

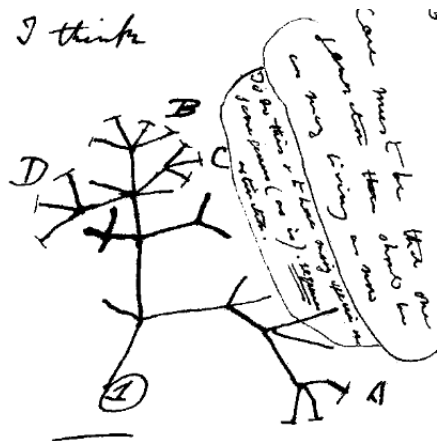
Lassoing phylogenetic trees

Katharina Huber,
University of East Anglia, UK

January 31, 2012

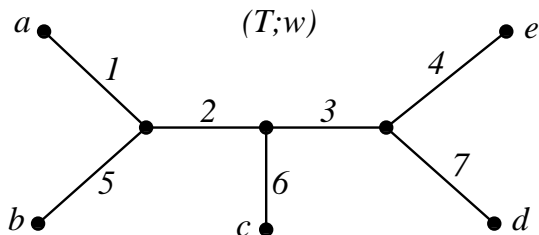


Darwin's tree



Phylogenetic tree $T = (V, E)$ on X

$$X = \{a, b, c, d, e\}$$



i.e.

- ▶ leaf set is X ,
- ▶ no vertices of degree 2,
- ▶ not necessarily binary,
- ▶ an edge weighting i.e. a map $\omega : E(T) \rightarrow \mathbb{R}_{\geq 0}$ that is strictly positive on all interior edges of T .

How many tree topologies

Number $b(n)$ of binary non-isomorphic tree topologies (all edges have weight 1) on n leaves, $n \geq 4$:

$$b(n) = (2n - 5)!! = 1 \times 3 \times 5 \times 7 \times \dots \times 2n - 5$$

e.g.

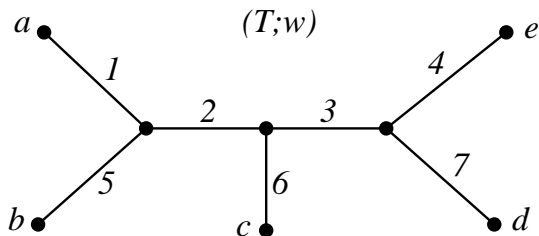
$$b(10) = 2027025 \text{ and } b(20) \sim 2 \times 10^{20}.$$

So asymptotic equivalence:

$$b(n) \sim \frac{1}{2\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-2}$$

Edge-weighted phylogenetic tree on X

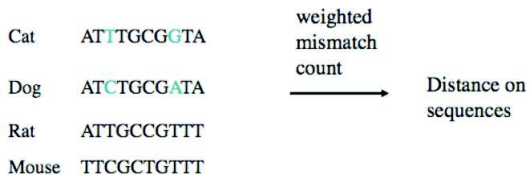
$$X = \{a, b, c, d, e\}$$



Shortest distances between leaves of tree induces distances between the elements on X e.g.

$$d_{(T;w)}(a, d) = 13.$$

Extracting distances from sequences, ...



d a *dissimilarity* on X , i.e. $d : X \times X \rightarrow \mathbb{R}$ such that $d(x, x) = 0$ and $d(x, y) = d(y, x)$, for all $x, y \in X$.

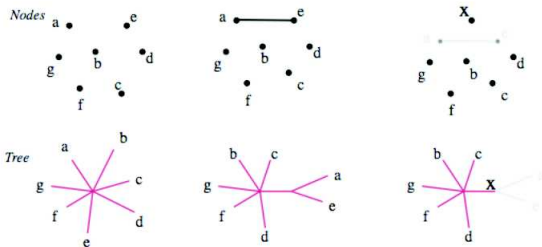
When does there exist a phylogenetic tree $(T; \omega)$ with edge-weight function ω such that:

$$d_{(T; \omega)}(x, y) = d(x, y),$$

for all $x, y \in X$?

