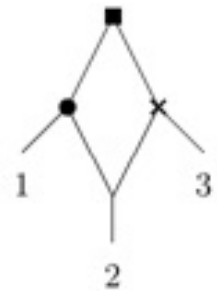
Orthology relations: From trees to networks

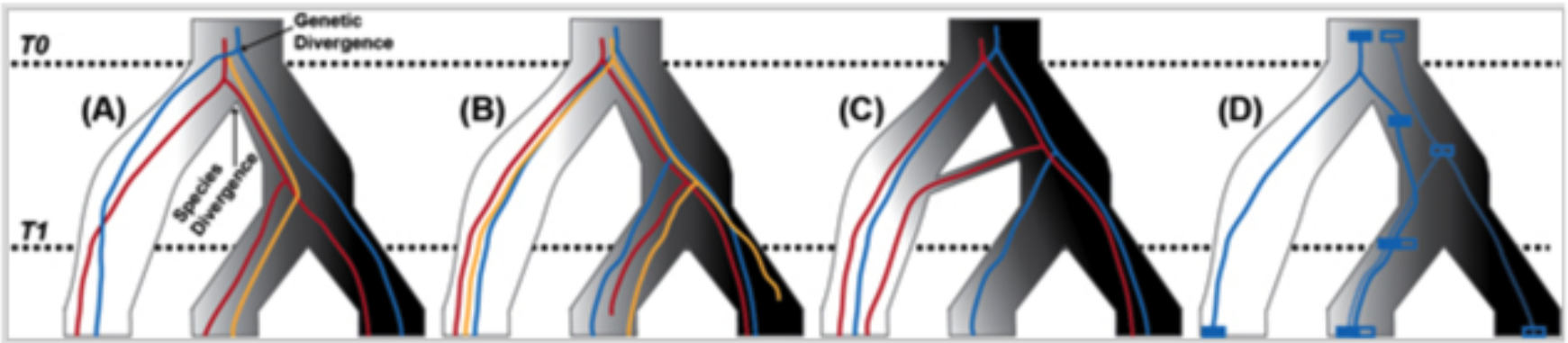
Katharina Huber

School of Computing Sciences,
University of East Anglia, UK.



