Recent Progress in Phylogenetic Combinatorics

Notes on joint work with **Katharina Huber**, **Jack Koolen**, and **Vincent Moulton**

to be presented Dec 18, 2007, at the workshop

Future Directions in Phylogenetic Methods and Models

The Isaac Newton Institute for Mathematical Sciences

What is Phylogenetic Combinatorics?

Phylogenetic combinatorics deals with the combinatorial aspects of phylogenetic-tree reconstruction.

What is Phylogenetic Combinatorics?

Phylogenetic combinatorics deals with the combinatorial aspects of phylogenetic-tree reconstruction.

A starting point was the following observation:

What is Phylogenetic Combinatorics?

Phylogenetic combinatorics deals with the combinatorial aspects of phylogenetic-tree reconstruction.

A starting point was the following observation:

Given a metric $D : X \times X \to \mathbf{R}$ representing the approximative genetic distances $D(x, y), \ldots$ between the members $x, y, \ldots$ of a finite collection X of taxa, it was shown in that the following assertions relating to the *object of desire*, a "phylogenetic X-tree", all are equivalent:

What is Phylogenetic Combinatorics? cont.

(M-i) The "tight span"

$$T_D := \left\{ f \in \mathbf{R}^X : \forall_{x \in X} f(x) = \sup(D(x,y) - f(y)) : y \in X \right\}$$

of D is an **R**-*tree*.

What is Phylogenetic Combinatorics? cont.

(M-i) The "tight span"

$$T_D := \left\{ f \in \mathbf{R}^X : \forall_{x \in X} f(x) = \sup(D(x,y) - f(y) : y \in X) \right\}$$

of D is an **R**-*tree*.

(M-ii) There exists an *edge-weighted* X*-tree* $(V, E; \ell)$ — i.e., a finite tree (V, E) with vertex set V and edge set E such that V contains X and every vertex in $V - X$ has degree at least 3, and a (positive) edge weighting $\ell : E \to \mathbf{R}_{>0}$ that assigns a positive length $\ell(e)$ to every edge e in E — such that D is the restriction to X of the *shortest-path metric* induced by ℓ on V (i.e., the largest metric d on V with $d(u,v) \leq \ell(\{u,v\})$ for all $\{u,v\} \in E$.

What is Phylogenetic Combinatorics? cont.

(M-iii) There exists a map $w : \mathcal{S}(X) \to \mathbf{R}_{\geq 0}$ from the set $\mathcal{S}(X)$ of all bi-partitions — or *splits* $S = \{A, B\}$ — of X into the set $\mathbf{R}_{\geq 0}$ of non-negative real numbers such that

What is Phylogenetic Combinatorics? cont.

(M-iii) There exists a map $w : \mathcal{S}(X) \to \mathbf{R}_{\geq 0}$ from the set $\mathcal{S}(X)$ of all bi-partitions — or *splits* $S = \{A, B\}$ — of X into the set $\mathbf{R}_{\geq 0}$ of non-negative real numbers such that

- given any two splits $S = \{A, B\}$ and $S' = \{A', B'\}$ in $\mathcal{S}(X)$ with $w(S), w(S') \neq 0$, at least one of the four intersections $A \cap A'$, $B \cap A'$, $A \cap B'$, and $B \cap B'$ is empty and

What is Phylogenetic Combinatorics? cont.

(M-iii) There exists a map $w : \mathcal{S}(X) \to \mathbf{R}_{\geq 0}$ from the set $\mathcal{S}(X)$ of all bi-partitions — or *splits* $S = \{A, B\}$ — of X into the set $\mathbf{R}_{\geq 0}$ of non-negative real numbers such that

- given any two splits $S = \{A, B\}$ and $S' = \{A', B'\}$ in $\mathcal{S}(X)$ with $w(S), w(S') \neq 0$, at least one of the four intersections $A \cap A'$, $B \cap A'$, $A \cap B'$, and $B \cap B'$ is empty and
- $D(x, y) = \sum_{S \in \mathcal{S}(X : x \leftrightarrow y)} w(S)$ holds where

$$\mathcal{S}(X : x \leftrightarrow y) := \left\{ \{A, B\} \in \mathcal{S}(X) : x \in A, y \in B \right\}$$

denotes the set of splits $S = \{A, B\} \in \mathcal{S}(X)$ that *separate* x and y.