Existence: Neighbor-Joining (Saito/Nei 1987)

- ❖ The most widely used distance based phylogenetic method.
- ❖ Happy to give a tree whatever the data
- ❖ Recovers distance correctly under a well-understood condition



BioNJ (Gascuel 1997)

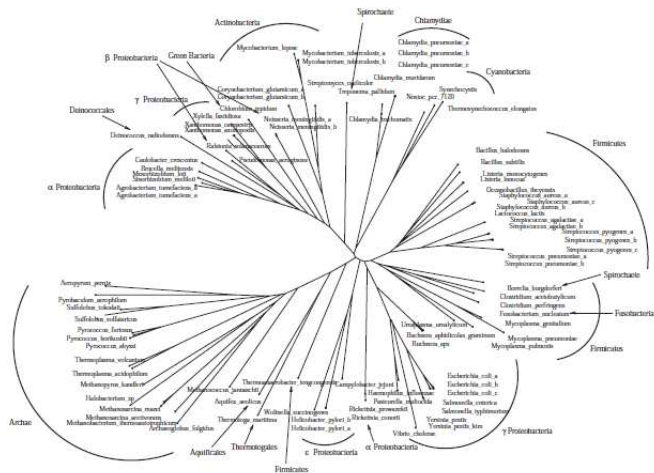


Fig. 2. Phylogenetic tree for 91 prokaryotic genomes produced by BioNJ on the 'matched distances' matrix.

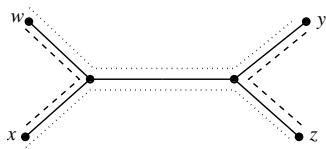
d a dissimilarity on X .

If there exists an edge-weighted phylogenetic tree $(T; \omega)$ such that $d_{(T; \omega)}(x, y) = d(x, y)$, for all $x, y \in X$, when is that edge-weighted phylogenetic tree unique, that is, up to isomorphism, $(T; \omega)$ is the only edge-weighted phylogenetic tree that satisfies this property?

Uniqueness: 4-point condition

d a dissimilarity on X . For all $w, x, y, z \in X$:

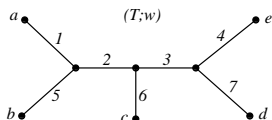
$$d(w, x) + d(y, z) \leq \max\{d(w, z) + d(y, x), d(x, z) + d(y, w)\}$$



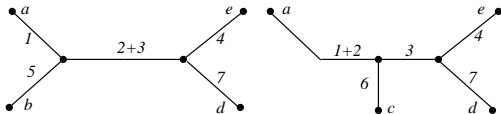
Note: If d satisfies this condition, then the edge-weighted phylogenetic tree is unique (up to isomorphism). Considered as a graph, the tight span associated to d is that tree.

From edge-weighted phylogenetic trees to collections of weighted quartets

$X = \{a, b, c, d, e\}$



induces weighted quartets:



So have a collection of weighted quartets!

Quartet weight functions

$\mu : \mathcal{Q}(X) := \{\text{quartet with leaf set in } X\} \rightarrow \mathbb{R}_{\geq 0}$ such that:

(T1) For all $a, b, c, d \in X$, at least two of $\mu(ab|cd)$, $\mu(ac|bd)$, and $\mu(ad|bc)$ are equal to 0.

(T2) For all $x \in X - \{a, b, c, d\}$, if $\mu(ab|cd) > 0$, then

$$\mu(ab|cx), \mu(ab|dx) > 0 \text{ or } \mu(ax|cd), \mu(bx|cd) > 0.$$

(T3) For all $a, b, c, d, e \in X$, if $\mu(ab|cd) > \mu(ab|ce) > 0$, then

$$\mu(ae|cd) = \mu(ab|cd) - \mu(ab|ce).$$

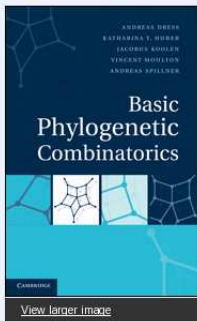
(T4) For all $a, b, c, d, e \in X$, if $\mu(ab|cd) > 0$ and $\mu(bc|de) > 0$, then

$$\mu(ab|de) = \mu(ab|cd) + \mu(bc|de).$$

Theorem (Grünewald, Huber, Moulton, Semple, Journal of Mathematical Biology (JMB), 2008)

Let $\mu : \mathcal{Q}(X) \rightarrow \mathbb{R}_{\geq 0}$ be a map. Then $\mu = \mu_{\mathcal{T}}$ for some edge-weighted phylogenetic tree \mathcal{T} on X if and only if μ satisfies conditions (T1)-(T4). Moreover, if such a tree exists, then, up to isomorphism and the weights of the pendant edges, \mathcal{T} is unique.

Note: Similar result for *binary* edge-weighted phylogenetic trees Erdős and Dress, Annals of Combinatorics, 2003.



