep coalescence
- duplication–loss



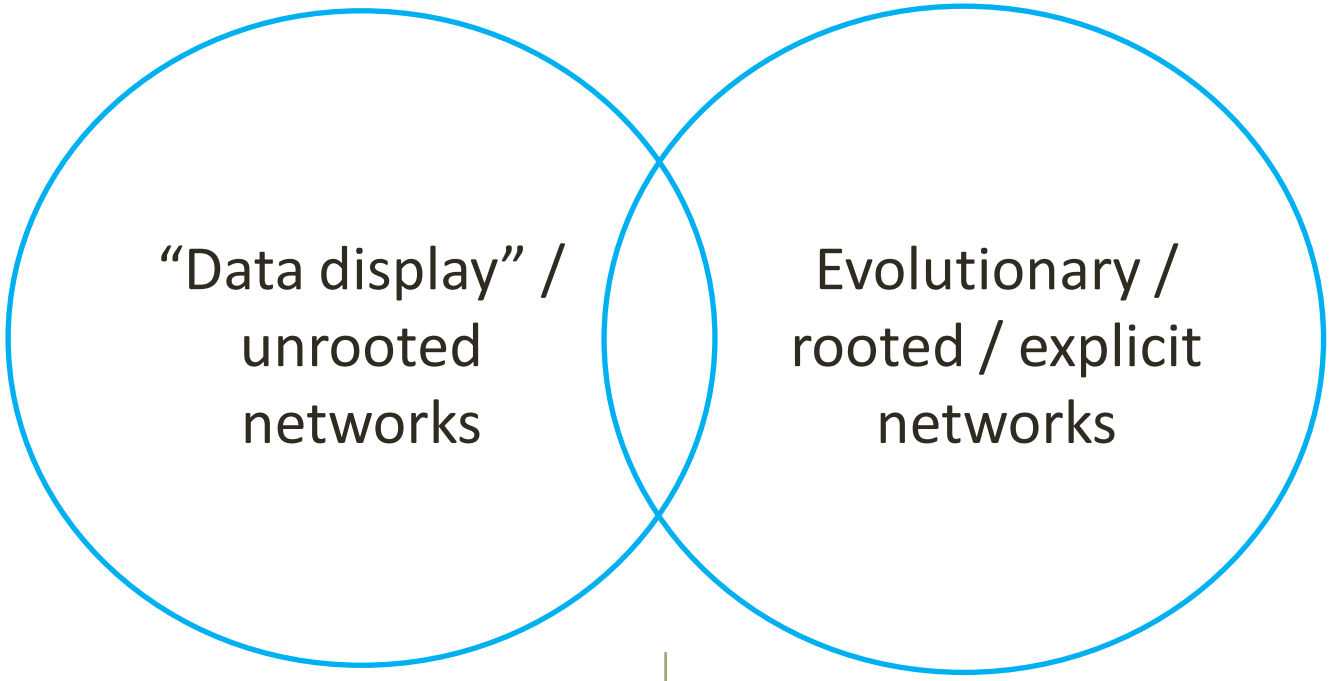
(v) homology

- hybridization
- introgression
- recombination
- horizontal gene transfer
- genome fusion
- deep coalescence (Incomplete Lineage Sorting)
- duplication–loss


(v) homology
hybridization
introgression
recombination
horizontal gene transfer
genome fusion
deep coalescence
duplication–loss
Horizontal / reticulate
vertical
(Incomplete Lineage Sorting)

Models and algorithms

Phylogenetic networks: 2 types



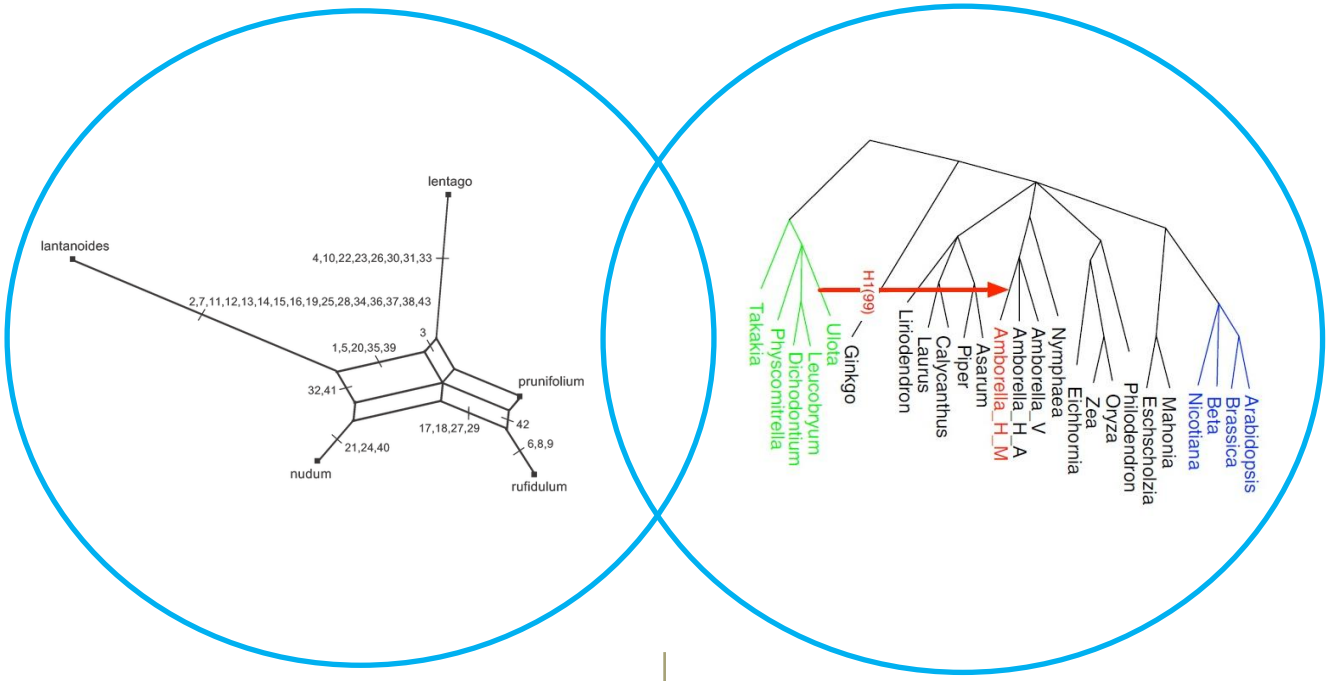
No (explicit) model of evolution: tries to graphically represent **where** the data is non-treelike.

Does not generate a hypothesis of "what happened".

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

Tries – in some limited way – to generate a hypothesis of "what happened".

Phylogenetic networks: 2 types



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

Does not generate a hypothesis of "what happened".

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

Tries – in some limited way – to generate a hypothesis of "what happened".

Briefly: data-display networks

*This example taken from **Primer of Phylogenetic Networks** by David Morrison, http://acacia.atspace.eu/Tutorial/Tutorial.html*

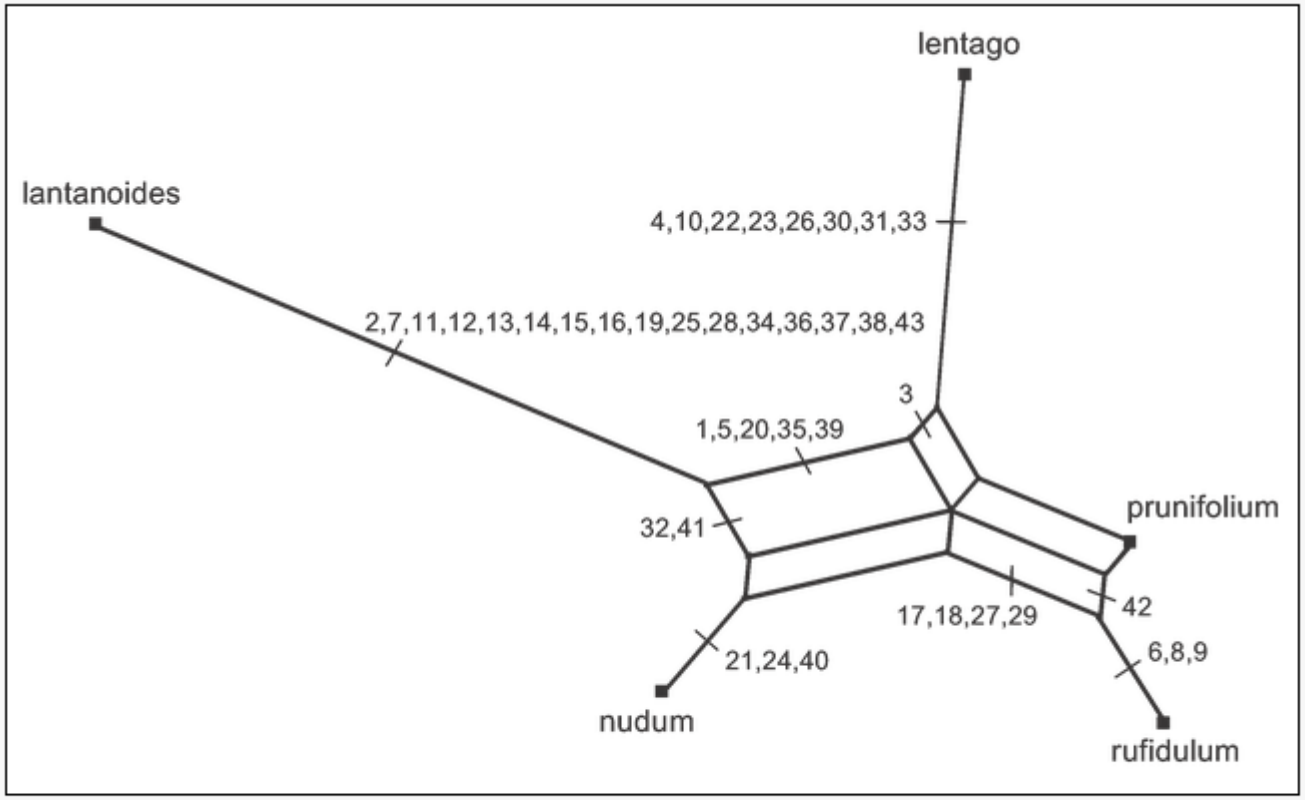


Figure 6. The Median Network for the *Viburnum* sequence, showing the edges (or sets of parallel edges) associated with each of the 43 characters.

Briefly: data-display networks

Characters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
Viburnum lantanoides	A	C	T	G	A	G	T	C	T	T	C	G	T	T	T	A	T	A	T	T	T	T	C	T	T	C	T	G	T	T	A	C	G	A	T	C	A	T	A	T	T	C	A
Viburnum nudum	A	T	T	G	A	G	C	C	T	T	T	A	C	C	A	G	T	A	C	T	C	T	C	C	C	C	T	C	T	T	A	T	G	C	T	A	G	C	A	C	C	T	G
Viburnum rufidulum	G	T	T	G	C	A	C	T	G	T	T	A	C	C	A	G	C	G	C	C	T	T	C	T	C	C	C	C	C	T	A	T	G	C	C	A	G	C	G	T	C	T	G
Viburnum prunifolium	G	T	C	G	C	G	C	C	T	T	T	A	C	C	A	G	C	G	C	C	T	T	C	T	C	C	C	C	C	T	A	T	G	C	C	A	G	C	G	T	C	C	G
Viburnum lentago	G	T	C	T	C	G	C	C	T	G	T	A	C	C	A	G	T	A	C	C	T	C	A	T	C	G	T	C	T	C	G	C	A	C	C	A	G	C	G	T	T	C	G

In each of the 43 DNA sites, at most two different DNA characters are observed. So each site induces a bipartition. In this way, there are 9 different bipartitions possible, shown below. (Note that the original numbering of the DNA sites is lost in the figure below).

Viburnum lantanoides	C	C	C	T	T	T	T	T	T	G	G	A	A	A	A	T	T	A	A	A	T	C	C	T	T	T	G	G	A	C	T	G	T	T	T	A	T	T	T	C	C	T
Viburnum nudum	T	T	A	C	C	C	C	C	C	A	A	C	C	G	G	G	T	T	A	A	A	T	C	C	T	T	T	G	G	A	C	T	G	T	T	T	A	C	C	C	T	T	C
Viburnum rufidulum	T	T	A	C	C	C	C	C	C	A	A	C	C	G	G	G	C	C	G	C	G	T	C	C	T	T	T	G	G	A	T	G	A	C	C	C	G	T	T	T	T	T	C
Viburnum prunifolium	T	T	A	C	C	C	C	C	C	A	A	C	C	G	G	G	C	C	G	C	G	C	C	C	T	T	T	G	G	A	C	T	G	C	C	C	G	T	T	T	C	T	C
Viburnum lentago	T	T	A	C	C	C	C	C	C	A	A	C	C	G	G	G	C	C	G	C	G	C	A	G	G	C	C	T	A	G	C	T	G	T	T	T	A	T	T	T	C	C	T

Each parallel set of edges in the network, represents one of these 9 bipartitions.

Briefly: data-display networks

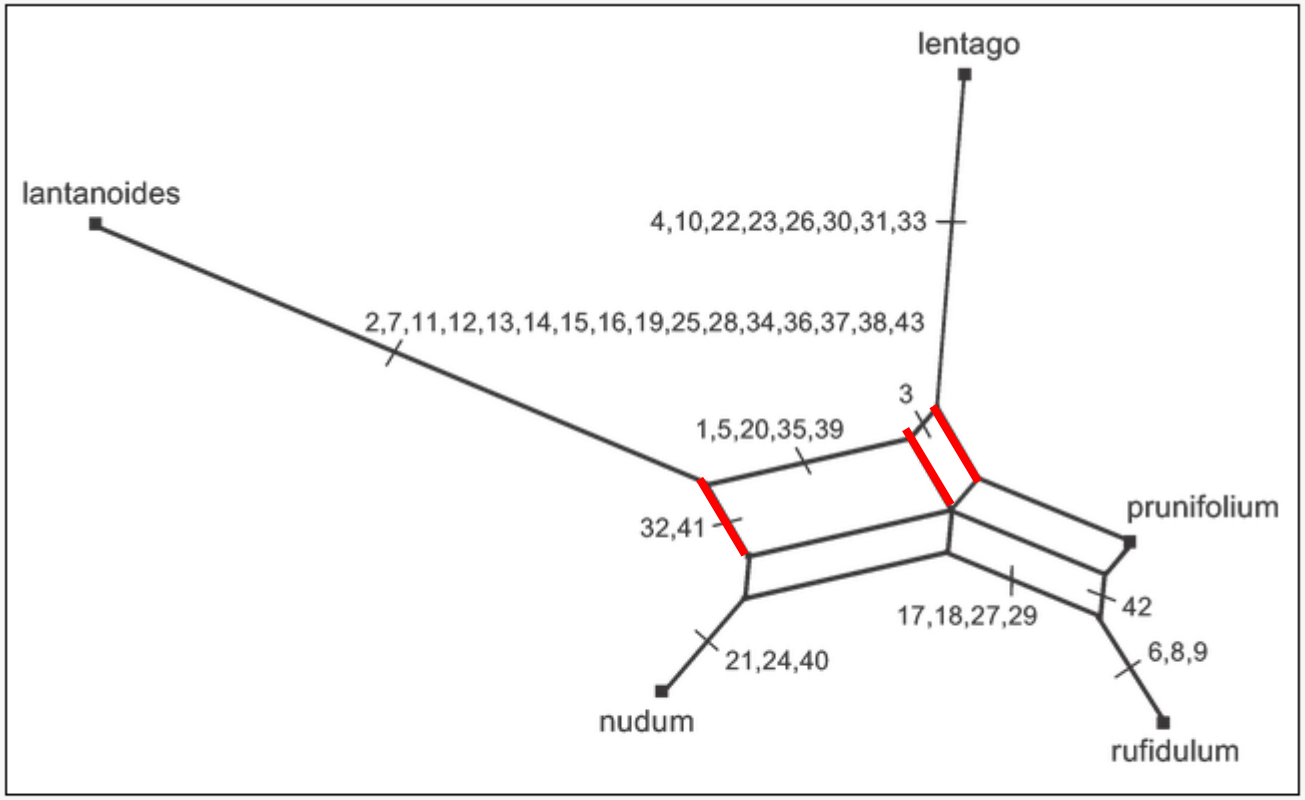


Figure 6. The Median Network for the *Viburnum* sequence, showing the edges (or sets of parallel edges) associated with each of the 43 characters.