What is Phylogenetic Combinatorics? cont.

(M-iii) There exists a map $w : \mathcal{S}(X) \to \mathbf{R}_{\geq 0}$ from the set $\mathcal{S}(X)$ of all bi-partitions — or *splits* $S = \{A, B\}$ — of X into the set $\mathbf{R}_{\geq 0}$ of non-negative real numbers such that

- given any two splits $S = \{A, B\}$ and $S' = \{A', B'\}$ in $\mathcal{S}(X)$ with $w(S), w(S') \neq 0$, at least one of the four intersections $A \cap A'$, $B \cap A'$, $A \cap B'$, and $B \cap B'$ is empty and
- $D(x, y) = \sum_{S \in \mathcal{S}(X : x \leftrightarrow y)} w(S)$ holds where

$$\mathcal{S}(X : x \leftrightarrow y) := \big\{\{A, B\} \in \mathcal{S}(X) : x \in A, y \in B\big\}$$

denotes the set of splits $S = \{A, B\} \in \mathcal{S}(X)$ that *separate* x and y.

(M-iv) $D(x, y) + D(u, v) \leq \max\big(D(x, u) + D(y, v), D(x, v) + D(y, u)\big)$ holds for all $x, y, u, v \in X$.

What is Phylogenetic Combinatorics? cont.

Moreover, the metric space T_D actually coincides in this case with the **R**-tree that is canonically associated with an edge-weighted X-tree $(V, E; \ell)$.

What is Phylogenetic Combinatorics? cont.

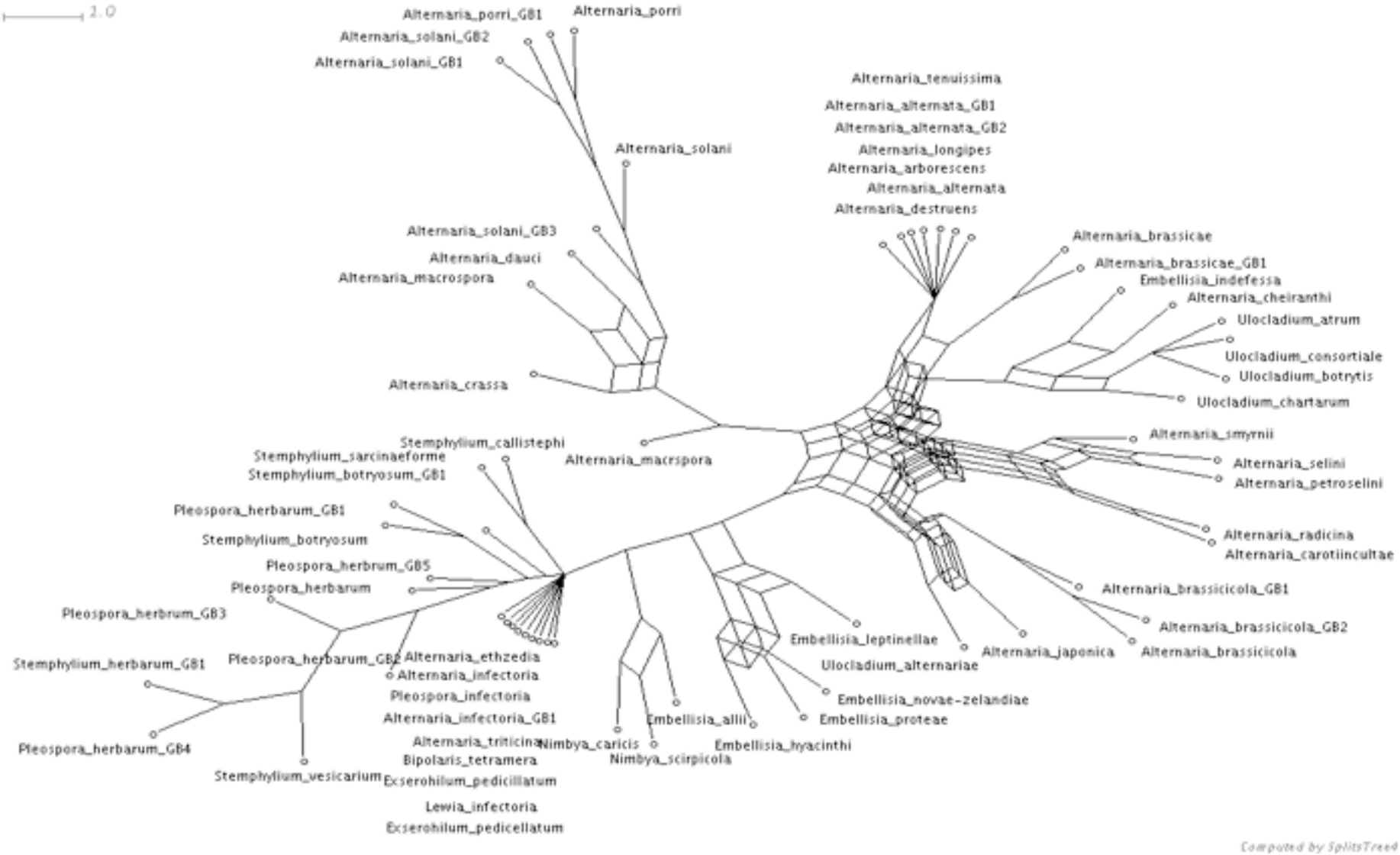
Moreover, the metric space T_D actually coincides in this case with the **R**-tree that is canonically associated with an edge-weighted X-tree $(V, E; \ell)$.