February 15, 2017

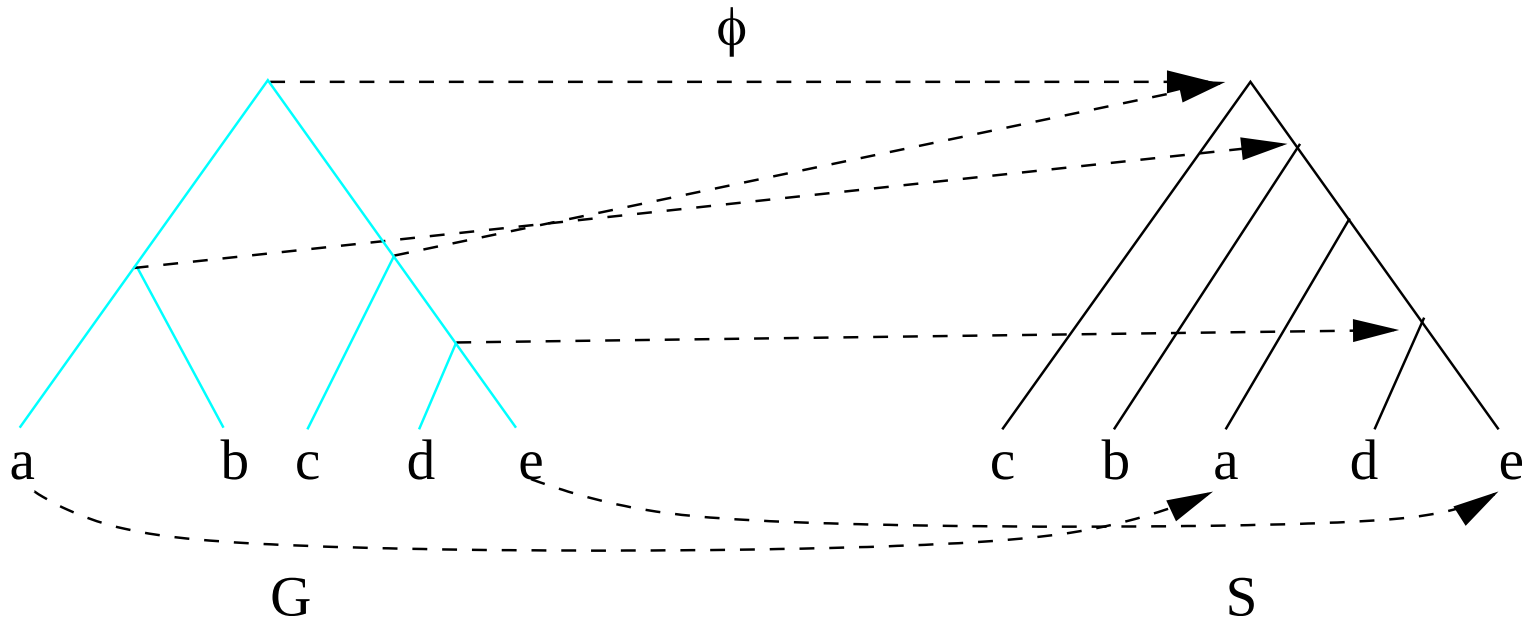
Gene trees vs species trees



Gene trees and species trees can be incongruent for many reasons.
Gene trees and species trees can be incongruent for many reasons. (A) Genes can have unequal rates of evolution. (B) Gene loss and gene duplication are common. (C) Gene flow can occur between lineages after their separation. (D) Recombination between neighboring regions can also lead to species phylogenies and gene histories that do not match.

M W Mitchell et al, Nature Education Knowledge 4(2):1

Standard approach



Reconciliation map ϕ for a gene tree G and a species tree S:

- $\phi : V(G) \rightarrow V(S)$
- Duplication: there exists at least one child v' of v such that $\phi(v)$ is above $\phi(v')$

Optimization problem

Task: Assume a list of allowed evolutionary processes and a cost for each of them and find a reconciliation mapping that is optimal (among potentially many optimal ones!).

- parsimony framework: Doyon et al (2009), Bansal et al (2012), Tofigh et al (2011), Jacox et al (2016) etc.
- Bayesian framework: Baudet et al (2015) etc.
- likelihood framework: Goreki et al (2011) etc.

Drawback: They all depend on the quality of the trees which is not guaranteed.

Alternative approach

Phylogenomics with paralogs

Marc Hellmuth[a,1,2], Nicolas Wieseke[b,1], Marcus Lechner[c], Hans-Peter Lenhof[a], Martin Middendorf[b], and Peter F. Stadler[d,e,f,g,h,i,j]

[a]Center for Bioinformatics, Saarland University, D-66041 Saarbrücken, Germany; [b]Parallel Computing and Complex Systems Group, Department of Computer Science, Leipzig University, D-04109 Leipzig, Germany; [c]Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, D-35032 Marburg, Germany; [d]Bioinformatics Group, Department of Computer Science, and [e]Interdisciplinary Center of Bioinformatics, Leipzig University, D-04107 Leipzig, Germany; [f]Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany; [g]Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany; [h]Institute for Theoretical Chemistry, University of Vienna, A-1090 Vienna, Austria; [i]Center for Non-Coding RNA in Technology and Health, University of Copenhagen, 1870 Frederiksberg C, Denmark; and [j]Santa Fe Institute, Santa Fe, NM 87501

Edited* by Peter Schuster, University of Vienna, Vienna, Austria, and approved January 9, 2015 (received for review July 7, 2014)

Phylogenomics heavily relies on well-curated sequence data sets that comprise, for each gene, exclusively 1:1 orthologos. Paralogs are treated as a dangerous nuisance that has to be detected and removed. We show here that this severe restriction of the data sets is not necessary. Building upon recent advances in mathematical phylogenetics, we demonstrate that gene duplications convey meaningful phylogenetic information and allow the inference of plausible phylogenetic trees, provided orthologs and

the true orthology relation Θ^*, which can be interpreted as a graph G_Θ whose vertices are genes and whose edges connect estimated (co)orthologs.