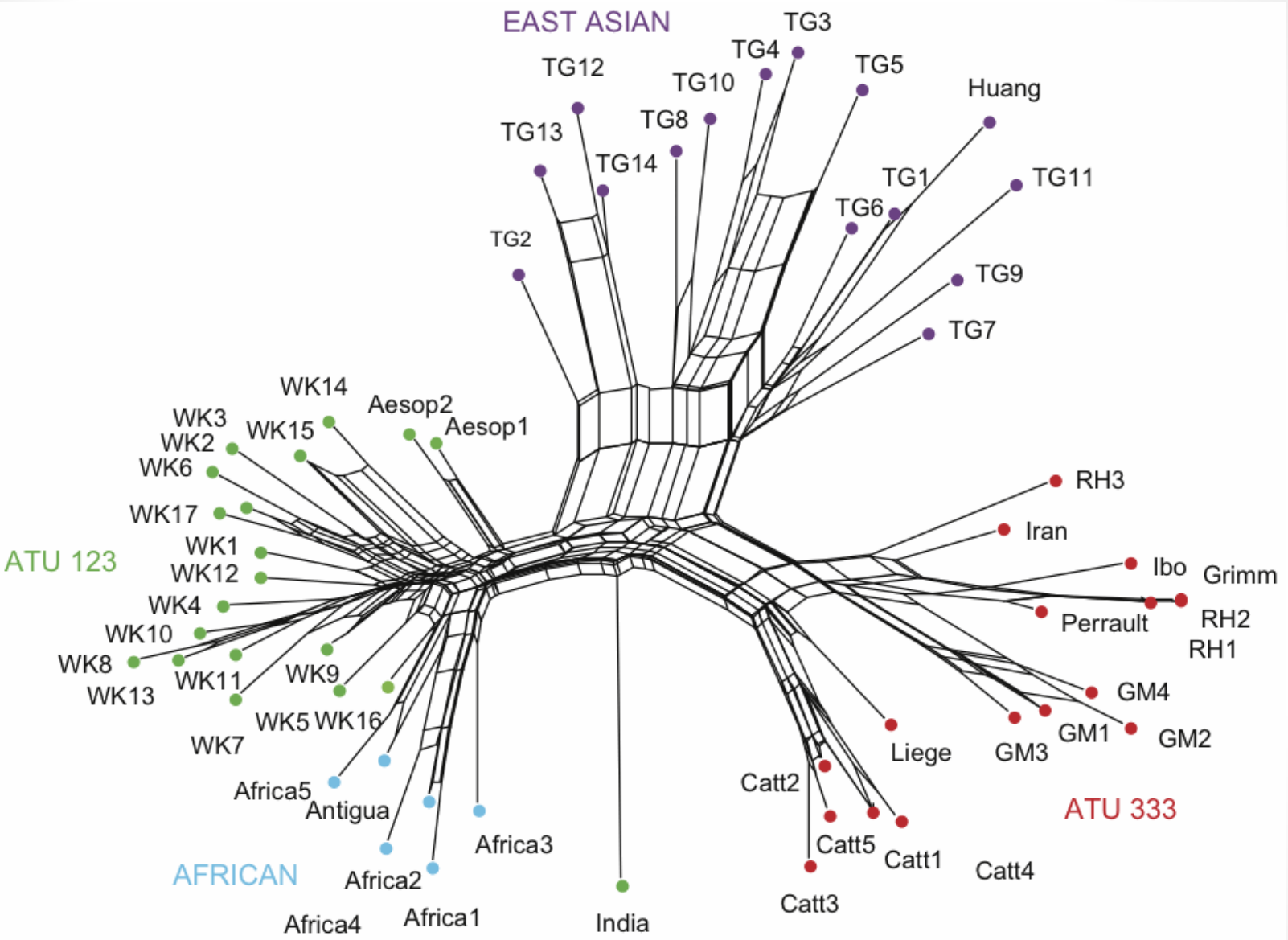
Briefly: data-display networks

- In practice data-display phylogenetic networks are still used far more than evolutionary phylogenetic networks.
- Why? Because they let the biologist *explore* the data, and to draw his/her own conclusions. They do not impose a (probably controversial…) model on the biologist.

Briefly: data-display networks

- In practice data-display phylogenetic networks are still used far more than evolutionary phylogenetic networks.
- Why? Because they let the biologist *explore* the data, and to draw his/her own conclusions. They do not impose a (probably controversial…) model on the biologist.

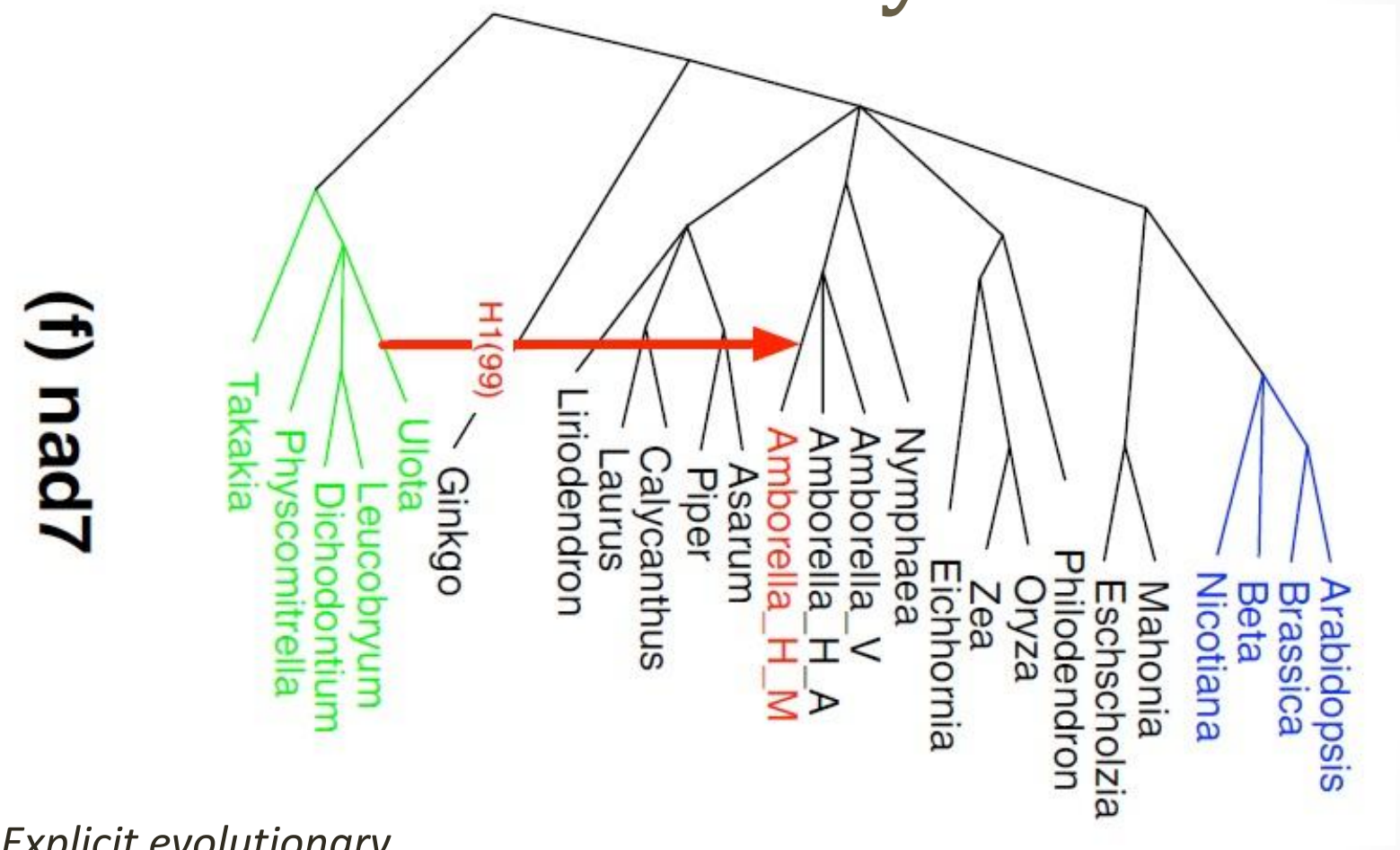
- **Note that nothing in this example is specific to biology! The characters are binary and unordered. Such data display networks are indeed already being used outside biology.**


EAST ASIAN
TG3
TG4
TG5
TG12
TG10
Huang
TG13
TG8
TG14
TG1
TG11
TG6
TG2
TG9
TG7
WK14
Aesop2
WK3
WK15
Aesop1
WK2
WK6
RH3
WK17
Iran
ATU 123
WK1
Ibo
Grimm
WK12
WK4
Perrault
RH2
WK10
RH1
WK8
WK11
WK9
GM4
WK13
WK5
WK16
GM1
GM2
WK7
Liege
GM3
Catt2
Africa5
ATU 333
Antigua
Catt5
Catt1
Africa3
AFRICAN
Catt4
Africa2
Catt3
Africa4
Africa1
India

The Phylogeny of Little Red Riding Hood, Jamshid J. Tehrani, PLOS One 2013

The Phylogeny of Little Red Riding Hood, Jamshid J. Tehrani, PLOS One 2013

Back to evolutionary networks



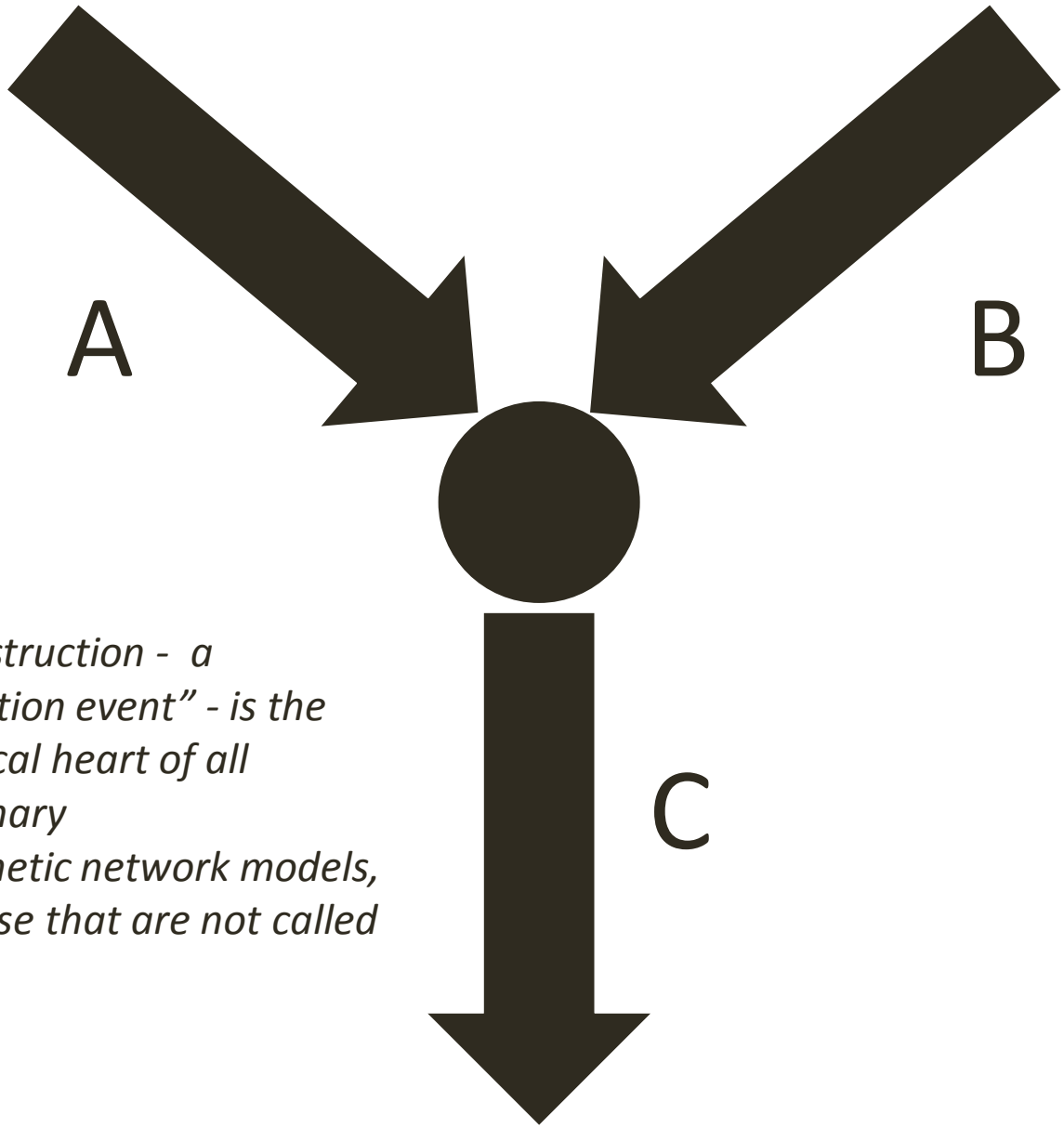
*Explicit evolutionary hypothesis combining vertical + **horizontal** events*

Figure from: Bootstrap-based Support of HGT Inferred by Maximum Parsimony, Park et al. BMC Evolutionary Biology 2010, 10:131

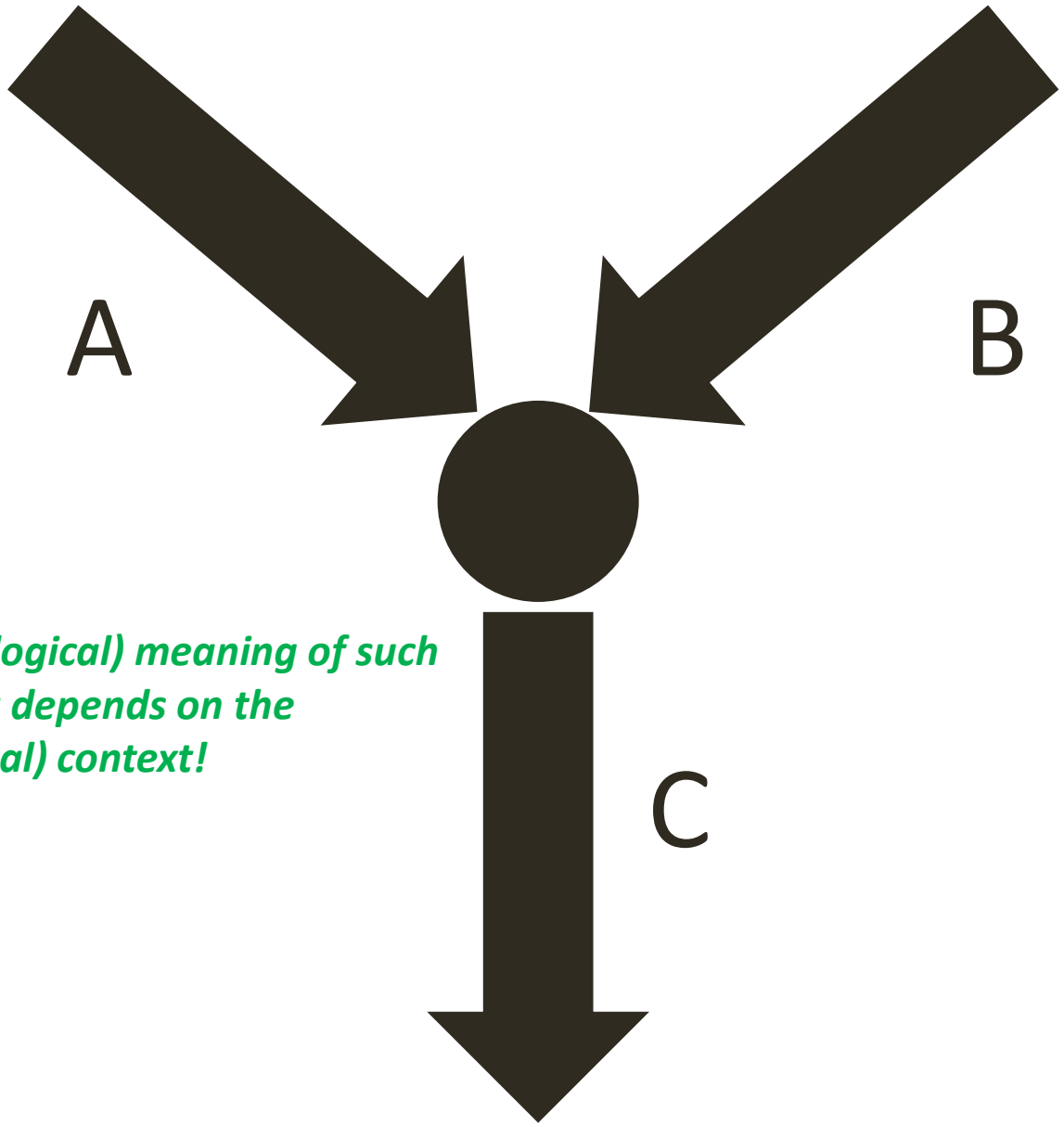
Back to evolutionary networks

- Source of confusion: evolutionary phylogenetic networks appear under various different names and guises in different branches of (comparative) evolutionary biology.
  - Trees with edges added to denote horizontal gene transfer events
  - Deletion-Loss-Transfer species-gene tree reconciliation scenarios
  - Ancestral Recombination Graphs (ARGs)
  - Hybridization networks
  - …