This observation suggested to further investigate (1) the tight-span construction and (2) representations of metrics by weighted split systems with various specific properties, even if the metric in question would not satisfy the very special properties described above. These investigations have, in turn, given rise to a full-fledged research program dealing many diverse aspects of these two topics.

What is Phylogenetic Combinatorics? cont.

Here, I will focus on rather new developments relating to *block decomposition* and *virtual cut points* of metric spaces that allow to canonically decompose any given finite metric space into a sum of pairwise compatible *block metrics*, thus providing a far-reaching generalization of the result recalled above.

An Instructive Example: A Network of Fungi



An Instructive Example: A Network of Fungi

Can we reconstruct this network and its "blocks" (or "two-connected components") from the metric it induces on the set X of fungi-labelled vertices?

An Instructive Example: A Network of Fungi

Can we reconstruct this network and its "blocks" (or "two-connected components") from the metric it induces on the set X of fungi-labelled vertices?

That's what block-decomposition theory is all about:

An Instructive Example: A Network of Fungi

Can we reconstruct this network and its "blocks" (or "two-connected components") from the metric it induces on the set X of fungi-labelled vertices?

That's what block-decomposition theory is all about:

Indeed, somehow in analogy to Dan Gusfield's theory of **galled trees** that identifies "blobs" in **sequence space** using (median) networks, block-decomposition theory allows us to identify the "blocks" within the framework of **metric-space** theory.

A Basic Concept: Block Realizations of Metrics

Given a (proper) metric

$$D : X^2 \to \mathbf{R} : (x, y) \mapsto xy$$

defined on a finite set X, a **graph realization** of D is a pair (G, ℓ) consisting of a finite connected graph $G = (V, E)$ and a **length-assigning map**

$$\ell : E \to \mathbf{R}_{>0} : \{u, v\} \mapsto \ell(u, v)$$

such that $X \subseteq V$ and $xy = d_\ell(x, y)$ holds for all $x, y \in X$, where d_ℓ denotes the metric induced by ℓ on V, i.e., the (necessarily unique and proper) largest metric defined on V for which $d_\ell(u, v) \leq \ell(u, v)$ holds for every edge $\{u, v\} \in E$.

A Basic Concept: Block Realizations of Metrics, cont.