Basic Phylogenetic Combinatorics

Andreas Dress, Universität Bielefeld, Germany

Katharina T. Huber, University of East Anglia

Jacobus Koolen, Pohang University of Science and Technology, South Korea

Vincent Moulton, University of East Anglia

Andreas Spillner, Ernst-Moritz-Arndt-Universität Greifswald, Germany

Hardback

ISBN:9780521768320

276pages

55 b/w illus.

Dimensions: 228 x 152 mm

Weight: 0.54kg

Not yet published - available from January 2012

\$65.00 (Z)

[Preorder this title >](#)

[Request an examination copy](#)

Phylogenetic combinatorics is a branch of discrete applied mathematics concerned with the combinatorial description and analysis of phylogenetic trees and related mathematical structures such as phylogenetic networks and tight spans. Based on a natural conceptual framework, the book focuses on the interrelationship between the principal options for encoding phylogenetic trees: split systems, quartet systems and metrics. Such encodings provide useful options for analyzing and manipulating phylogenetic trees

But what happens if we have missing distances?



Available online at www.sciencedirect.com



Discrete Mathematics 276 (2004) 229–248

DISCRETE
MATHEMATICS

www.elsevier.com/locate/disc

On the extension of a partial metric to a tree metric

Alain Guénoche^a, Bruno Leclerc^b, Vladimir Makarenkov^{c,d}

^a*Institut de Mathématiques de Luminy, 163 avenue de Luminy,
F-13009 Marseille, France*

^b*Centre d'Analyse et de Mathématiques Sociales, Ecole des Hautes Études en Sciences Sociales,
54 bd Raspail, F-75270 Paris Cedex 06, France*

^c*Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succ. Centre-ville,
Montréal, Québec, Canada H3C 3J7*

^d*Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia*

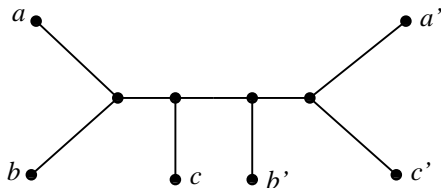
Abstract

Farach et al. (Algorithmica 13 (1995) 155–179) defined problem MCA (matrix completion to additive) and proved it to be NP-complete: given a partial dissimilarity d on a finite set X , does there exist a tree metric extending d to all pairs of elements of X . We use a previously described simple method of phylogenetic reconstruction, and its extension to partial dissimilarities, to characterize some classes of polynomial instances of MCA and of a related problem. We point out that these problems admit many other polynomial instances. We focus particularly on two classes of generalized cycles, together with the corresponding maximal acyclic graphs (2-trees and 2d-trees).

© 2003 Elsevier B.V. All rights reserved.

Uniqueness: Two phenomena (I)

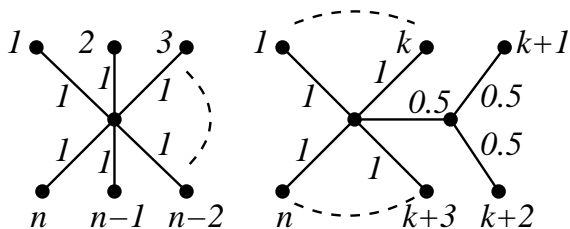
$$X = \{a, b, c, a', b', c'\}; \mathcal{L} = (\{a, b, c\}) \cup (\{a', b', c'\}).$$



Note: Central interior edge can be any positive real number

Uniqueness: Two phenomena (II)

$$X = \{1, 2, \dots, n\}, n \geq 4, \text{ and } \mathcal{L} = \binom{X}{2} - \{\{k+1, k+2\}\}.$$

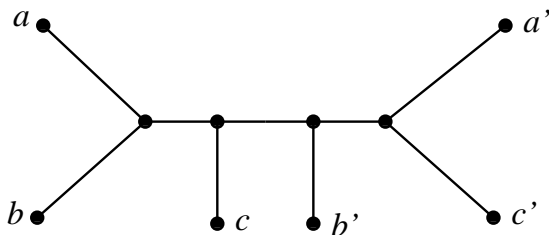


Note: Distance between a and b with $ab := \{a, b\} \in \mathcal{L}$ are the same!

Regarding phenomenon (I): Edge-weight lasso

T a phylogenetic tree on X and $\mathcal{L} \subseteq \binom{X}{2}$ with $\bigcup_{A \in \mathcal{L}} A = X$. Then we say that \mathcal{L} is an *edge-weight lasso* for T if $\omega = \omega'$ holds for all edge weightings ω, ω' of T with $d_{(T, \omega)}|_{\mathcal{L}} = d_{(T, \omega')}|_{\mathcal{L}}$.