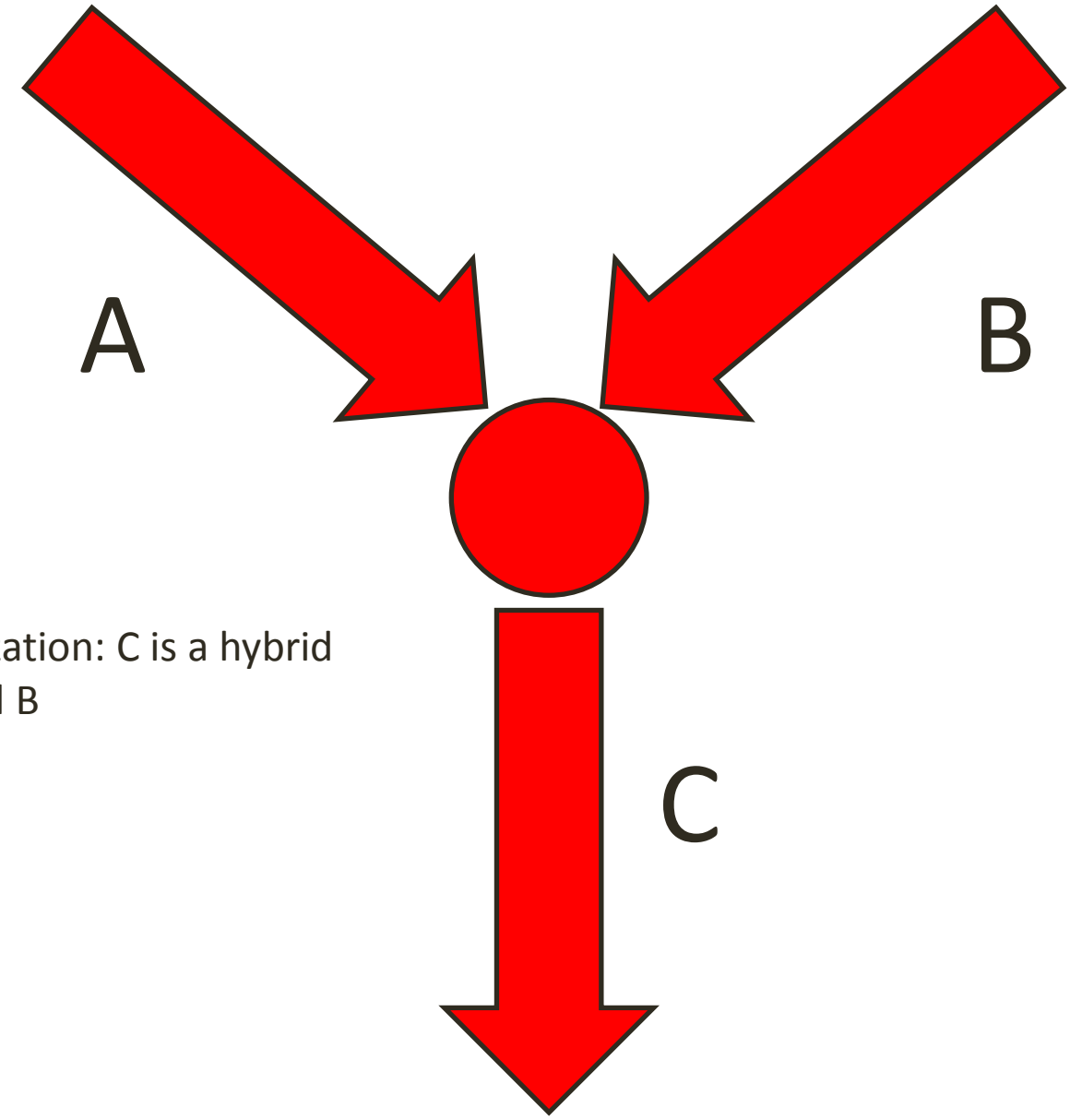
“A rose by any other name, would smell as sweet…”



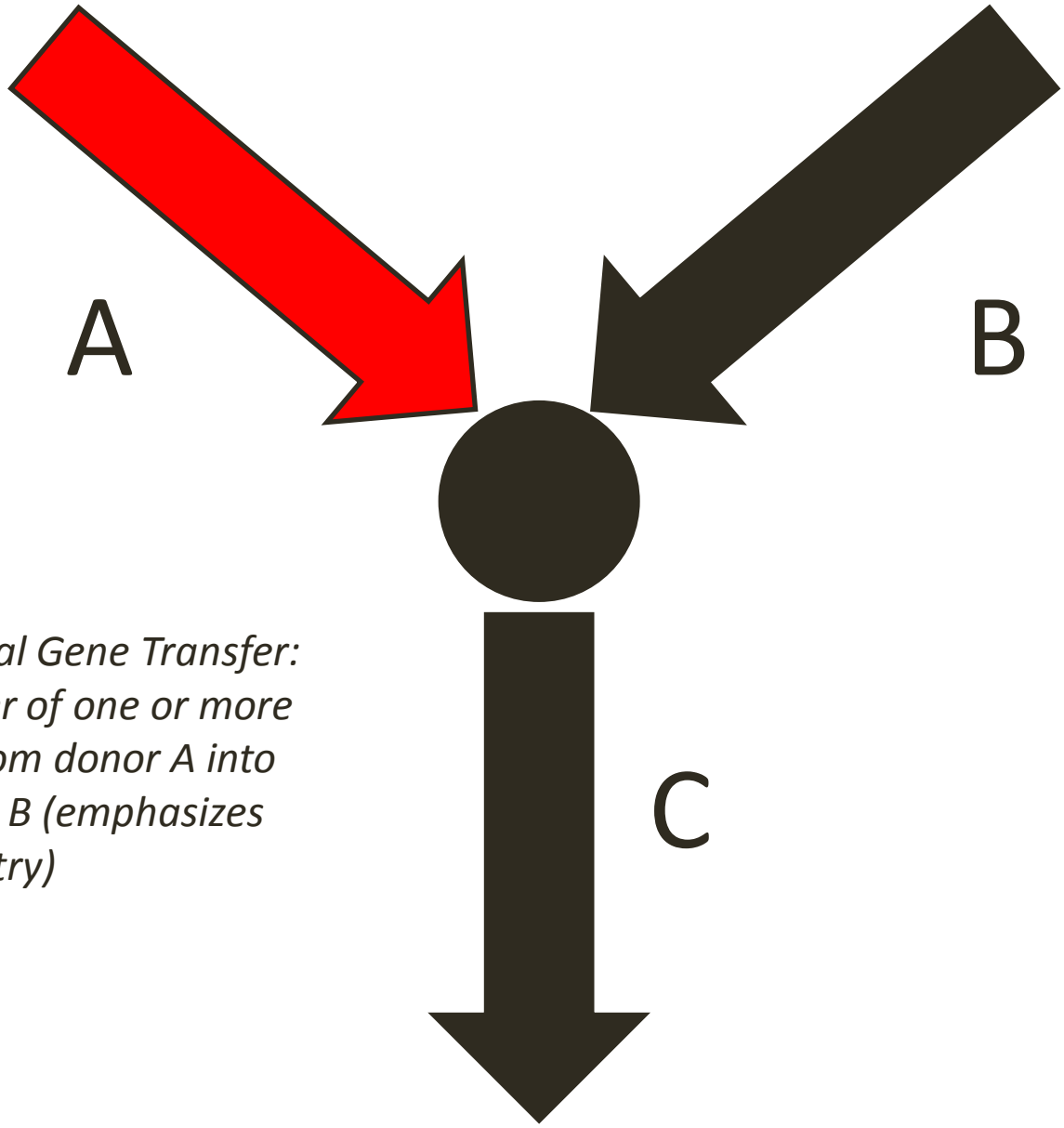
This construction - a "reticulation event" - is the topological heart of all evolutionary phylogenetic network models, even those that are not called as such…



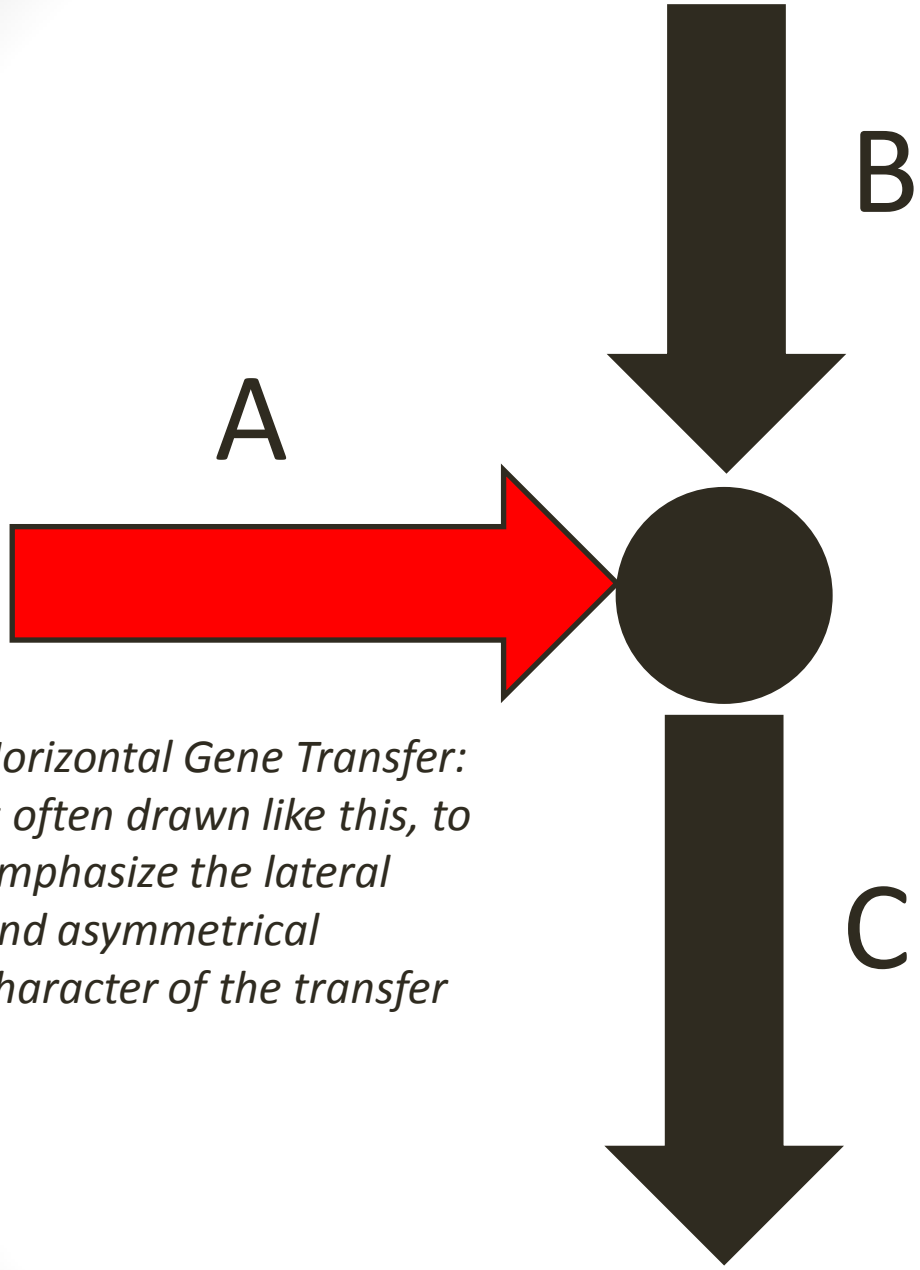
The (biological) meaning of such an event depends on the (biological) context!