While a metric can have several (non-equivalent) graph realizations (even if shortest total length $\ell(G) := \sum_{e \in E} \ell(e)$ is required), it has been observed occasionally that graph realizations satisfying certain additional, rather specific constraints (mostly structural constraints combined with some shortest-length requirements, but not necessarily implying shortest total length) can sometimes be shown to be uniquely determined — up to canonical isomorphism — by such constraints.

A Basic Concept: Block Realizations of Metrics, cont.

Here, we will focus on a specific class of graph realizations, the **block realizations**:

A Basic Concept: Block Realizations of Metrics, cont.

Here, we will focus on a specific class of graph realizations, the **block realizations**:

We define a graph realization (G, ℓ) of a finite metric D to be a **block realization** of D if the graph $G = (V, E)$ is a **block graph** (i.e., a connected graph whose 2-connected components are cliques) and every vertex v in $V - X$ has degree at least 3 and is a **cut point** of G, that is, it is a vertex in V such that the graph $G - v$ induced by G on $V - \{v\}$ (i.e., the graph $G - v := (V - \{v\}, E \cap \binom{V - \{v\}}{2})$) is disconnected.

A Second Basic Concept: Compatible Decompositions of Metrics

Next, we consider decompositions

$$D = d_1 + d_2 + \cdots + d_k$$

of a given metric

$$D : X \times X \to \mathbf{R} : (x, y) \mapsto xy$$

defined on a finite set X into a sum of metrics $d_1, d_2, \ldots, d_k$ also defined on X that will be required to satisfy a very specific condition:

Compatible Decompositions of Metrics, cont.

- Given any finite index set I, we define an I-indexed family $\mathcal{D} = (d_i)_{i \in I}$ of metrics on X to be a **compatible decomposition** of D if

Compatible Decompositions of Metrics, cont.

- Given any finite index set I, we define an I-indexed family $\mathcal{D} = (d_i)_{i \in I}$ of metrics on X to be a **compatible decomposition** of D if

  – any two metrics d_i and d_j for distinct $i, j \in I$ are linearly independent (considered as vectors in $\mathbf{R}^{X \times X}$),

Compatible Decompositions of Metrics, cont.

- Given any finite index set I, we define an I-indexed family $\mathcal{D} = (d_i)_{i \in I}$ of metrics on X to be a **compatible decomposition** of D if

  – any two metrics d_i and d_j for distinct $i, j \in I$ are linearly independent (considered as vectors in $\mathbf{R}^{X \times X}$),

  – $D = \sum_{i \in I} d_i$ holds, and

Compatible Decompositions of Metrics, cont.

- Given any finite index set I, we define an I-indexed family $\mathcal{D} = (d_i)_{i \in I}$ of metrics on X to be a **compatible decomposition** of D if

  – any two metrics d_i and d_j for distinct $i, j \in I$ are linearly independent (considered as vectors in $\mathbf{R}^{X \times X}$),

  – $D = \sum_{i \in I} d_i$ holds, and

  – there exist points x_{ij}, x_{ji} in X for any two distinct indices i, j in I such that $d_i(x_{ij}, y)\, d_j(x_{ji}, y) = 0$ holds for every $y \in X$.

Compatible Decompositions of Metrics, cont.

- Given any finite index set I, we define an I-indexed family $\mathcal{D} = (d_i)_{i \in I}$ of metrics on X to be a **compatible decomposition** of D if

  – any two metrics d_i and d_j for distinct $i, j \in I$ are linearly independent (considered as vectors in $\mathbf{R}^{X \times X}$),

  – $D = \sum_{i \in I} d_i$ holds, and

  – there exist points x_{ij}, x_{ji} in X for any two distinct indices i, j in I such that $d_i(x_{ij}, y)\, d_j(x_{ji}, y) = 0$ holds for every $y \in X$.

  I.e., every $y \in X$ has distance 0 to

Compatible Decompositions of Metrics, cont.