Recent advances in mathematical phylogenetics suggest that the estimated orthology relation Θ contains information on the structure of the species tree. To make this connection, we combine here three abstract mathematical results that are made precise in *Materials and Methods* below.

Hellmuth et al., PNAS, 2015, 112(7)

A formalization

Given a set X of species, a set $\mathcal{E}$ of events (e.g. speciation, duplication, etc) and a symbolic map $\delta : \binom{X}{2} \to \mathcal{E}$, can we construct a phylogenetic tree T on X and a labeling $t : V_{int}(T) \to \mathcal{E}$ that *represents* δ, that is, $\delta(x, y)$ equals the label of $lca(x, y)$ under t?

A formalization

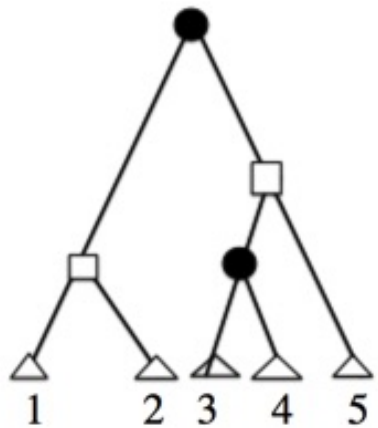
Given a set X of species, a set $\mathcal{E}$ of events (e.g. speciation, duplication, etc) and a symbolic map $\delta : \binom{X}{2} \to \mathcal{E}$, can we construct a phylogenetic tree T on X and a labeling $t : V_{int}(T) \to \mathcal{E}$ that *represents* δ, that is, $\delta(x, y)$ equals the label of $lca(x, y)$ under t?

Example: $X = \{1, \ldots, 5\}$, $\mathcal{E} = \{\bullet, \square\}$ and $\delta(1, 2) = \delta(3, 5) = \delta(4, 5) = \square$, and $\delta(x, y) = \bullet$ otherwise.

A formalization

Given a set X of species, a set $\mathcal{E}$ of events (e.g. speciation, duplication, etc) and a symbolic map $\delta : \binom{X}{2} \to \mathcal{E}$, can we construct a phylogenetic tree T on X and a labeling $t : V_{int}(T) \to \mathcal{E}$ that *represents* δ, that is, $\delta(x, y)$ equals the label of $lca(x, y)$ under t?

Example: $X = \{1, \ldots, 5\}$, $\mathcal{E} = \{\bullet, \square\}$ and $\delta(1, 2) = \delta(3, 5) = \delta(4, 5) = \square$, and $\delta(x, y) = \bullet$ otherwise.



A characterization

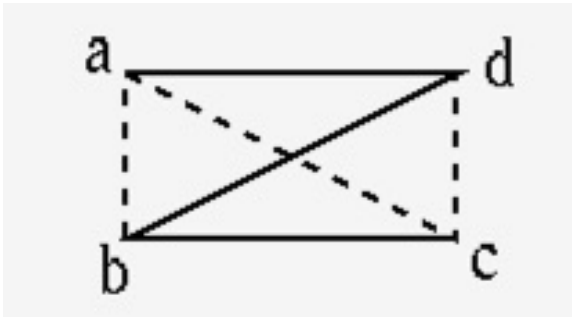
Theorem (Böcker and Dress, 1998)

Let $\delta : \binom{X}{2} \to \mathcal{E}$ *be a symbolic map. Then there exists a labelled phylogenetic tree on* X *representing* δ *if and only if* δ *is a* symbolic ultrametric *that is,* δ *satisfies*