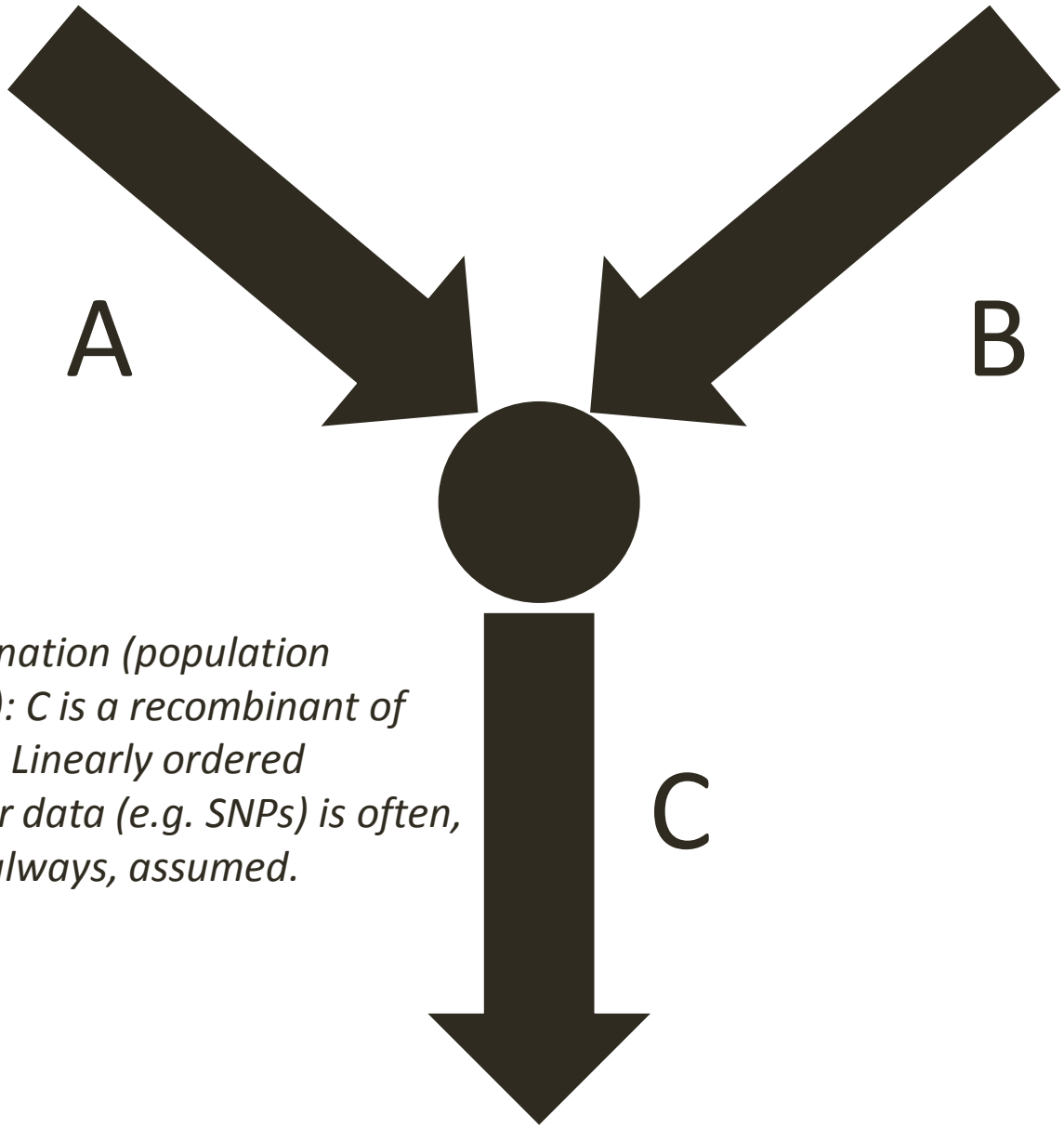
Hybridization: C is a hybrid of A and B



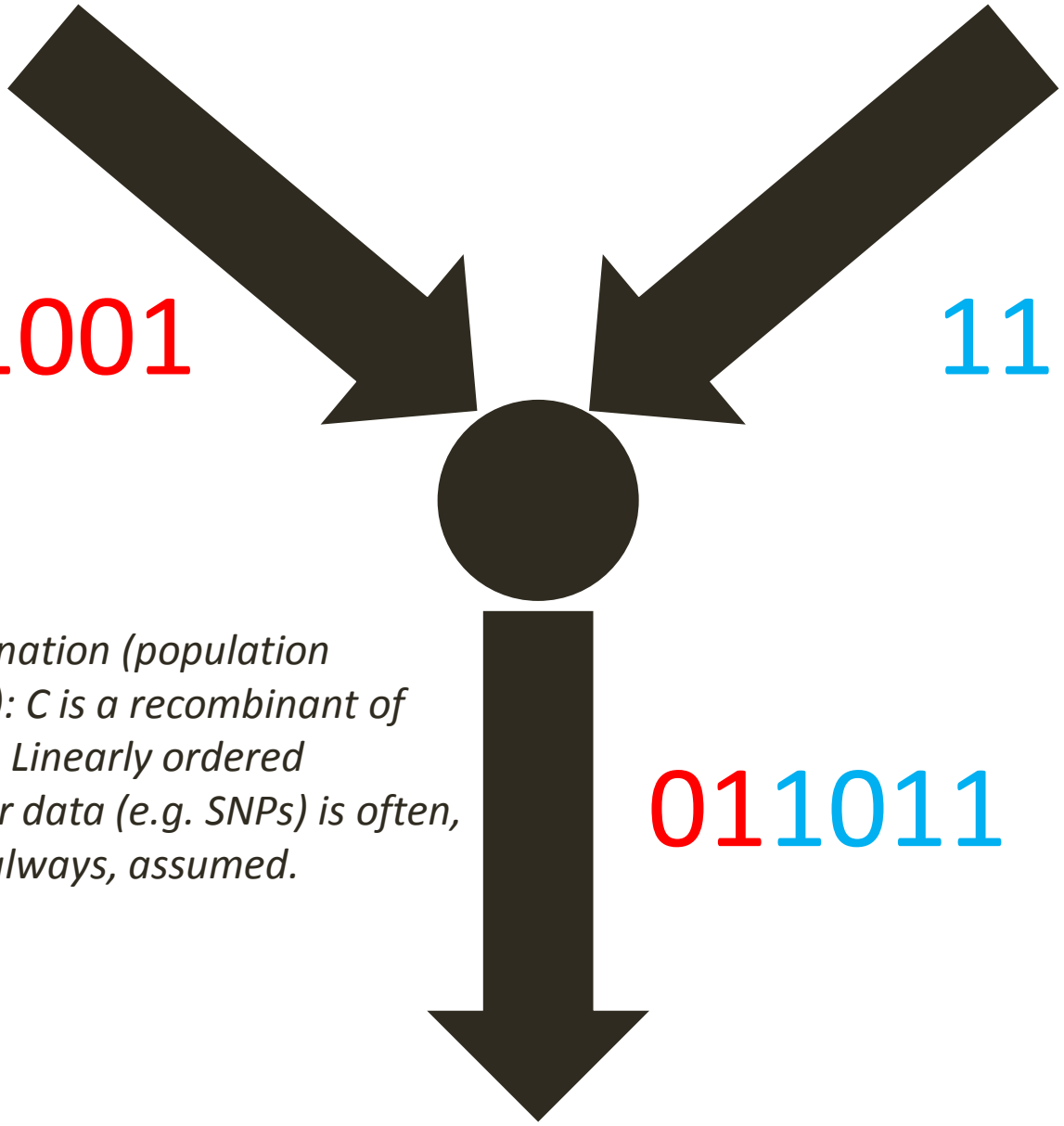
Horizontal Gene Transfer: a transfer of one or more genes from donor A into recipient B (emphasizes asymmetry)



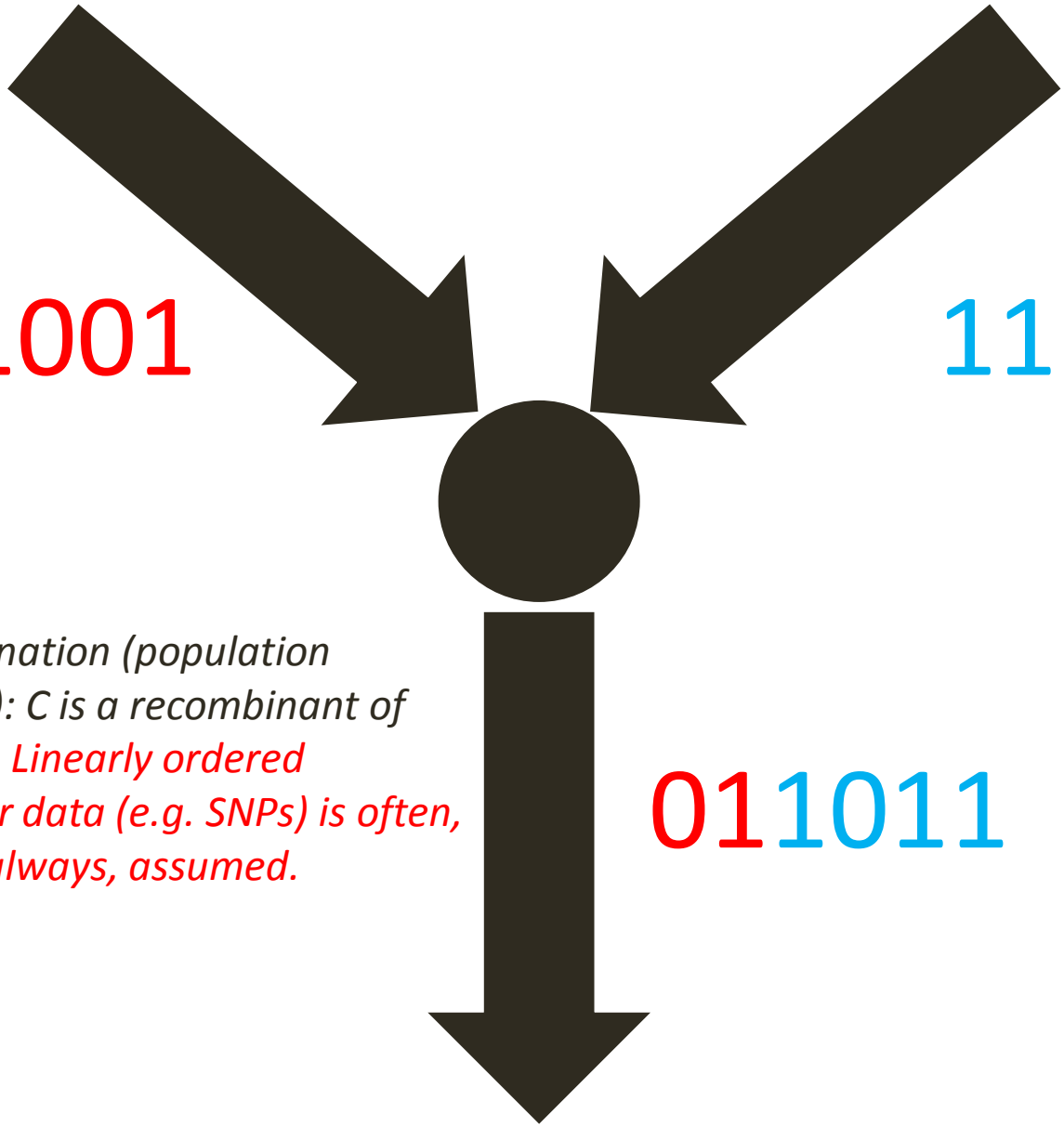
Horizontal Gene Transfer: is often drawn like this, to emphasize the lateral and asymmetrical character of the transfer



Recombination (population genetics): C is a recombinant of A and B. Linearly ordered character data (e.g. SNPs) is often, but not always, assumed.



Recombination (population genetics): C is a recombinant of A and B. Linearly ordered character data (e.g. SNPs) is often, but not always, assumed.

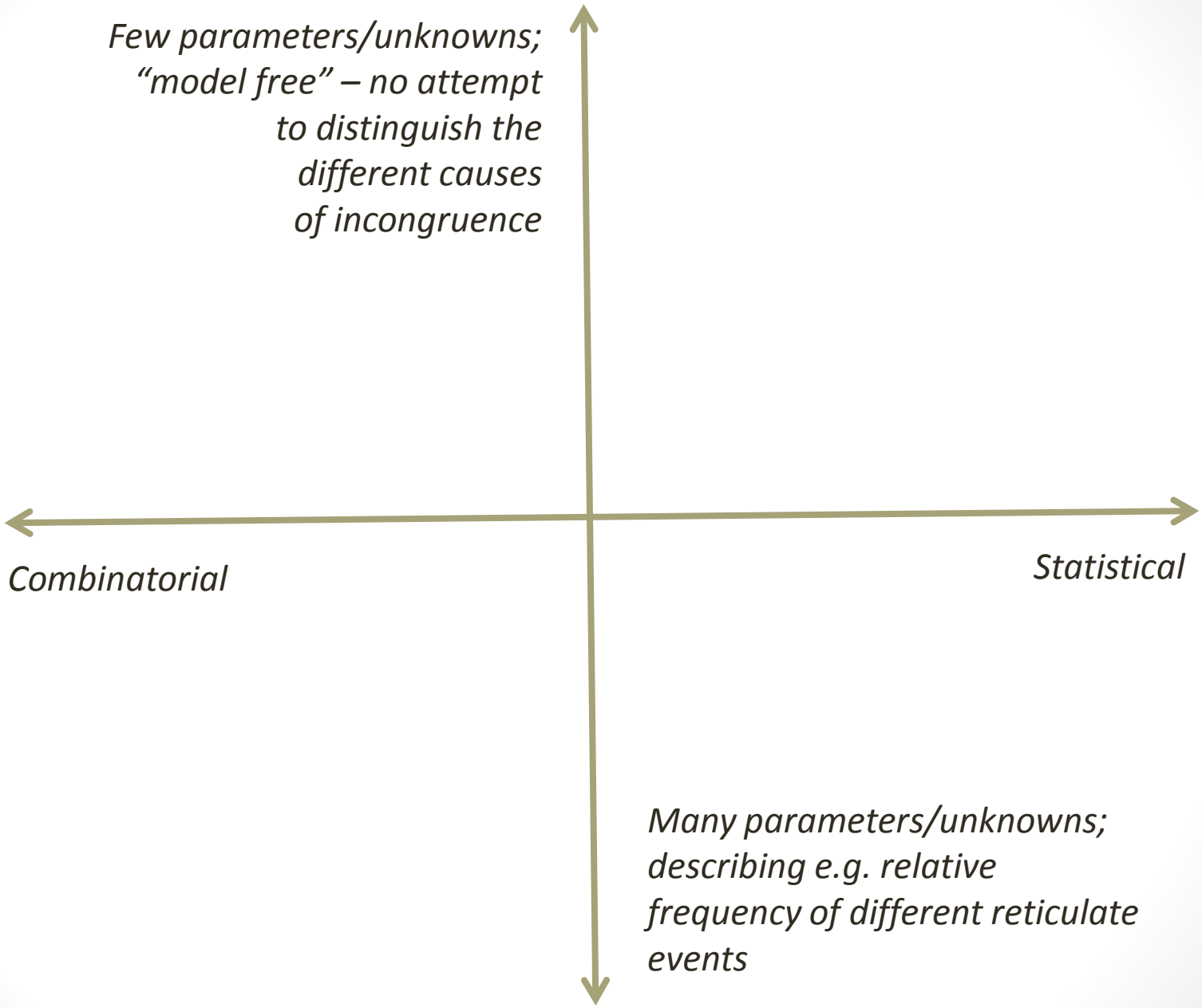


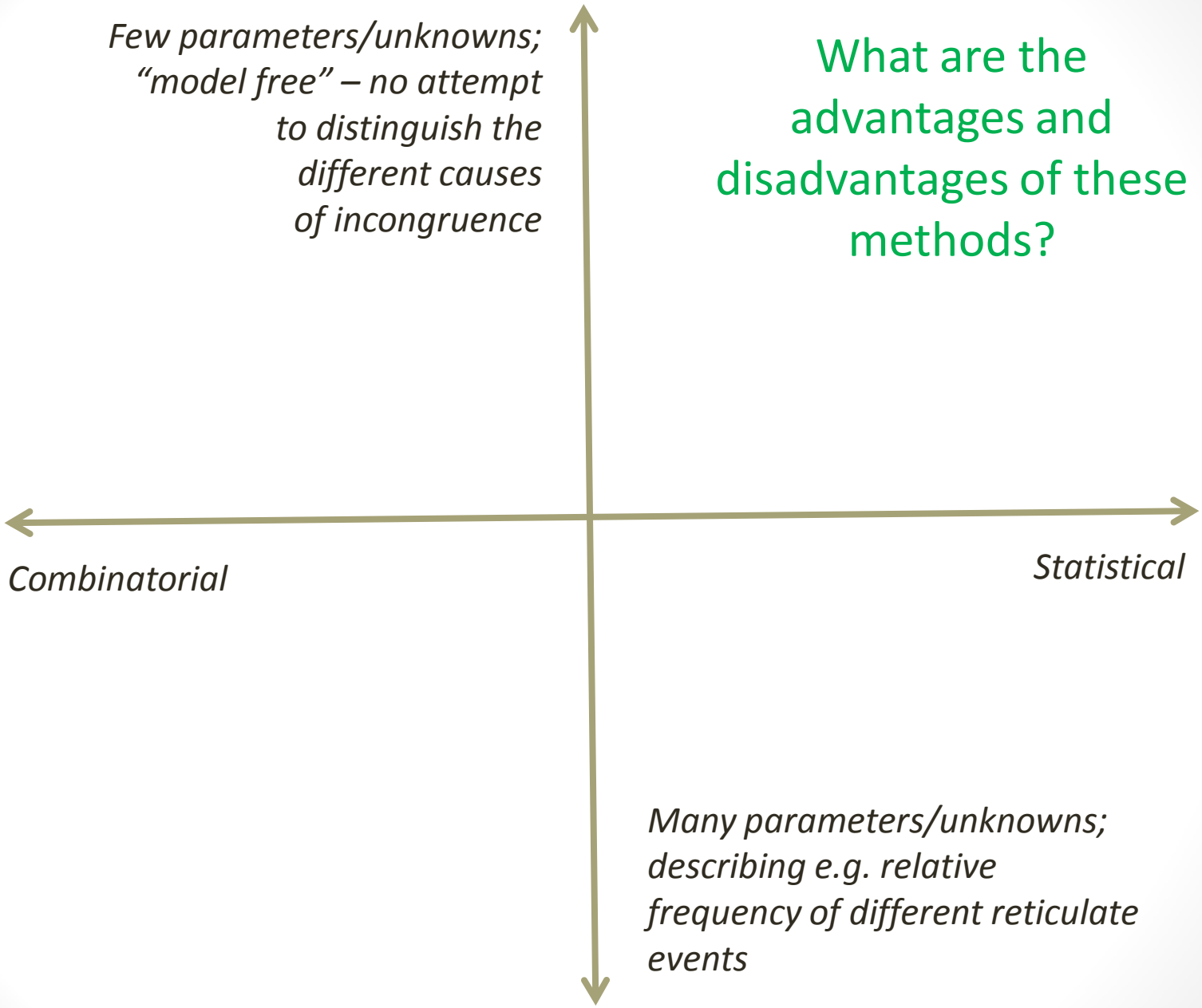
Recombination (population genetics): C is a recombinant of A and B. Linearly ordered character data (e.g. SNPs) is often, but not always, assumed.