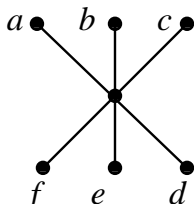
$X = \{a, b, c, a', b', c'\}$; $\mathcal{L} = (\{a, b, c\}) \cup (\{a', b', c'\}) \cup \{aa', bb', cc'\}$.



Regarding phenomenon (II): Topological lasso

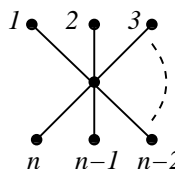
T a phylogenetic tree on X and $\mathcal{L} \subseteq \binom{X}{2}$ with $\bigcup_{A \in \mathcal{L}} A = X$. Then we say that \mathcal{L} is a *topological lasso for T* if $T \simeq T'$ holds for any phylogenetic tree T' on X for which there exist edge weightings ω of T and ω' of T' with $d_{(T,\omega)}|_{\mathcal{L}} = d_{(T',\omega')}|_{\mathcal{L}}$.

$X = \{a, b, c, d, e, f, g, h\}$ and $\mathcal{L} = \binom{X}{2}$.

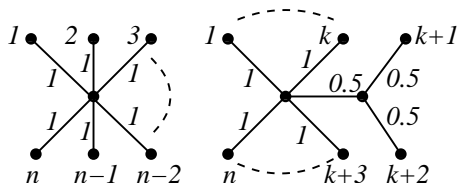


Not a topological lasso

$$X = \{1, 2, \dots, n\}, n \geq 4, \text{ and } \mathcal{L} = \binom{X}{2} - \{\{k+1, k+2\}\}.$$

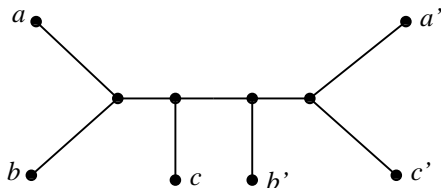


But, ...

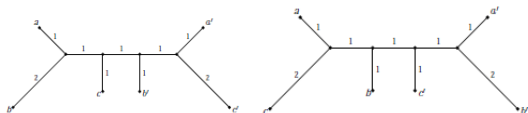


Edge-weight lasso $\not\Rightarrow$ Topological lasso

$$X = \{a, b, c, a', b', c'\}; \mathcal{L} = (\{a, b, c\}) \cup (\{a', b', c'\}) \cup \{aa', bb', cc'\}.$$

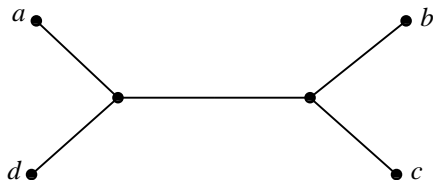


\mathcal{L} is not a topological lasso for above phylogenetic tree since



Topological lasso $\not\Rightarrow$ Edge-weight lasso

$X = \{a, b, c, d\}$ and $\mathcal{L} = \{ab, bc, cd, da\}$.

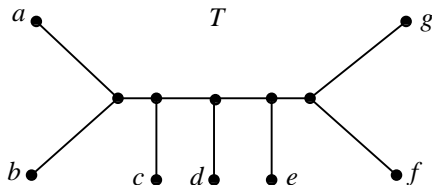


Note: The weight of the pendant edges is not fixed by the given distances. However the weight of the interior edge is.

Strong lasso

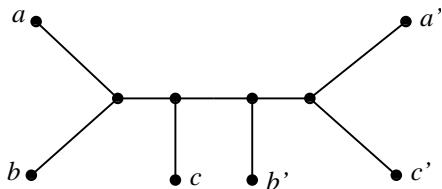
T a phylogenetic tree on X with $\bigcup_{A \in \mathcal{L}} A = X$. Then we say that \mathcal{L} is a *strong lasso* of T if $(T, \omega) \equiv (T', \omega')$ holds (i.e. there exists a graph isomorphism from T to T' that respects X and the edge-weights), for every given edge weighting ω of T , for any phylogenetic tree T' on X and any edge weighting ω' of T' with $d_{(T, \omega)}|_{\mathcal{L}} = d_{(T', \omega')}|_{\mathcal{L}}$.