(C1) *For all three elements* $x, y, z \in X$*, at least two of* $\delta(x, y)$*,* $\delta(y, z)$ *and* $\delta(x, z)$ *are the same.*

(C2) δ *is symmetric.*

(C3) *There exist no four elements* $a, b, c, d \in X$ *such that:*



Bottom-Up (Hellmuth, Hernandez-Rosales, H., Moulton, Stadler, Wieseke, 2011)

Input: Sets X and $\mathcal{E}$ and a symbolic map $\delta : \binom{X}{2} \to \mathcal{E}$.
Output: A unique labelled phylogenetic tree (T, t) that represents δ or the statement "δ is not a symbolic ultrametric".

(1) Construct a forest F consisting of singleton trees labelled by the elements in X.

(2) Iteratively look for *pseudo-cherries* (i.e. maximal subsets of X that have the same parent in the tree thus far constructed). In each iteration, collapse the found pseudo-cherries into a vertex and adjust δ accordingly.

A characterization

Corollary (Hellmuth, Hernandez-Rosales, H., Moulton, Stadler, Wieseke, 2011)

Suppose $\delta : \binom{X}{2} \to \{\square, \bullet\}$ *is a symbolic map. Then the following are equivalent:*

(i) BOTTOM-UP *completes when given* δ.

(ii) δ *is a symbolic ultrametric.*

(iii) *The graph with vertex set* X *and any two* $x, y \in X$ *joined by an edge if* $\delta(x, y) = \square$ *is a* co-graph, *that is, no four vertices induce a subgraph that is a path of length 3.*

(iii') *The graph with vertex set* X *and any two* $x, y \in X$ *joined by an edge if* $\delta(x, y) = \bullet$ *is a co-graph.*

Exploiting link with co-graphs

Corollary (Hellmuth, Hernandez-Rosales, H., Moulton, Stadler, Wieseke, 2011)

Let $\delta : X \times X \to \mathcal{E} = \{\square, \bullet\}$ be a symbolic map, and let K be a positive integer. Then the problem of deciding if there is a map $\delta^* : X \times X \to \mathcal{E}$ such that δ^* is a symbolic ultrametric and differs from δ in fewer than K values is NP-complete.

Furthermore, ...

Orthology Relation and Gene Tree Correction: Complexity Results

Manuel Lafond(✉) and Nadia El-Mabrouk

Department of Computer Science, Université de Montréal, Montréal, QC, Canada
lafonman@iro.umontreal.ca

Abstract. Tree-oriented methods for inferring orthology and paralogy relations between genes are based on reconciling a gene tree with a species tree. On the other hand, many tree-free methods, mainly based on sequence similarity, are also available. The link between orthology relations and gene trees has been formally considered recently from the angle of reconstructing phylogenies from orthology relations. Here, we rather consider this link from a correction point of view. While a gene tree induces a set of relations, the converse is not always true, as a set

WABI 2015: Algorithms in Bioinformatics pp 66-79

From trees to networks

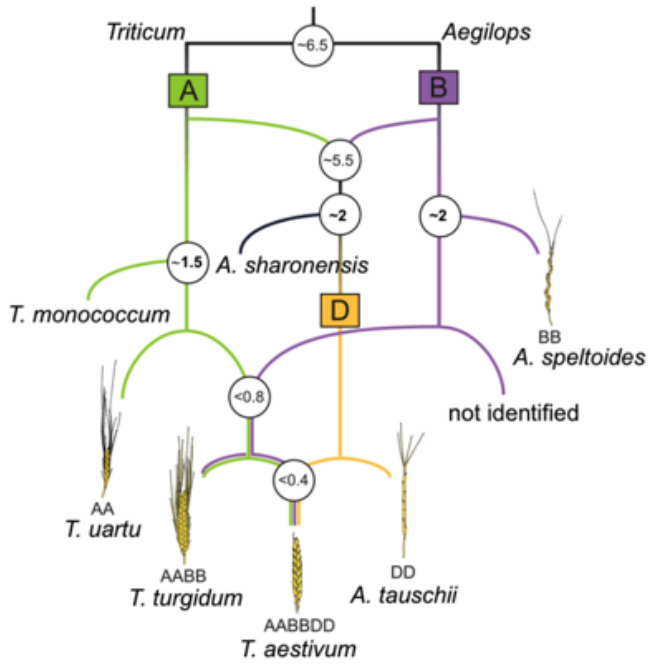
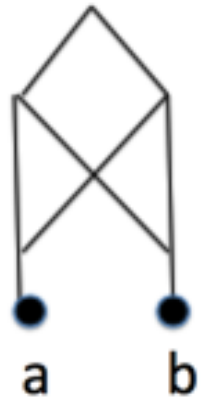