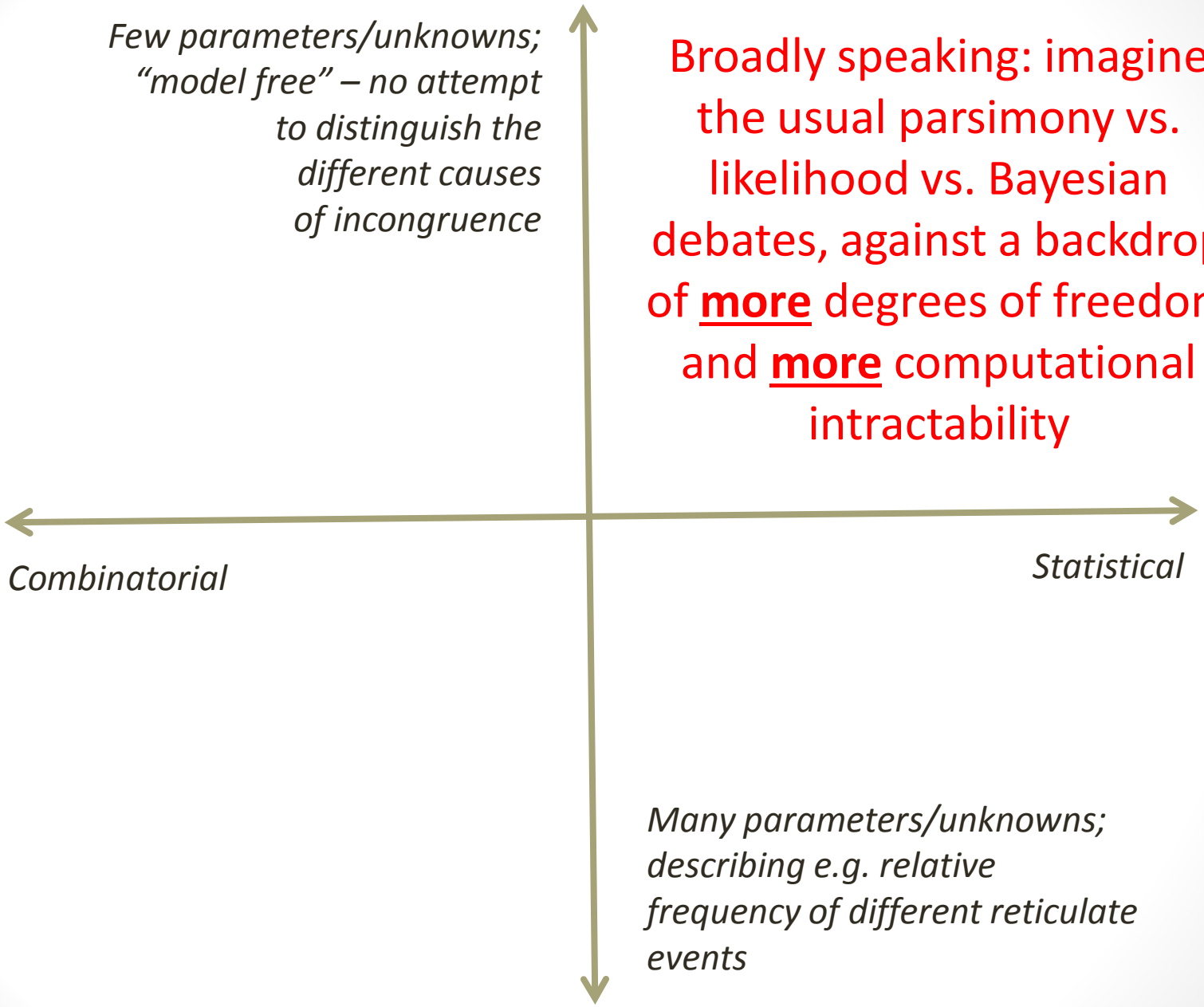
Model assumptions

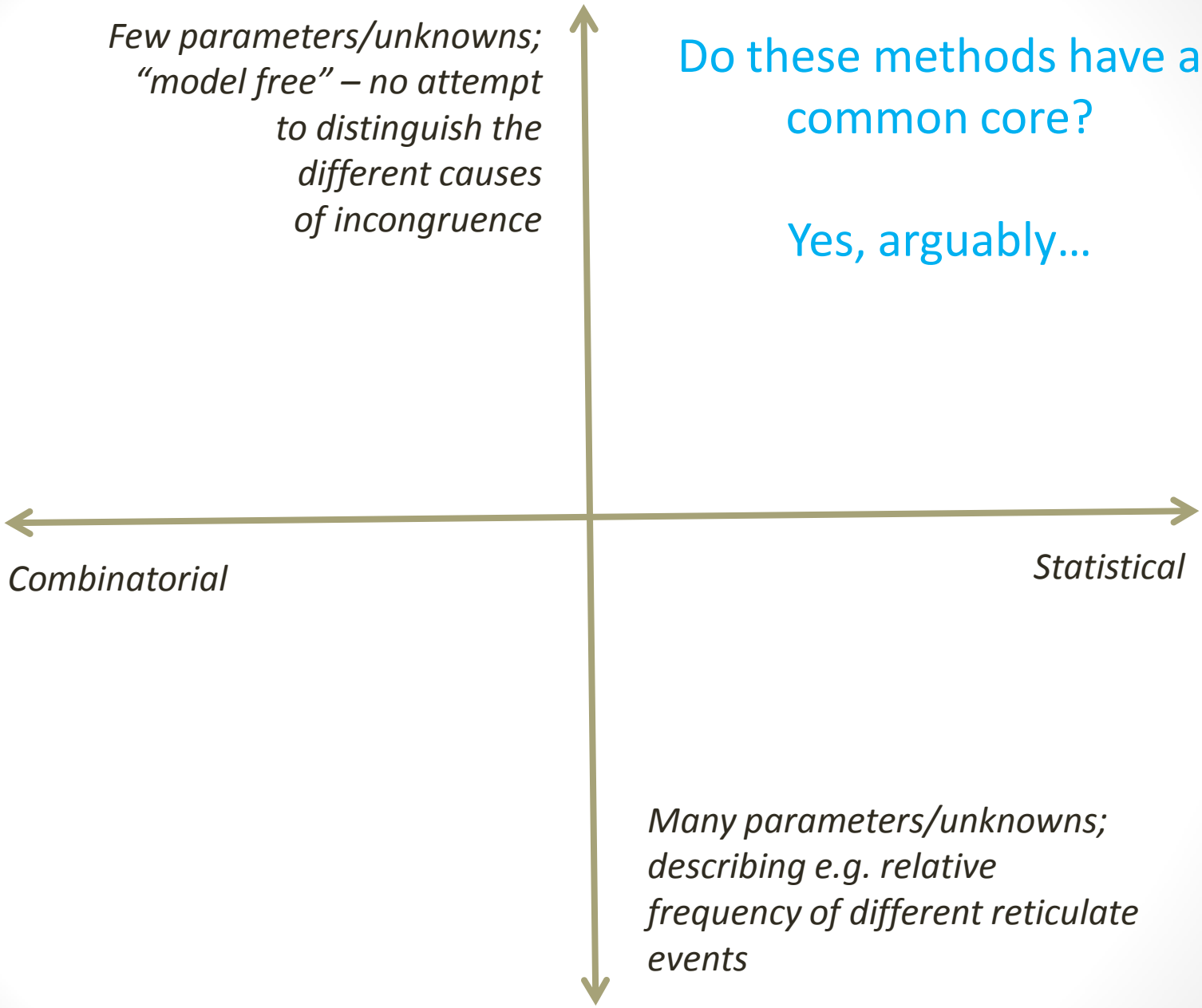
- Some evolutionary phylogenetic network models use assumptions that might not hold in the linguistic context
  - Linearly ordered genome
  - Infinite sites model (each character mutates at most once)

- **But many do not!**
  - **Methods that work directly with sequence data (i.e. multiple alignments) often do not use the linear ordering**
  - **Many methods are (for computational complexity reasons and due to modelling uncertainty) indirect, based on puzzling together (fragments of) tree-like signals that have already been obtained "elsewhere"…**


Few parameters/unknowns;
"model free" – no attempt
to distinguish the
different causes
of incongruence
Combinatorial
Statistical
Many parameters/unknowns;
describing e.g. relative
frequency of different reticulate
events


Few parameters/unknowns; "model free" – no attempt to distinguish the different causes of incongruence
What are the advantages and disadvantages of these methods?
Combinatorial
Statistical
Many parameters/unknowns; describing e.g. relative frequency of different reticulate events

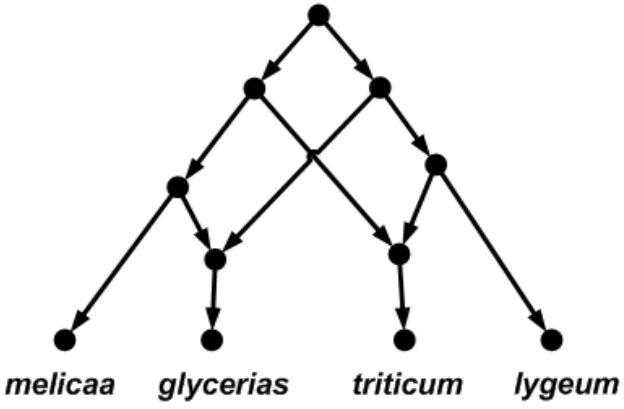

Few parameters/unknowns; "model free" – no attempt to distinguish the different causes of incongruence
Broadly speaking: imagine the usual parsimony vs. likelihood vs. Bayesian debates, against a backdrop of **more** degrees of freedom and **more** computational intractability
Combinatorial
Statistical
Many parameters/unknowns; describing e.g. relative frequency of different reticulate events

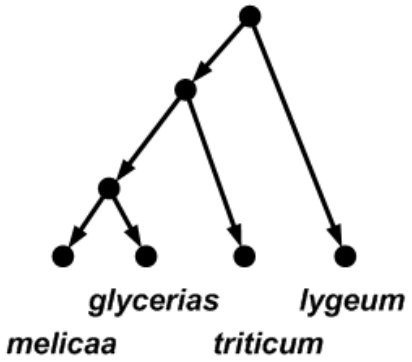

Few parameters/unknowns; "model free" – no attempt to distinguish the different causes of incongruence
Do these methods have a common core?
Yes, arguably…
Combinatorial
Statistical
Many parameters/unknowns; describing e.g. relative frequency of different reticulate events

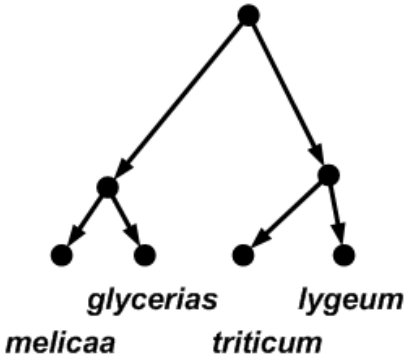
Sets of trees

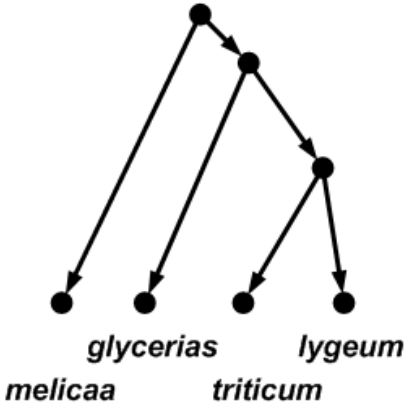
- A recurring theme - sometimes implicit - is the idea that an evolutionary phylogenetic network has many different trees (or more generally: tree-like signals) topologically embedded within it.

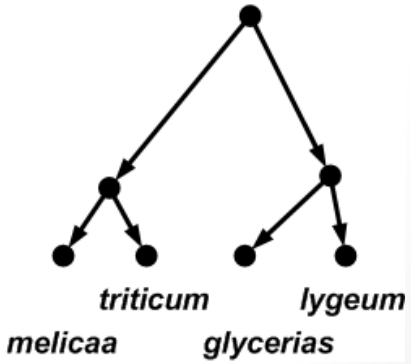
- That is: it is the simultaneous representation of the multiple distinct tree signals that can be present in a genome.