- Given any finite index set I, we define an I-indexed family $\mathcal{D} = (d_i)_{i \in I}$ of metrics on X to be a **compatible decomposition** of D if

  – any two metrics d_i and d_j for distinct $i, j \in I$ are linearly independent (considered as vectors in $\mathbf{R}^{X \times X}$),

  – $D = \sum_{i \in I} d_i$ holds, and

  – there exist points x_{ij}, x_{ji} in X for any two distinct indices i, j in I such that $d_i(x_{ij}, y)\, d_j(x_{ji}, y) = 0$ holds for every $y \in X$.

  I.e., every $y \in X$ has distance 0 to

  - either the point x_{ij} relative to d_i

Compatible Decompositions of Metrics, cont.

- Given any finite index set I, we define an I-indexed family $\mathcal{D} = (d_i)_{i \in I}$ of metrics on X to be a **compatible decomposition** of D if

  – any two metrics d_i and d_j for distinct $i, j \in I$ are linearly independent (considered as vectors in $\mathbf{R}^{X \times X}$),

  – $D = \sum_{i \in I} d_i$ holds, and

  – there exist points x_{ij}, x_{ji} in X for any two distinct indices i, j in I such that $d_i(x_{ij}, y)\, d_j(x_{ji}, y) = 0$ holds for every $y \in X$.

  I.e., every $y \in X$ has distance 0 to

  - either the point x_{ij} relative to d_i
  - or the point x_{ji} relative to d_j.

From Block Realizations to Compatible Decompositions

Given any block graph $G = (V, E)$, let $\mathcal{B}(G)$ denote the collection of all **blocks** $B \subseteq V$ of G (i.e., all those subsets B of the vertex set V that make up the vertex set of a 2-connected component of G) and, given a block realization (G, ℓ) of a metric D defined on a finite set X, associate to any **block** $B \in \mathcal{B}(G)$ of G the metric $d_{(B|\ell)}$ defined on X by

$$d_{(B|\ell)} : X \times X \to \mathbf{R} : (x, y) \mapsto d_\ell(x_{(B|\ell)}, y_{(B|\ell)})$$

where $x_{(B|\ell)}$ denotes, for any $x \in X$, the (necessarily unique!) point in B that minimizes the distance (relative to d_ℓ) to x.

The Main Result

Referring to these concepts, the following result can be established:

Theorem

Associating, to any block realization (G, ℓ) of a metric D defined on a finite set X, the collection