Figure: Marcussen et al, Science, 2014, 345:250092

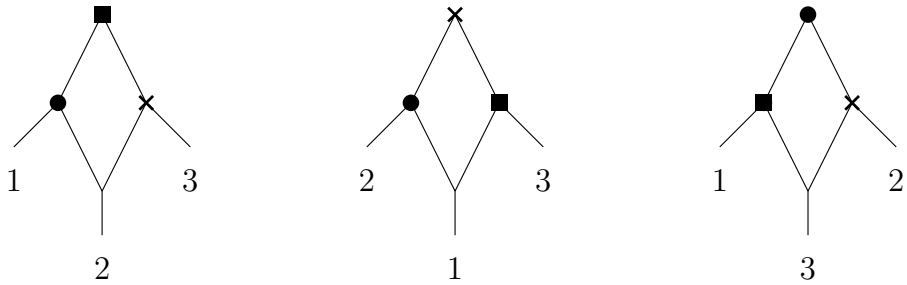
A *phylogenetic network* on a set X is a rooted directed acyclic graph with leaf set X such that the root has indegree 0, the leaves have indegree 1 and all other vertices have degree three.

Two main problems

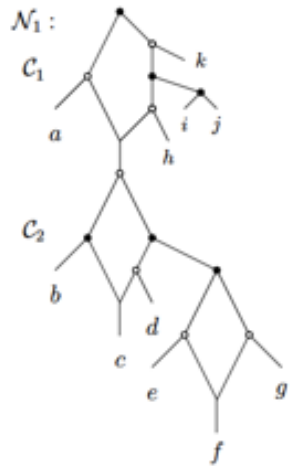
- non-uniqueness of lca



- non-identifiability using lca



We use level-1 networks



A *level-1 network* on X is a phylogenetic network on X such that when ignoring directions no two cycles share a vertex.

Note: Last common ancestor $lca_N(x, y)$ of any two leaves x and y in a level-1 network N is unique!

We use trivariate symbolic maps

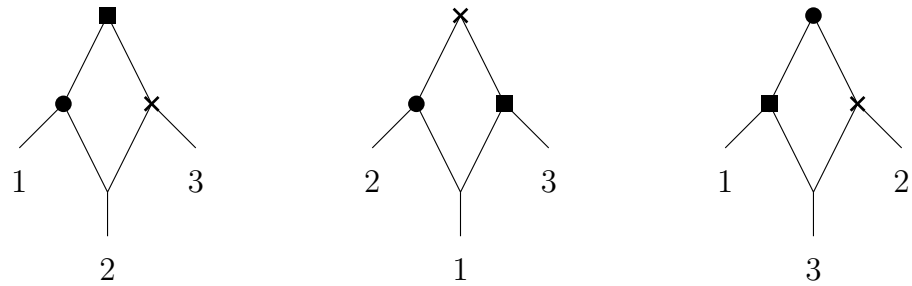


Figure: Label of $lca(1, 2, 3)$ is different in all 3 cases.