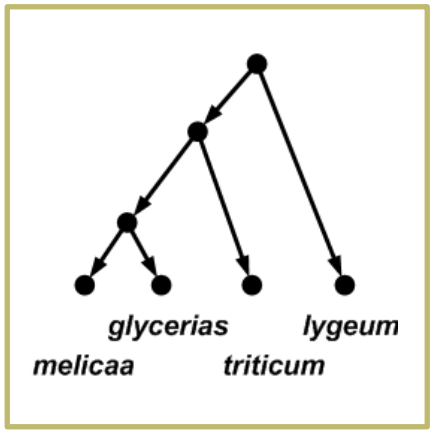

melicaa
glycerias
triticum
lygeum

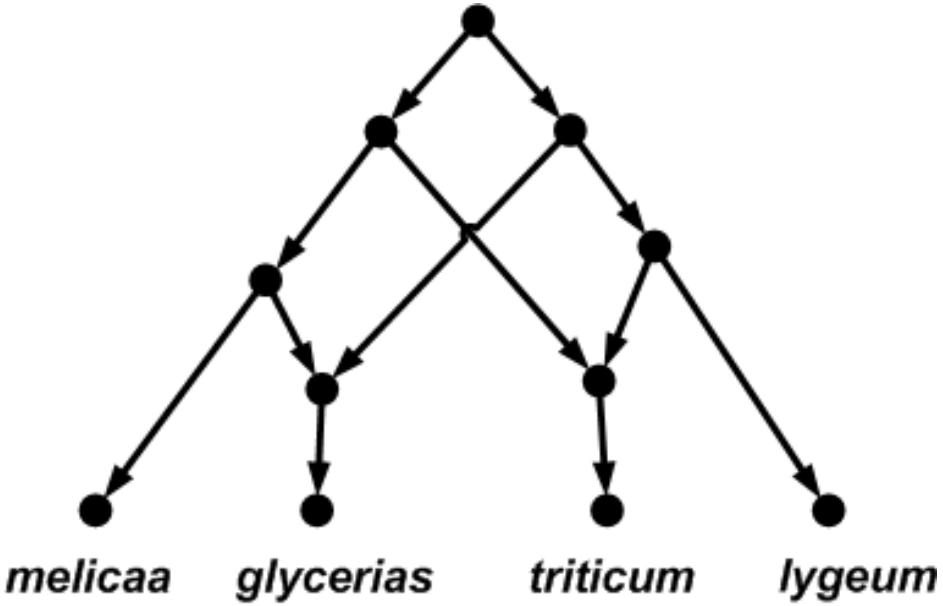

glycerias
lygeum
melicaa
triticum

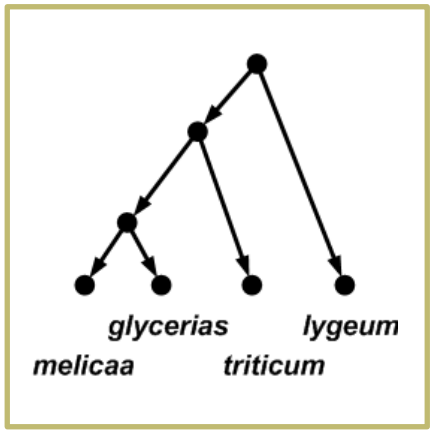

glycerias
lygeum
melicaa
triticum

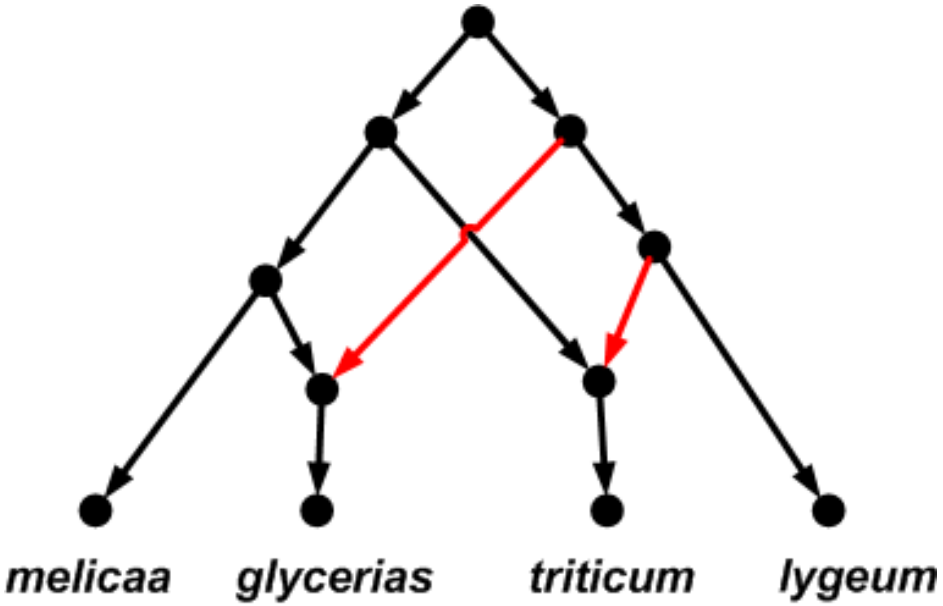

glycerias
lygeum
melicaa
triticum

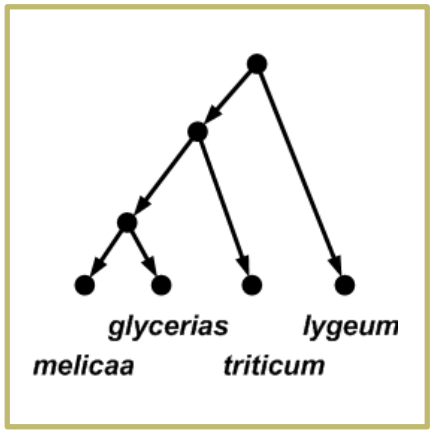

triticum
lygeum
melicaa
glycerias

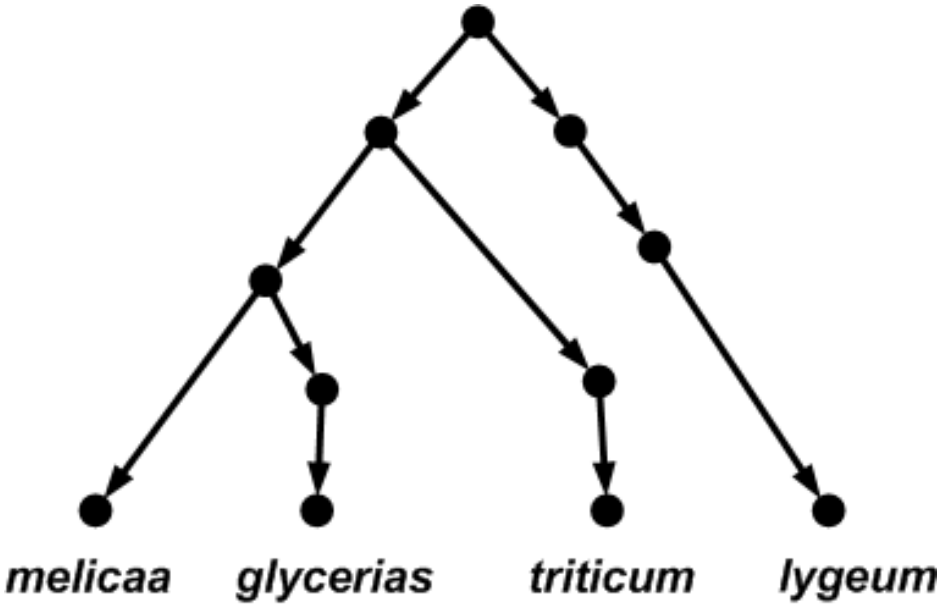

glycerias
lygeum
melicaa
triticum


melicaa
glycerias
triticum
lygeum


glycerias
lygeum
melicaa
triticum


melicaa
glycerias
triticum
lygeum


glycerias
lygeum
melicaa
triticum


melicaa
glycerias
triticum
lygeum

Overall framework

Multiple conflicting tree hypotheses/signals

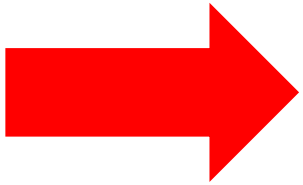
+

Appropriate choice of model (main decision: how to measure the goodness-of-fit of the tree signals to the network?)

+

Appropriate choice of parameters

Sets of (gene) trees
Species tree + sets of gene trees
Subtrees
Monophyletic clades
Characters (e.g. DNA alignment)

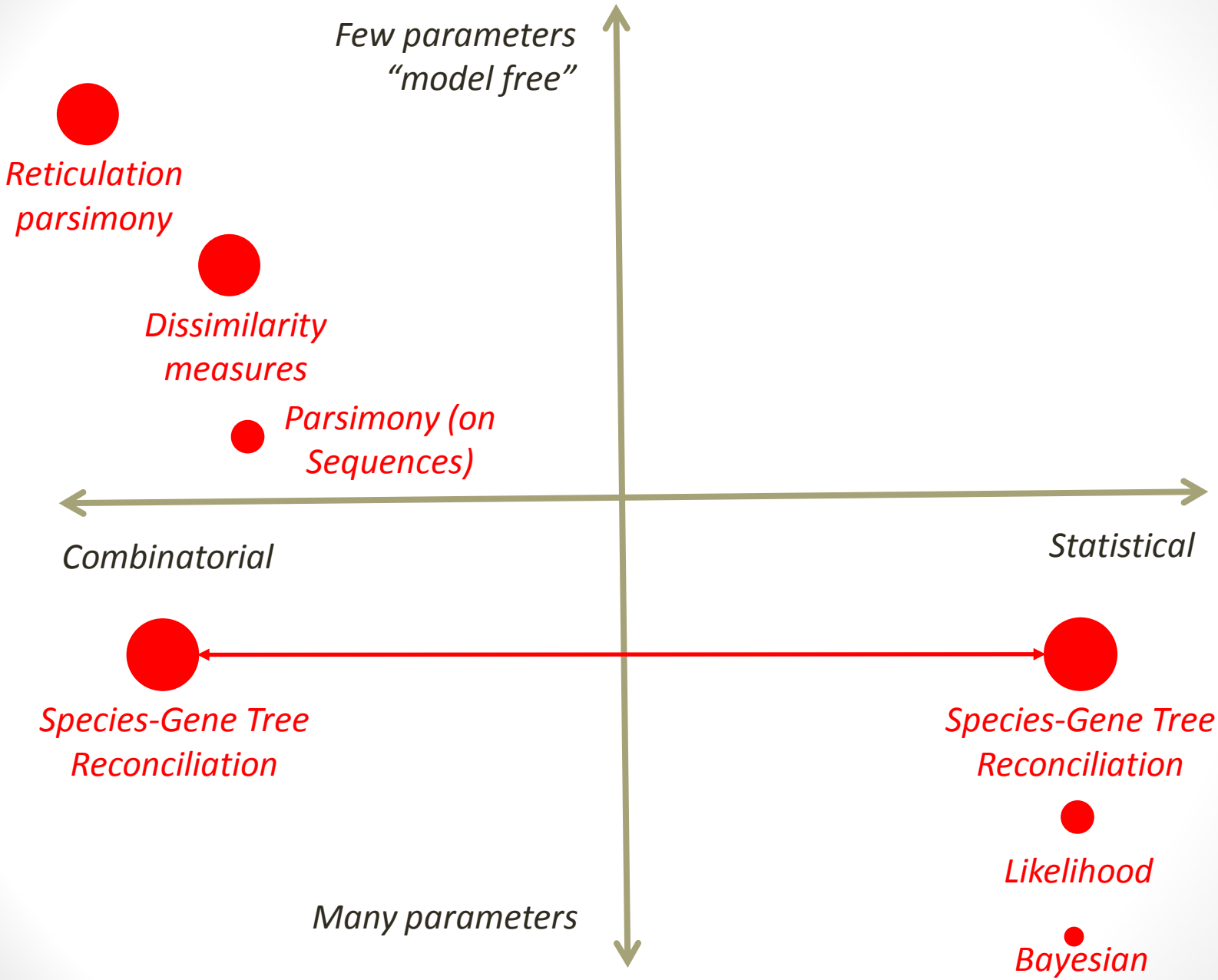


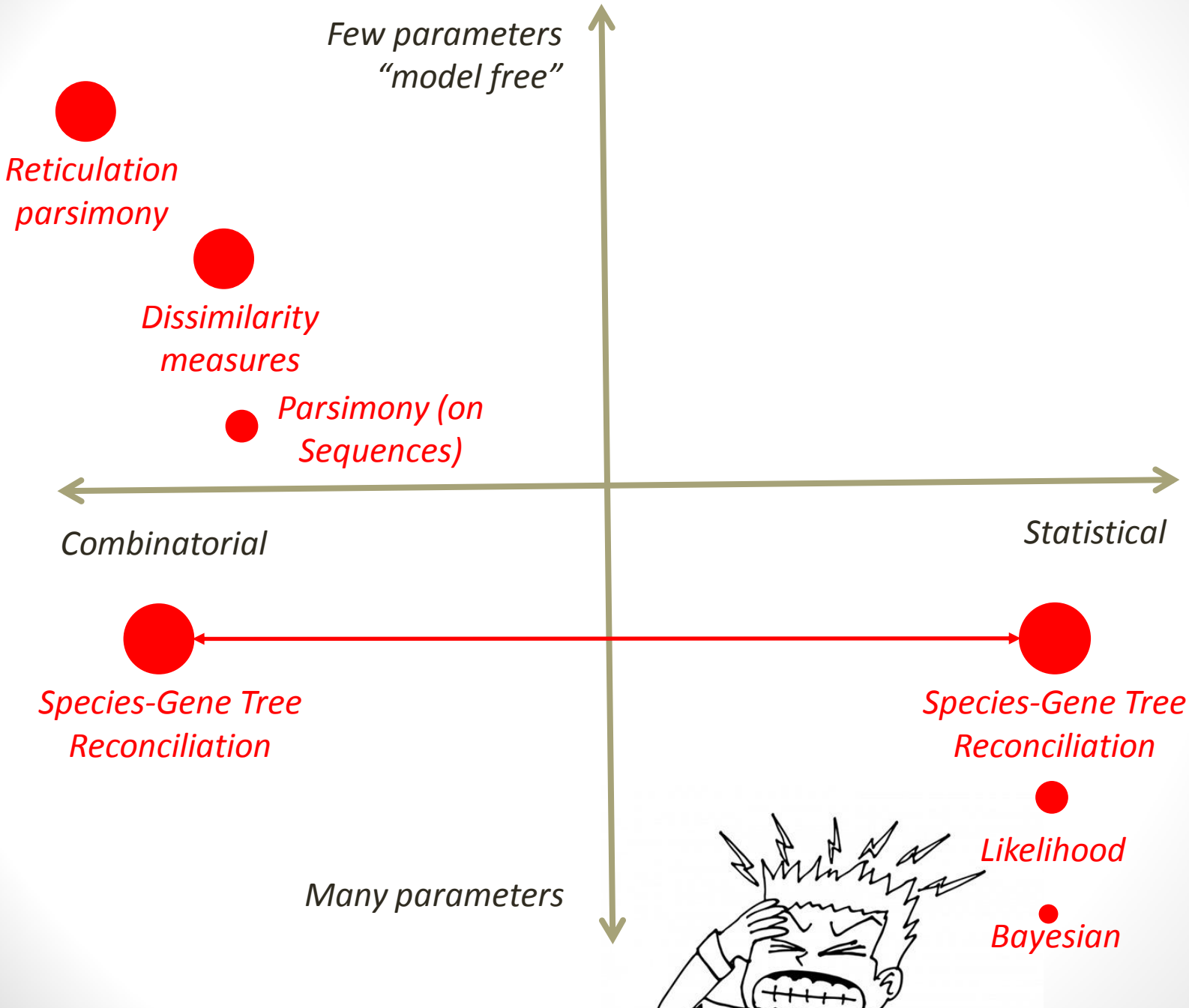
Inference

Phylogenetic network that fits the input tree signals “well”

Summary of methods…

- Reticulation parsimony
  - *Minimizing number of horizontal events*
- Reconciliation
  - *Mapping gene tree(s) onto given species tree subject to minimizing cost model (parsimony) or statistical model*
- Topological dissimilarity
  - *E.g. SPR distance or incompatible quartet topologies as proxy for presence and location of horizontal events*
- “Highways”
  - *Only focus on horizontal events which many gene trees seem to want to use*
- Character-based (e.g. on alignments or SNP data)
  - *Parsimony, Likelihood, Bayesian*


Few parameters
"model free"
Reticulation parsimony
Dissimilarity measures
Parsimony (on Sequences)
Combinatorial
Statistical
Species-Gene Tree Reconciliation
Species-Gene Tree Reconciliation
Likelihood
Many parameters
Bayesian


Few parameters
"model free"
Reticulation parsimony
Dissimilarity measures
Parsimony (on Sequences)
Combinatorial
Statistical
Species-Gene Tree Reconciliation
Species-Gene Tree Reconciliation
Likelihood
Bayesian
Many parameters

Maximum likelihood inference of reticulate evolutionary histories

Yun Yu[a,1], Jianrong Dong[a], Kevin J. Liu[a,b], and Luay Nakhleh[a,b,1]

Departments of [a]Computer Science and [b]Ecology and Evolutionary Biology, Rice University, Houston, TX 77005

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved October 7, 2014 (received for review April 30, 2014)

Hybridization plays an important role in the evolution of certain groups of organisms, adaptation to their environments, and diversification of their genomes. The evolutionary histories of such groups are reticulate, and methods for reconstructing them are still in their infancy and have limited applicability. We present a maximum likelihood method for inferring reticulate evolutionary histories while accounting simultaneously for incomplete lineage sorting. Additionally, we propose methods for assessing confidence in the amount of reticulation and the topology of the inferred evolutionary history. Our method obtains accurate estimates of reticulate evolutionary histories on simulated datasets. Furthermore, our method provides support for a hypothesis of a reticulate evolutionary history inferred from a set of house mouse (*Mus musculus*) genomes. As evidence of hybridization in eukaryotic groups accumulates, it is essential to have methods that infer reticulate evolutionary histories. The work we present here allows for such inference and provides a significant step toward putting phylogenetic networks on par with phylogenetic trees as a model of capturing evolutionary relationships.

reticulate evolution | incomplete lineage sorting | phylogenetic networks |

To the best of our knowledge, the first method to conduct a search of the phylogenetic network space in search of optimal phylogenies is described in a study by our group (18). However, this method is based on the maximum parsimony criterion: It seeks a phylogenetic network that minimizes the number of "extra lineages" resulting from embedding the set of gene tree topologies within its branches.

Progress with phylogenetic network inference notwithstanding, methods of inferring reticulate evolutionary histories while accounting for ILS are still considered to be in their infancy and inapplicable broadly (9). This inapplicability stems mainly from two major issues: the lack of a phylogenetic network inference method and the lack of a method to assess the confidence in the inference. Here, we develop methods that resolve both issues and carry phylogenetic networks into the realm of practical phylogenomic applications. For the inference, we propose operations for traversing the phylogenetic network space, as well as methods for assessing the complexity of a network. For measuring branch support of inferred networks, we use the bootstrap method. Furthermore, we derive, for the first time to our knowledge, the distribution (density) of gene trees with branch lengths given

Likelihood function…

$$L(\Psi, \Gamma | \mathscr{S}) = \prod_{i=1}^{m} \int_{g} \mathbf{P}(S_i | g) p(g | \Psi, \Gamma) dg,$$

- Likelihood of a parameterized network topology (i.e. a network topology augmented with a certain set of branch lengths and inheritance probabilities at the horizontal events) *given* a set of alignments (one per locus/gene) is equal to….

- The product (ranging over each input alignment S_i) of,
  - The integral ranging over all possible gene trees *g, of*
    - The probability of observing S_i given *g,* multiplied by the probability of observing *g* within the network

Likelihood function…

Aaaaaaaaaaaaaaaaaaaargh……

$$L(\Psi, \Gamma | \mathscr{S}) = \prod_{i=1}^{m} \int_{g} \mathbf{P}(S_i | g) p(g | \Psi, \Gamma) dg,$$

- Likelihood of a parameterized network topology (i.e. a network topology augmented with a certain set of branch lengths and inheritance probabilities at the horizontal events) *given* a set of alignments (one per locus/gene) is equal to….