$X = \{a, b, \dots, g\}$; $\mathcal{L} = \{ab, ad, bc, be, cd, cf, de, dg, ef, fg\}$



Not a strong lasso

$$X = \{a, b, c, a', b', c'\}; \mathcal{L} = \left(\binom{\{a,b,c\}}{2}\right) \cup \left(\binom{\{a',b',c'\}}{2}\right) \cup \{aa', bb', cc'\}.$$



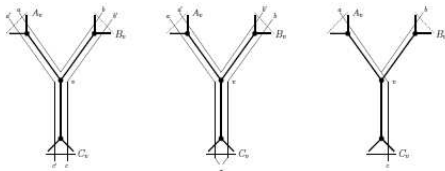
Note: Not a topological lasso but an edge-weight lasso

Theorem (Dress, Huber, Steel, JMB, 2011)

- (i) *If $n \geq 4$ holds and \mathcal{L} is a topological lasso for T , then the graph (X, \mathcal{L}) must be connected.*
- (ii) *If \mathcal{L} is an edge-weight lasso for T , then the graph (X, \mathcal{L}) must be strongly non-bipartite i.e., every connected component of the graph (X, \mathcal{L}) is not bipartite.*
- (iii) *In particular, the graph (X, \mathcal{L}) must be connected and non-bipartite if \mathcal{L} is a strong lasso for T .*

Covers of binary phylogenetic trees

- T a binary phylogenetic tree on X . Then $\mathcal{L} \subseteq \binom{X}{2}$ is called a
- ▶ *pointed cover* if there exists some $x \in X$ such that for all interior vertices v of T the situation indicated in the figure below (left) holds
 - ▶ *triplet cover* if for every interior vertex v of T the situation indicated in the figure below (right) holds

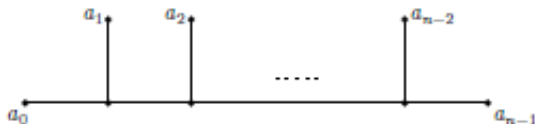


Note: Left: pointed cover, Right: triplet cover.

An example

$$X = \{a_0, a_1, \dots, a_{n-2}, a_{n-1}\}, \quad n \geq 4,$$

$$\mathcal{L} := \{a_0 a_i : i = 1, \dots, n-1\} \cup \{a_{i-1} a_i : i = 2, \dots, n-1\}.$$



Note: \mathcal{L} is a pointed cover as well as a triplet cover. – see Chaiken, Dewdney, Slater, Alg. Disc. Math. 1983, and Barthelemy, Guenoah, Trees and Proximity Representations, 1991, for more on this type of triplet cover.

A combinatorial characterization of triplet covers

Triplet cover \mathcal{L} induces a set $\mathcal{C}(\mathcal{L}) \subseteq \binom{X}{3}$ by putting

$$C(ab, ac, bc) := \{a, b, c\}, \text{ for } ab, ac, bc \in \mathcal{L}.$$

$\mathcal{C} \subseteq \binom{X}{3}$ with $\bigcup_{C \in \mathcal{C}} C = X$. Then \mathcal{C} is a triplet cover of some binary phylogenetic tree on X if and only if

$$\left| \bigcup_{C \in \mathcal{C}'} C \right| \geq |\mathcal{C}'| + 2.$$

for all non-empty subsets $\mathcal{C}' \subseteq \mathcal{C}$. (Dress, Steel, Applied Mathematics Letters, 2009)

Theorem (Dress, Huber, Steel, JMB, 2011)

- (i) Every triplet cover \mathcal{L} of a binary phylogenetic tree T on X lassos the edge weights for T . Furthermore, $(T, \omega) \equiv (T', \omega')$ must hold for every edge weighting ω of T and every pair (T', ω') that consists of a phylogenetic T' on X and an edge weighting (where interior edges may have weight zero) ω' of T' such that \mathcal{L} is also a triplet cover of T' and $d_{(T, \omega)}|_{\mathcal{L}} = d_{(T', \omega')}|_{\mathcal{L}}$ holds.
- (ii) If a subset \mathcal{L} of $\binom{X}{2}$ is a pointed cover of a binary phylogenetic tree T on X , then \mathcal{L} is a strong lasso for T .

Theorem (Dress, Huber, Steel, JMB, 2011)

Given a phylogenetic tree T on X and a bipartition $\{A, B := X - A\}$ of X . Then the following assertions are equivalent:

- (i) The subset $\{ab : a \in A \text{ and } b \in B\}$ of $\binom{X}{2}$ is a topological lasso for T .
- (iii) $A \cap c \neq \emptyset \neq B \cap c$ holds for every 2-subset c of X whose elements form a cherry in T .

More to come but many thanks for listening for now!