6 Phylogenetic Networks

Real evolutionary data often contains a number of different and sometimes conflicting phylogenetic signals, and thus do not always clearly support a unique tree. To address this problem, Hans-Jürgen Bandelt and Andreas Dress developed the method of *split decomposition*.

For ideal data, this method gives rise to a tree, whereas less ideal data are represented by a tree-like network that may indicate evidence for different and conflicting phylogenies.

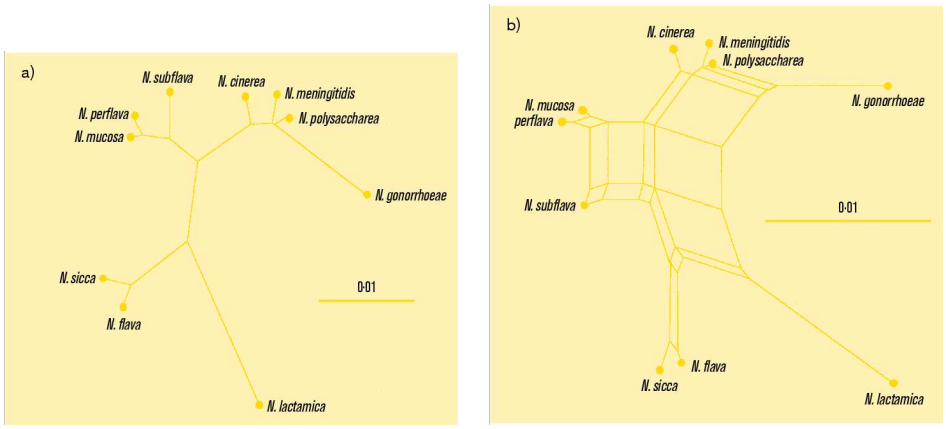
The following lectures are based on:

Hans-Jürgen Bandelt and Andreas W. M. Dress. *A canonical decomposition theory for metrics on a finite set*, Advances in Mathematics, 92(1):47-105 (1992)

Daniel H. Huson, *SplitsTree: analyzing and visualizing evolutionary data*, Bioinformatics, 14(10):68-73 (1998).

6.1 Trees vs networks

Here is (a) the unrooted neighbor-joining tree for 16S rRNA sequences (1355 bp) from ten species of *Neisseria* and (b) a splits graph computed from the same distance matrix:



(a) (b)

(Source: Eddie C. Holmes. Genomics, phylogenetics and epidemiology, Microbiology Today, 26:162-163 (1999).)

6.2 Representing distances using trees

Given a set X of taxa. For our purposes, a *phylogenetic tree* T is a tree such that:

- all leaves, and perhaps some of the internal nodes, too, are (multi-)labeled by elements

of X, such that each taxon appears exactly once, and

- every edge e has a weight d_e associated with it.

Given a set of taxa X and a distance matrix $\{d_{ab}\}$ (i.e., a *dissimilarity function* or *pseudo metric*) describing "evolutionary distances" between the different taxa, obtained in some way, e.g. Hamming or Jukes-Cantor distances.

Any distance-based tree building method attempts to represent a given distance matrix d as well as possible using a phylogenetic tree T, i.e. for any two taxa $a, b \in X$ we approximate $d_{ab} \approx \sum_{e \in P} d_e$, where P is the unique path of edges in T that connects the nodes with labels a and b.

6.3 Main goal

Given a distance matrix d, a tree building method such as neighbor-joining will compute a phylogenetic tree T for d, no matter how "untree-like" the distance matrix d may be.

(Recall that the four-point condition determines whether a given distance matrix d is *additive* or not, i.e. whether it has an exact representation as by a phylogenetic tree, or not.)

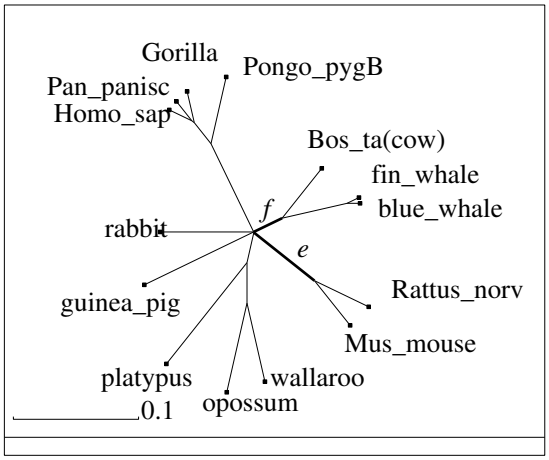
Our goal is to use more general graphs to represent distances, so-called *splits graphs*. As we will see, the graph will be a tree, whenever the given distances are tree-like (i.e., additive, or close to additive).

To reach this goal, we proceed indirectly by discussing sets of splits and introducing the notation of *weak-compatibility*.

Just as a set of compatible splits can be represented by a phylogenetic tree, we will see that a weakly-compatible set of splits can be represented by a splits graph.

6.4 Tree and splits

Here is an example of a phylogenetic tree T:



Each edge in T defines a split of the set of taxa X. For example, the edge labeled e separates rat and mouse from all other taxa, and the edge f separates cow, fin whale and blue whale from all others.

6.5 Tree distance and Σ-distance

Given a phylogenetic tree T with edge weights. We define the *tree distance* between two taxa a and b as

$$d_T(a,b) := \sum_{e \in P} d_e,$$

where P denotes the set of edges along the unique simple path from the node labeled a to the node labeled b.

We set $d_S := d_e$, if S is the split corresponding to e. We define the Σ-*distance* between two taxa a and b as

$$d_\Sigma(a,b) := \sum_{S \in \Sigma(a,b)} d_S,$$

where $\Sigma(a,b)$ is the set of all splits in Σ that separate a and b.