- The product (ranging over each input alignment S_i) of,
  - The integral ranging over all possible gene trees *g, of*
    - The probability of observing S_i given *g,* multiplied by the probability of observing *g* within the network

Likelihood function…second try!

$$L(\Psi, \Gamma | \mathcal{G}) = \prod_{i=1}^{m} p(G_i | \Psi, \Gamma),$$

- Likelihood of a parameterized network topology (i.e. a network topology augmented with a certain set of branch lengths and inheritance probabilities at the horizontal events) *given* **a set of previously inferred gene trees, is…**
- The product (ranging over each gene tree G_i) of,
  - **The probability of observing G_i given the network**
- Topological computational "short-cut"

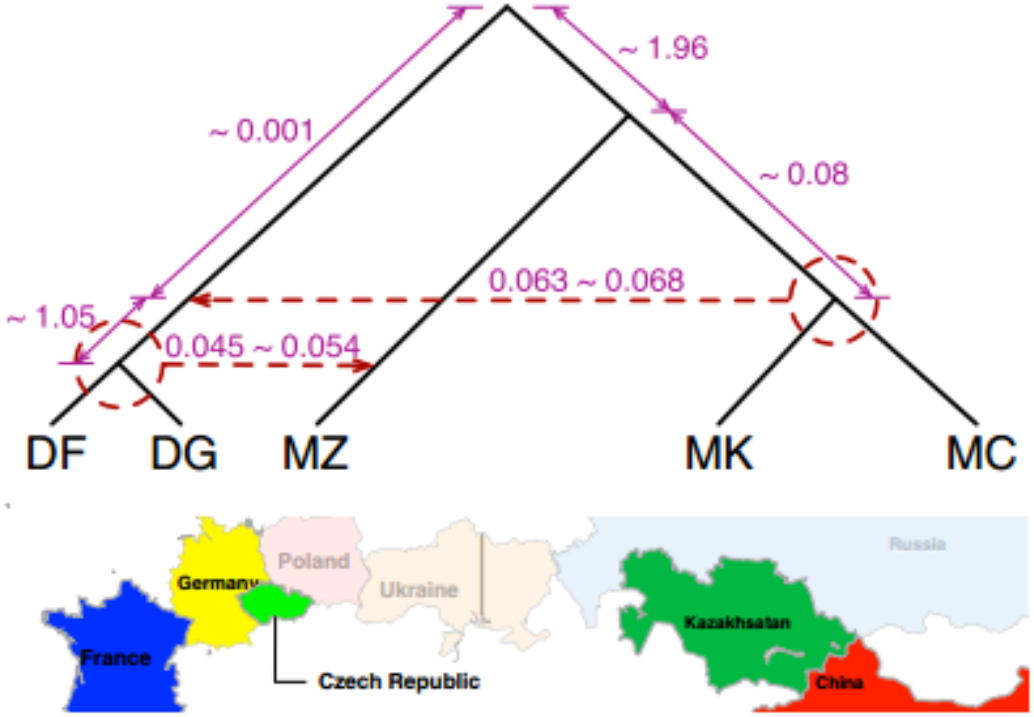


Fig. 3. Optimal phylogenetic network inferred on the house mouse (*M. musculus*) dataset. A single individual was sampled from each of five populations: *M. m. domesticus* from France (DF), *M. m. domesticus* from Germany (DG), *M. m. musculus* from the Czech Republic (MZ), *M. m. musculus* from Kazakhstan (MK), and *M. m. musculus* from China (MC). The analysis found multiple, almost equally optimal, phylogenetic networks with two reticulation events. These multiple networks all agreed on the recipient populations but disagreed on the donor populations. One hybridization (the top dashed horizontal arrow) involves the MRCA of DF and DG as a recipient population, yet seems to have involved MK, MC, or their MRCA as the donor population. The second hybridization (the bottom dashed horizontal arrow) involves MZ as a recipient population, yet seems to have involved DF, DG, or their MRCA as the donor population. Branch lengths in coalescent units (on the tree branches) and inheritance probabilities (on the horizontal edges) are shown (full details of the data and results are provided in *SI Appendix*).

Trends in phylogenetic networks

- Pragmatic combinations of parsimony-based and statistical methods: comparative speed + resolution
- Constructive statistical methods (e.g. A Bayesian Method for Analyzing Lateral Gene Transfer, Sjöstrand et al, Systematic Biology 2014)
- Multi-event models (D-T-L-H-ILS….)
- Robustness/stability analysis (noise, uncertainty, multiple optima)
- Getting the huge size of the network search space under control (…)
- Solving “small” problems (small parsimony, small likelihood) more efficiently (unlike on trees these problems are NP-hard)
- Identifiability / Reconstructability issues

Looking forwards...

Use outside biology?

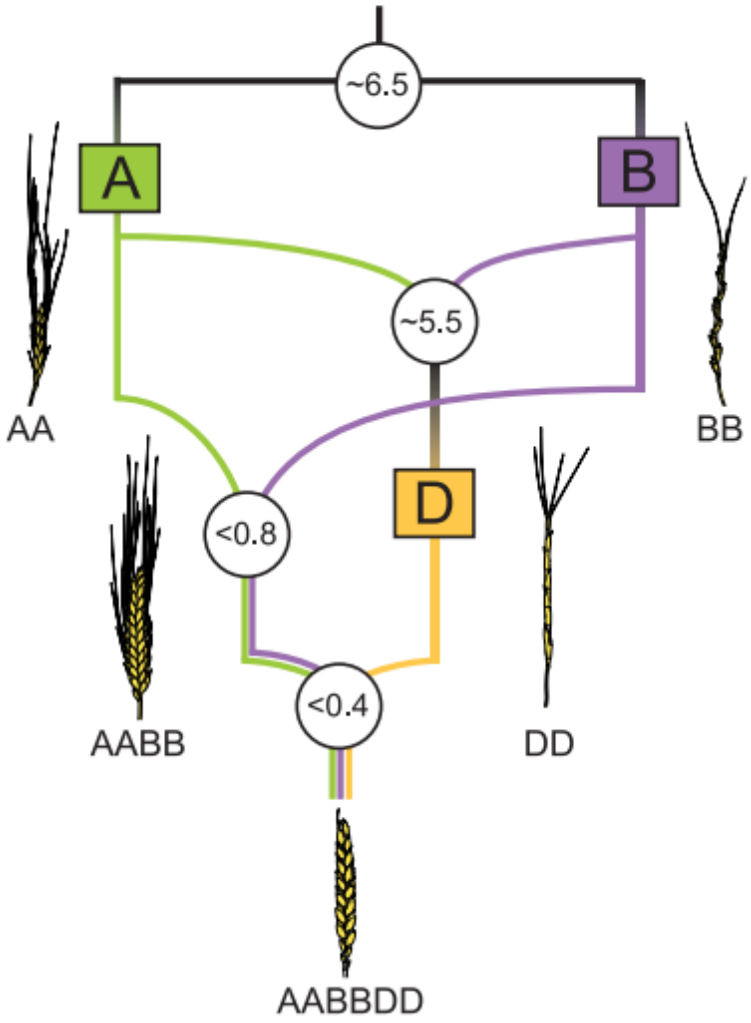
- Computational intractability has forced us to use all kinds of (topological) short-cuts…e.g. working with previously inferred tree topologies (or fragments of trees) instead of directly on sequence data. **Can non-biological fields exploit this abstraction layer?**

- Don't worry if your "gene trees" are bad or incomplete or noisy, they are in biology too! **Many techniques aim at trying to compensate for this (i.e. focussing only on the strongest signal)**

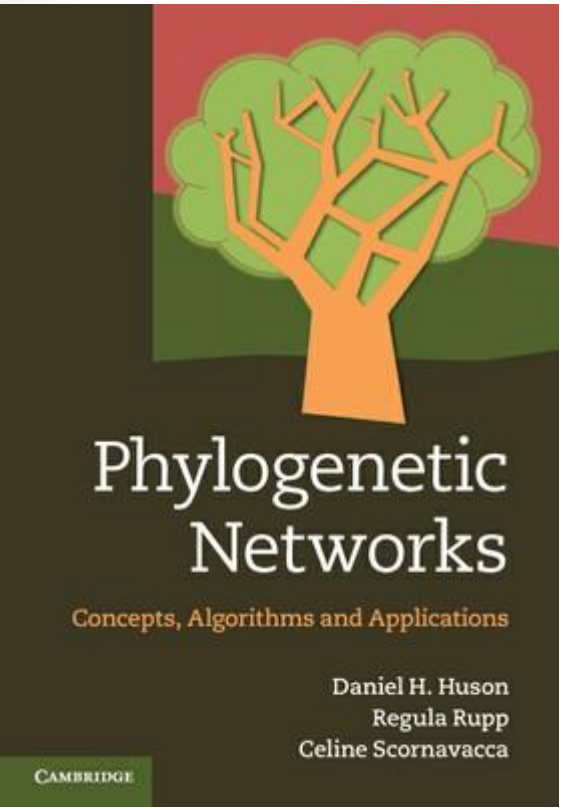
- Automatic network methods are in their infancy in biology too, there is no silver bullet. **Methods will remain semi-automated / part of an ad-hoc experimental pipeline for the foreseeable future. Make sure you understand *exactly* what software does…**

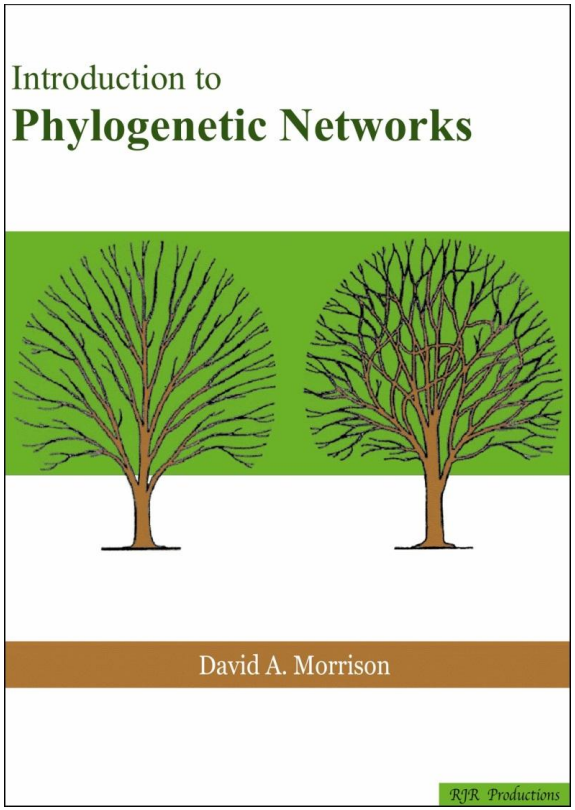
- Don't bother trying to infer network topologies with lots of horizontal events (either in biology or linguistics) – **keep it simple, i.e. at most "a few" horizontal events (cf. mouse, wheat) ➔**

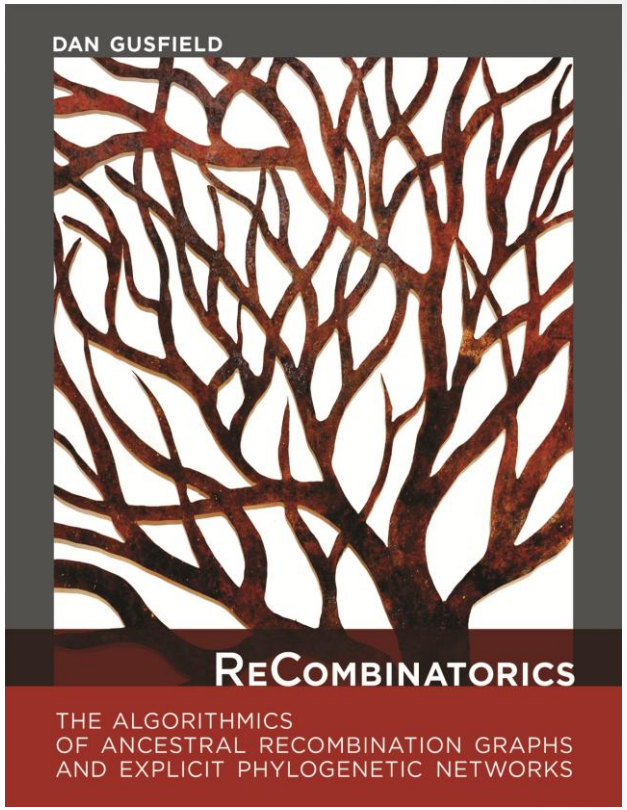
Fig. 3. Model of the phylogenetic history of bread wheat (*Triticum aestivum*; AABBDD). Approximate dates for divergence and the three hybridization events are given in white circles in units of million years ago. Differentiation of the wheat lineage (*Triticum* and *Aegilops*) from a common ancestor into the A and B genome lineages began ~6.5 Ma. The first hybridization occurred ~5.5 Ma between the A and B genome lineages and led to the origin of the D genome lineage by homoploid hybrid speciation. The second hybridization, between a close relative (BB) of *Ae. speltoides* and *T. urartu* (AA), gave rise to the allotetraploid emmer wheat (*T. turgidum*; AABB) by polyploidization. Bread wheat originated by allopolyploidization from a third hybridization, between emmer wheat and *Ae. tauschii* (DD). The three diploid lineages are indicated with color and labels. Inflorescences (spikes) illustrate extant species closely related to those involved in the polyploidizations.



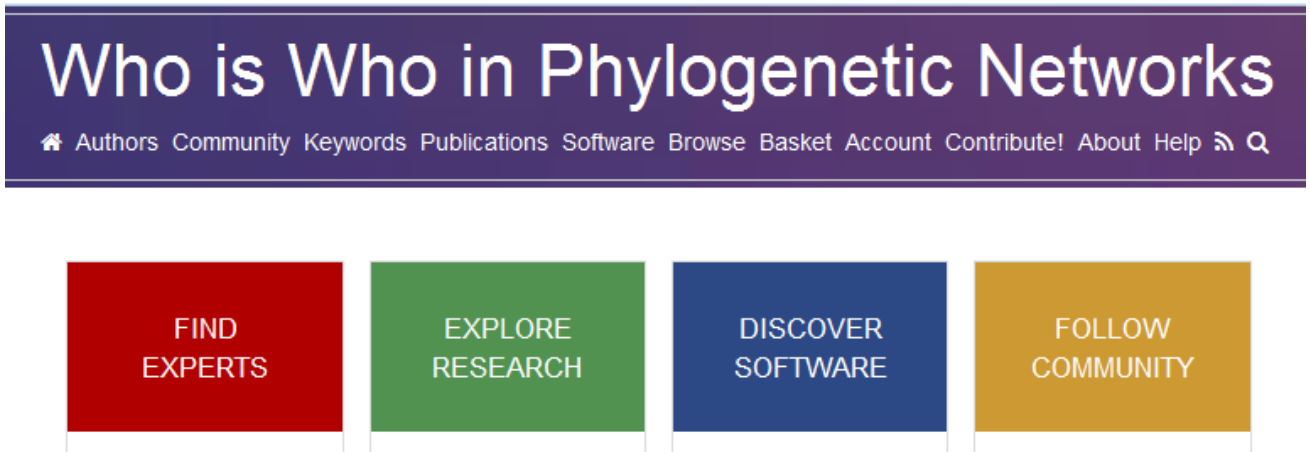
Ancient hybridizations among the ancestral genomes of bread wheat, Marcuse et al, Science 2014


Phylogenetic Networks
Concepts, Algorithms and Applications
Daniel H. Huson
Regula Rupp
Celine Scornavacca
CAMBRIDGE

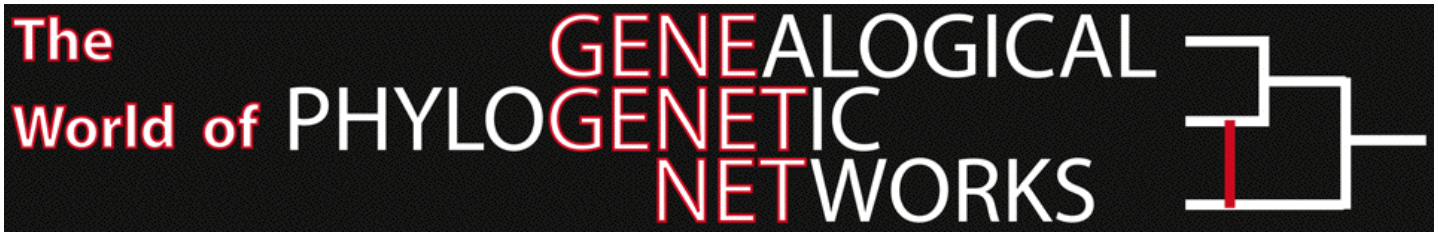

Introduction to
Phylogenetic Networks
David A. Morrison
RJR Productions


DAN GUSFIELD
RECOMBINATORICS
THE ALGORITHMICS
OF ANCESTRAL RECOMBINATION GRAPHS
AND EXPLICIT PHYLOGENETIC NETWORKS

Some books…

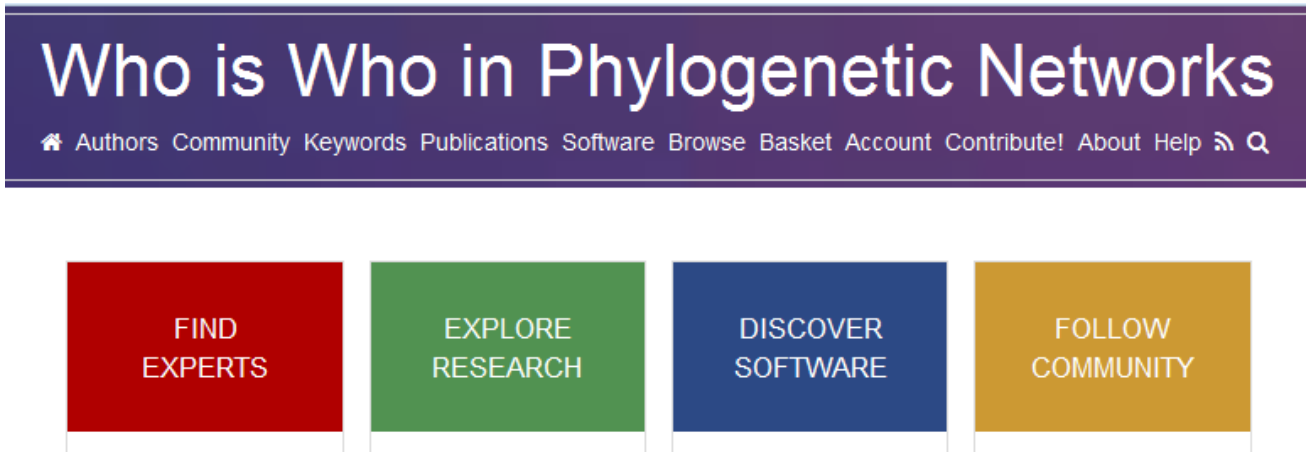


http://phylnet.univ-mlv.fr/

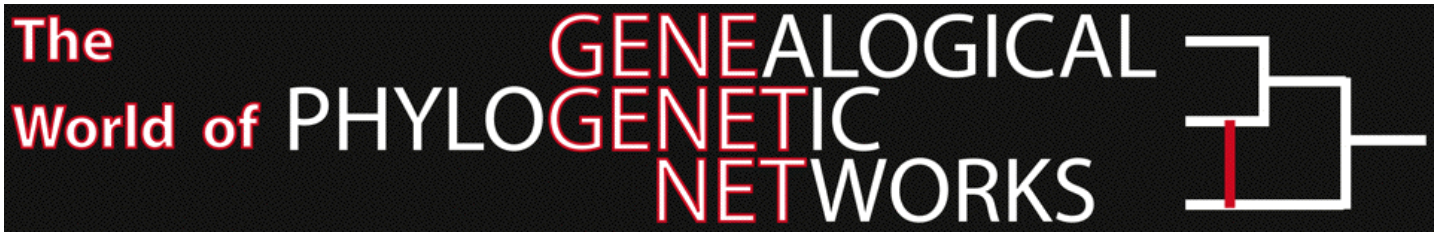


http://phylonetworks.blogspot.com

Some websites…



http://phylnet.univ-mlv.fr/



http://phylonetworks.blogspot.com

Thank you for listening ☺

Park *et al. BMC Evolutionary Biology* 2010, **10**:131
http://www.biomedcentral.com/1471-2148/10/131