We use trivariate symbolic maps

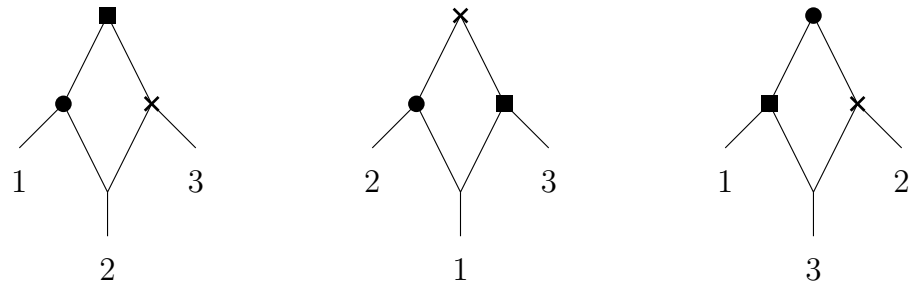


Figure: Label of $lca(1,2,3)$ is different in all 3 cases.

symbolic 3-dissimilarities i.e. maps $\delta : \binom{X}{\leq 3} \to \mathcal{E}$.

Note: This definition generalizes the concept of a symbolic 2-dissimilarity.

level-1 representations

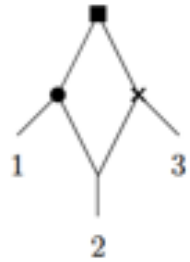


Figure: A level-1 representation of $\delta : \binom{X}{3} \to \{\bullet, \square, \times\}$ given by $\delta(x, y) = \bullet$, $\delta(2, 3) = \times$ and $\delta(1, 3) = \delta(1, 2, 3) = \square$.

Suppose $\delta : \binom{X}{\leq 3} \to \mathcal{E}$ is a symbolic 3-dissimilarity, N is a level-1 network on X, and $t : V_{int}(N)^- \to \mathcal{E}$ is a labelling from the set $V_{int}(N)^-$ of all interior vertices of N excluding those with indegree 2 in terms of $\mathcal{E}$. Then (N, t) is called a *level-1 representation of* δ if $\delta(x, y, z) = t(lca_N(x, y, z))$, for all $x, y, z \in X$.

Network-Popping (H., Scholz, 2016)

Input: A symbolic 3-dissimilarity $\delta : \binom{X}{\leq 3} \to \mathcal{E}$
Output: level-1 representation of δ or statement "δ does not have a level-1 representation".

(1) Find and build cycles supported by δ.

(2) Recurse:

(2a) Associate to δ and X a 2-dissimilarity δ' on a partitioning X' of X induced by the found cycles.

(2b) Apply Bottom-Up algorithm (Hellmuth *et. al.*, 2011) to δ' and X' to obtain a rooted phylogenetic tree T'.

(2c) Replace a leaf of T' with a cycle (if appropriate).

Theorem (H., Scholz, 2016)

Let $\delta : \binom{X}{3} \to M$ be a symbolic 3-dissimilarity. Then the following statements are equivalent:
(i) There exists a level-1 network that represents δ.
(ii) For input δ, Network-Popping *returns a labelled level-1 network which represents δ and which is unique up to isomorphism.*

Theorem (H., Scholz, 2016)

Let $\delta : \binom{X}{3} \to M$ *be a symbolic map,* $|X| \geq 6$*. Then* δ *can be represented by a labelled level-1 network if and only if for all subsets* $Y \subseteq X$ *of size* $|X| - 1$*, the restriction* $\delta|_Y$ *of* δ *to* Y *is representable by a level-1 network.*

Future directions

- Make the evolutionary scenario more realistic by
  - (i) including more events such as e.g. loss;
  - (ii) considering other classes of phylogenetic networks. This might involve going from a 3-dissimilarity to a k-dissimilarity, $k \geq 4$.
- Implement a divide an conquer version of our algorithm. – Runtime of NETWORK-POPPING is $\mathcal{O}(|X|^6)$.
- Is there a co-graph result analogue for level-1 networks?
- etc.