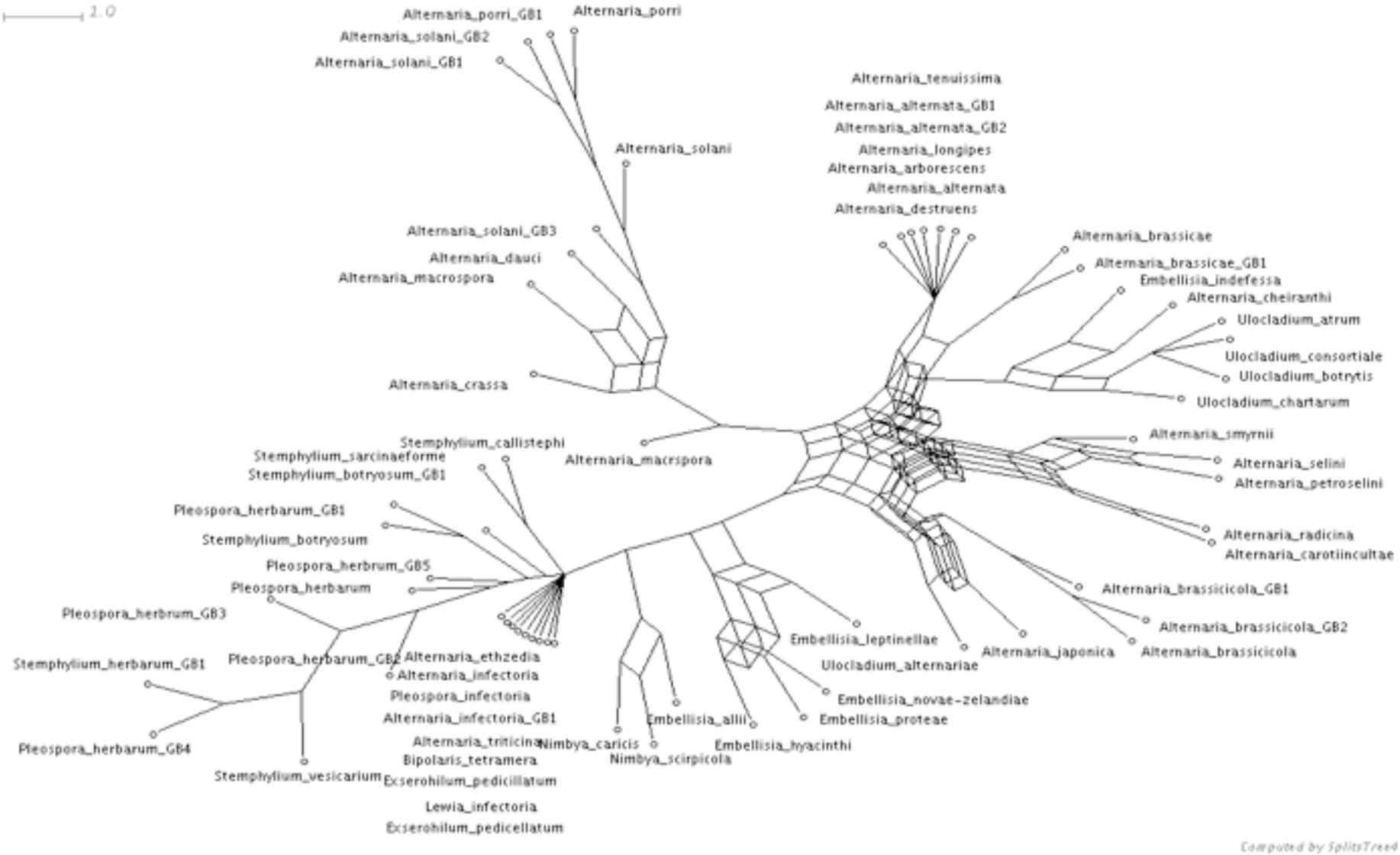
$$\mathcal{D}(G|\ell) := \{d_{(B|\ell)} : B \in \mathcal{B}(G)\}$$

sets up a canonical one-to-one correspondence between (isomorphism classes of) block realizations and compatible decompositions of D.

The Main Result

This result can be used to establish that **shortest block realizations** of a finite metric are (essentially) unique — a fact that allows us to generalize the well-known fact quoted above regarding the unique decomposition of **tree metrics** into a sum of pairwise compatible split metrics to arbitrary metrics, noting that the components in the shortest block realizations of a finite metric D are split metrics if and only if D satisfies the four-point condition (M-iv).

The Fungal Network again



Compatible Decompositions of Metrics, cont.

And we define D to be **block indecomposable** — or a **block metric** — if there exists no compatible decomposition $\mathcal{D}$ of D containing two or more members.

The Block-Decomposition Theorem for Finite Metrics

Theorem

(i) *Given any metric D defined on a finite set X, there exists a unique compatible decomposition $\mathcal{D} = \mathcal{D}_D$ of D, called* **the block decomposition** *of D, all of whose summands are block metrics.*

The Block-Decomposition Theorem for Finite Metrics, cont.

(ii) The blocks in that decomposition $\mathcal{D} = \mathcal{D}_D$ correspond in a one-to-one fashion to the 2-**connected components** of the tight span

$$T(D) := \{f \in \mathbf{R}^X : \forall_{x \in X} f(x) = \sup_{y \in X} \big(xy - f(y)\big)\}$$

of D, i.e., the connected components of the complement of the subset $T(D)_{cut}$ of $T(D)$ consisting of all $f \in T(D)$ for which $T_f(D) := T(D) - \{f\}$ is disconnected, and there is no neighbourhood (U, f) of f in $T(D)$ that is homeomorphic to the interval $([-1, 1], 0)$ except in case $f(x) = 0$ holds for some $x \in X$.