BMC Evolutionary Biology

METHODOLOGY ARTICLE Open Access

Bootstrap-based Support of HGT Inferred by Maximum Parsimony

Hyun Jung Park†, Guohua Jin† and Luay Nakhleh*†

Abstract

Background: Maximum parsimony is one of the most commonly used criteria for reconstructing phylogenetic trees. Recently, Nakhleh and co-workers extended this criterion to enable reconstruction of *phylogenetic networks*, and demonstrated its application to detecting reticulate evolutionary relationships. However, one of the major problems with this extension has been that it favors more complex evolutionary relationships over simpler ones, thus having the potential for overestimating the amount of reticulation in the data. An *ad hoc* solution to this problem that has been used entails inspecting the improvement in the parsimony length as more reticulation events are added to the model, and stopping when the improvement is below a certain threshold.

Results: In this paper, we address this problem in a more systematic way, by proposing a nonparametric bootstrap-based measure of support of inferred reticulation events, and using it to determine the number of those events, as well as their placements. A number of samples is generated from the given sequence alignment, and reticulation events are inferred based on each sample. Finally, the support of each reticulation event is quantified based on the inferences made over all samples.

Conclusions: We have implemented our method in the NEPAL software tool (available publicly at http://bioinfo.cs.rice.edu/), and studied its performance on both biological and simulated data sets. While our studies show very promising results, they also highlight issues that are inherently challenging when applying the maximum parsimony criterion to detect reticulate evolution.

BIOINFORMATICS ORIGINAL PAPER

2013, pages 1–9
doi:10.1093/bioinformatics/btt021

Phylogenetics

Advance Access publication January 17, 2013

Systematic inference of highways of horizontal gene transfer in prokaryotes

Mukul S. Bansal[1,†], Guy Banay[1], Timothy J. Harlow[2], J. Peter Gogarten[2] and Ron Shamir[1,*]

[1]The Blavatnik School of Computer Science, Tel-Aviv University, Ramat Aviv, Tel Aviv 69978, Israel and [2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA

Associate Editor: David Posada

Syst. Biol. 59(1):27–41, 2010
© The Author(s) 2009. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org
DOI:10.1093/sysbio/syp076
Advance Access publication on November 9, 2009

Unifying Vertical and Nonvertical Evolution: A Stochastic ARG-based Framework

ERIK W. BLOOMQUIST[1] AND MARC A. SUCHARD[1,2,3,*]

[1]*Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095, USA; and*
[2]*Department of Biomathematics and* [3]*Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095-1766, USA;*
**Correspondence to be sent to: Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095-1766, USA;*
E-mail: msuchard@ucla.edu.

Received 12 January 2009; reviews returned 8 April 2009; accepted 21 September 2009
Associate Editor: Laura Kubatko

van Iersel *et al. BMC Bioinformatics* 2014, **15**:127
http://www.biomedcentral.com/1471-2105/15/127

BMC Bioinformatics

RESEARCH ARTICLE Open Access

A practical approximation algorithm for solving massive instances of hybridization number for binary and nonbinary trees

Leo van Iersel[1*], Steven Kelk[2], Nela Lekić[2] and Celine Scornavacca[3]

Abstract

Background: Reticulate events play an important role in determining evolutionary relationships. The problem of computing the minimum number of such events to explain discordance between two phylogenetic trees is a hard computational problem. Even for binary trees, exact solvers struggle to solve instances with reticulation number larger than 40-50.

Results: Here we present CycleKiller and NonbinaryCycleKiller, the first methods to produce solutions verifiably close to optimality for instances with hundreds or even thousands of reticulations.

Conclusions: Using simulations, we demonstrate that these algorithms run quickly for large and difficult instances, producing solutions that are very close to optimality. As a spin-off from our simulations we also present TerminusEst, which is the fastest exact method currently available that can handle nonbinary trees: this is used to measure the accuracy of the NonbinaryCycleKiller algorithm. All three methods are based on extensions of previous theoretical work (SIDMA 26(4):1635-1656, TCBB 10(1):18-25, SIDMA 28(1):49-66) and are publicly available. We also apply our methods to real data.

Keywords: Hybridization number, Phylogenetic networks, Approximation algorithms, Directed feedback vertex set

OPEN ACCESS Freely available online

PLOS | ONE

Genealogy-Based Methods for Inference of Historical Recombination and Gene Flow and Their Application in *Saccharomyces cerevisiae*

Paul A. Jenkins[1], Yun S. Song[1,2], Rachel B. Brem[3]*

Bioinformatics, 31(6), 2015, 841–848
doi: 10.1093/bioinformatics/btu728
Advance Access Publication Date: 6 November 2014
Original Paper

OXFORD

Phylogenetics

Joint amalgamation of most parsimonious reconciled gene trees

Celine Scornavacca[1,2,*], Edwin Jacox[1] and Gergely J. Szöllősi[3,*]

[1]ISEM, UM2-CNRS-IRD, Place Eugène Bataillon 34095 Montpellier, France, [2]Institut de Biologie Computationnelle (IBC), 95 rue de la Galéra, 34095 Montpellier, France and [3]ELTE-MTA 'Lendület' Biophysics Research Group 1117 Bp., Pázmány P. stny. 1A., Budapest, Hungary

*To whom correspondence should be addressed.
Associate Editor: David Posada

Received on May 27, 2014; revised on October 28, 2014; accepted on October 29, 2014

Abstract

Motivation: Traditionally, gene phylogenies have been reconstructed solely on the basis of molecular sequences; this, however, often does not provide enough information to distinguish between statistically equivalent relationships. To address this problem, several recent methods have incorporated information on the species phylogeny in gene tree reconstruction, leading to dramatic improvements in accuracy. Although probabilistic methods are able to estimate all model parameters but are computationally expensive, parsimony methods—generally computationally more efficient—require a prior estimate of parameters and of the statistical support.

Results: Here, we present the Tree Estimation using Reconciliation (TERA) algorithm, a parsimony based, species tree aware method for gene tree reconstruction based on a scoring scheme combining duplication, transfer and loss costs with an estimate of the sequence likelihood. TERA explores all reconciled gene trees that can be amalgamated from a sample of gene trees. Using a large scale simulated dataset, we demonstrate that TERA achieves the same accuracy as the corresponding probabilistic method while being faster, and outperforms other parsimony-based methods in both accuracy and speed. Running TERA on a set of 1099 homologous gene families from complete cyanobacterial genomes, we find that incorporating knowledge of the species tree results in a two thirds reduction in the number of apparent transfer events.

Availability and implementation: The algorithm is implemented in our program *TERA*, which is freely available from http://mbb.univ-montp2.fr/MBB/download_sources/16__TERA.

Contact: celine.scornavacca@univ-montp2.fr, ssolo@angel.elte.hu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

BIOINFORMATICS

Vol. 28 ECCB 2012, pages i409–i415
doi:10.1093/bioinformatics/bts386

Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees

Maureen Stolzer[1,*], Han Lai[1], Minli Xu[2], Deepa Sathaye[3], Benjamin Vernot[4] and Dannie Durand[1,3]

[1]Department of Biological Sciences, [2]Lane Center for Computational Biology, [3]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA and [4]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations

Gergely J. Szöllősi[a,b], Bastien Boussau[a,b,c], Sophie S. Abby[d,e], Eric Tannier[a,b,f], and Vincent Daubin[a,b,1]

[a]Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France; [b]Université de Lyon, F-69000 Lyon, France; [c]Department of Integrative Biology, University of California, Berkeley, CA 94720-3140; [d]Microbial Evolutionary Genomics, Département Génomes et Génétique, Institut Pasteur, F-75015 Paris, France; [e]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 3525, F-75015 Paris, France; and [f]Institut National de Recherche en Informatique et en Automatique Rhône-Alpes, F-38334 Montbonnot, France

Edited by Marc A. Suchard, University of California, Los Angeles, CA, and accepted by the Editorial Board September 7, 2012 (received for review February 20, 2012)

Syst. Biol. 64(1):102–111, 2015
© The Author(s) 2014. Published by Oxford University Press, on behalf of the Society of Systematic Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syu076
Advance Access publication September 18, 2014

How Much Information is Needed to Infer Reticulate Evolutionary Histories?

KATHARINA T. HUBER[1], LEO VAN IERSEL[2], VINCENT MOULTON[1,*], AND TAOYANG WU[1]

[1]*School of Computing Sciences, University of East Anglia, Norwich, UK, and* [2]*Centrum Wiskunde & Informatica (CWI), Amsterdam, Netherlands*
**Correspondence to be sent to: School of Computing Sciences, University of East Anglia, Norwich, UK; E-mail: vincent.moulton@cmp.uea.ac.uk.*

Received 14 February 2014; reviews returned 31 August 2014; accepted 10 September 2014
Associate Editor: Tania Stadler

Abstract.—Phylogenetic networks are a generalization of evolutionary trees and are an important tool for analyzing reticulate evolutionary histories. Recently, there has been great interest in developing new methods to construct rooted phylogenetic networks, that is, networks whose internal vertices correspond to hypothetical ancestors, whose leaves correspond to sampled taxa, and in which vertices with more than one parent correspond to taxa formed by reticulate evolutionary events such as recombination or hybridization. Several methods for constructing evolutionary trees use the strategy of building up a tree from simpler building blocks (such as triplets or clusters), and so it is natural to look for ways to construct networks from smaller networks. In this article, we shall demonstrate a fundamental issue with this approach. Namely, we show that even if we are given *all* of the subnetworks induced on all proper subsets of the leaves of some rooted phylogenetic network, we still do not have all of the information required to completely determine that network. This implies that even if *all* of the building blocks for some reticulate evolutionary history were to be taken as the input for any given network building method, the method might still output an incorrect history. We also discuss some potential consequences of this result for constructing phylogenetic networks. [Evolutionary tree; network reconstruction; phylogenetic network; reticulate evolution.]

Syst. Biol. 0(0):1–10, 2015
© The Author(s) 2015. Published by Oxford University Press on behalf of the Society of Systematic Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syv037

Which Phylogenetic Networks are Merely Trees with Additional Arcs?

ANDREW R. FRANCIS[1] AND MIKE STEEL[2,*]

[1]*Centre for Research in Mathematics, School of Computing, Engineering and Mathematics, University of Western Sydney, Australia;*
[2]*Biomathematics Research Centre, University of Canterbury, New Zealand*
**Correspondence to be sent to: Biomathematics Research Centre, University of Canterbury, Christchurch, 8041, Christchurch; E-mail: mike.steel@canterbury.ac.nz*

Received 13 April 2015; reviews returned 13 May 2015; accepted 20 May 2015
Associate Editor: Mark Holder

Abstract.—A binary phylogenetic network may or may not be obtainable from a tree by the addition of directed edges (arcs) between tree arcs. Here, we establish a precise and easily tested criterion (based on "2-SAT") that efficiently determines whether or not any given network can be realized in this way. Moreover, the proof provides a polynomial-time algorithm for finding one or more trees (when they exist) on which the network can be based. A number of interesting consequences are presented as corollaries; these lead to some further relevant questions and observations, which we outline in the conclusion. [Algorithm, Antichain, Phylogenetic network, phylogenetic tree, reticulate evolution, 2-SAT.]

Syst. Biol. 0(0):1–12, 2014
© The Author(s) 2014. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syu007

A Bayesian Method for Analyzing Lateral Gene Transfer

JOEL SJÖSTRAND[1,2], ALI TOFIGH[3], VINCENT DAUBIN[4], LARS ARVESTAD[1,2,5], BENGT SENNBLAD[1,6], and JENS LAGERGREN[1,7,*]

[1]*Science for Life Laboratory, Tomtebodavägen 23A, 17165 Solna, Sweden,* [2]*Department of Numerical Analysis and Computer Science, Stockholm University, Sweden,* [3]*McGill Centre for Bioinformatics, 4th floor, Bellini Building, Life Sciences Complex, 3649 Promenade Sir William Osler, Montreal, Quebec, Canada, H3G 0B1,* [4]*UMR CNRS 5558 - LBBE, "Biométrie et Biologie évolutive", UCB Lyon 1 - Bât. Grégor Mendel, 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex,* [5]*Department of Mechanics, Osquars Backe 18, KTH, SE-100 44 Stockholm, Sweden,* [6]*Karolinska University Hospital, CMM L8:03, Solna, SE-171 76 Stockholm, Sweden and* [7]*The School of Computer Science and Communication, Lindstedtsvägen 3, 5, KTH CSC, SE-100 44 Stockholm, Sweden*

**Correspondence to be sent to: Science for Life Laboratory, KISP (Karolinska Institutet Science Park), Tomtebodavägen 23A, 17165 Solna, Sweden, E-mail: jens.lagergren@scilifelab.se*

Joel Sjöstrand and Ali Tofigh contributed equally to this article.

Received 1 March 2013; reviews returned 24 May 2013; accepted 10 February 2014
Associate Editor: Peter Foster

Abstract.—Lateral gene transfer (LGT)—which transfers DNA between two non-vertically related individuals belonging to the same or different species—is recognized as a major force in prokaryotic evolution, and evidence of its impact on eukaryotic evolution is ever increasing. LGT has attracted much public attention for its potential to transfer pathogenic elements and antibiotic resistance in bacteria, and to transfer pesticide resistance from genetically modified crops to other plants. In a wider perspective, there is a growing body of studies highlighting the role of LGT in enabling organisms to occupy new niches or adapt to environmental changes. The challenge LGT poses to the standard tree-based conception of evolution is also being debated. Studies of LGT have, however, been severely limited by a lack of computational tools. The best currently available LGT algorithms are parsimony-based phylogenetic methods, which require a pre-computed gene tree and cannot choose between sometimes wildly differing most parsimonious solutions. Moreover, in many studies, simple heuristics are applied that can only handle putative orthologs and completely disregard gene duplications (GDs). Consequently, proposed LGT among specific gene families, and the rate of LGT in general, remain debated. We present a Bayesian Markov-chain Monte Carlo-based method that integrates GD, gene loss, LGT, and sequence evolution, and apply the method in a genome-wide analysis of two groups of bacteria: Mollicutes and Cyanobacteria. Our analyses show that although the LGT rate between distant species is high, the net combined rate of duplication and close-species LGT is on average higher. We also show that the common practice of disregarding reconcilability in gene tree inference overestimates the number of LGT and duplication events. [Bayesian; gene duplication; gene loss; horizontal gene transfer; lateral gene transfer; MCMC; phylogenetics.]