These definitions imply

$$d_T(a,b) = d_\Sigma(a,b)$$

for all taxa $a, b \in X$.

6.6 Weak compatibility

Compatibility is a requirement defined on any *two* splits. A relaxed concept is that of *weak compatibility*, which is a condition placed on any *three* splits.

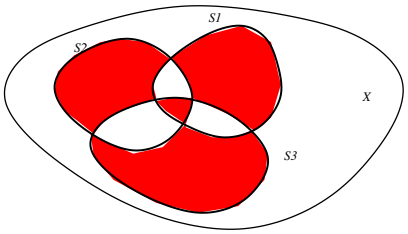
A triplet $S_1 = \{A_1, \bar{A}_1\}$, $S_2 = \{A_2, \bar{A}_2\}$ and $S_3 = \{A_3, \bar{A}_3\}$ is called *weakly compatible*, if at least one of the four intersections of

$$A_1 \cap A_2 \cap A_3,\ A_1 \cap \bar{A}_2 \cap \bar{A}_3,\ \bar{A}_1 \cap A_2 \cap \bar{A}_3,\ \text{or } \bar{A}_1 \cap \bar{A}_2 \cap A_3,$$

and of

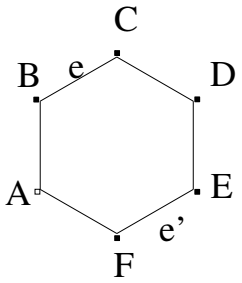
$$\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3,\ \bar{A}_1 \cap A_2 \cap A_3,\ A_1 \cap \bar{A}_2 \cap A_3,\ \text{or } A_1 \cap A_2 \cap \bar{A}_3,$$

is empty. This means that at least one shaded and one unshaded region of the following diagram must be empty:



Note that if any pair of the three splits is compatible, then all three are weakly-compatible: E.g., if for $S_1 = \{A_1, \bar{A}_1\}$ and $S_2 = \{A_2, \bar{A}_2\}$ we have $A_1 \cap A_2 = \emptyset$, then $A_1 \cap A_2 \cap A_3 = \emptyset$ and $A_1 \cap A_2 \cap \bar{A}_3 = \emptyset$.

On the other hand, it is possible that *every* pair of the three weakly-compatible splits is incompatible:



Here, a split of $X = \{A, B, C, D, E, F\}$ if given by each pair of parallel edges, e.g. edges e and e' define the split $\{\{A, B, F\}, \{C, D, E\}\}$.

6.7 Weak compatibility and splits graphs

As discussed above, any given set of splits Σ can be represented by a tree $T(\Sigma)$, if and only if Σ is compatible.

A weakly compatible split system $\mathcal{S}$ can be represented by a *splits graph* $G(\Sigma)$ that has the following properties:

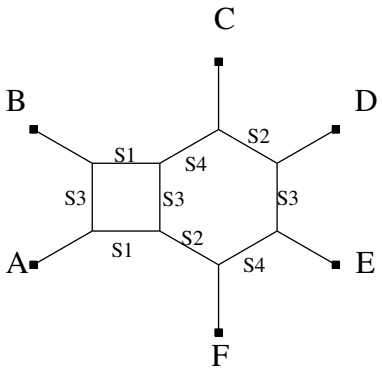
- all leaves (and, additionally, some internal nodes, perhaps) are multi-labeled by taxa so that each taxon appears exactly once,
- edges are labeled by splits such that each split appears at least once,
- deleting all edges labeled by any given split $S = \{A, \bar{A}\}$ produces precisely two components, one containing all nodes with labels in A and the other containing all nodes with labels in $\bar{A}$, and

- the graph is minimal with these properties.

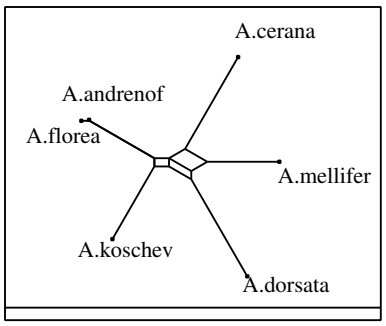
Given the following splits:

$$S_1 = \{\{A, B\}, \{C, D, E, F\}\}, \quad S_2 = \{\{A, B, C\}, \{D, E, F\}\},$$
$$S_3 = \{\{A, F, E\}, \{B, C, D\}\}, \quad S_4 = \{\{A, B, F\}, \{C, D, E\}\},$$
and all *trivial* splits A vs $B - F$, etc.

They can be represented as follows:



Here is another example of a splits graph:



This graph is based on DNA obtained from bees. It indicates that is there is some evidence that groups A.cerana and A.meillifer together, and conflicting evidence that groups A.mellifer with A.dorsata, for example.

6.8 Splits graphs and distances

Given a set of taxa X, a set of weakly compatible splits Σ of X and a value $d_S \geq 0$ for each split S.

As above, we define the Σ-distance between taxa a and b simply as $d_\Sigma(a, b) := \sum_{S \in \Sigma(a,b)} d_S$, where $\Sigma(a, b)$ is the set of all splits that separate a and b.

Assume we are given a corresponding splits graph G. In G, each split S is represented by a band of parallel edges and each such edge e has weight $d_e = d_S$.

Consider any two taxa $a, b \in X$. We define

$$d_G(a,b) := \min\{\sum_{e \in P} d_e \mid P \text{ is a simple path from } a \text{ to } b\}.$$

Lemma We have $d_\Sigma(a,b) = d_G(a,b)$ for all $a, b \in X$.

(Proof: need to show that a minimum path from a to b uses precisely one edge for every split that separates a and b.)

6.9 Two main questions

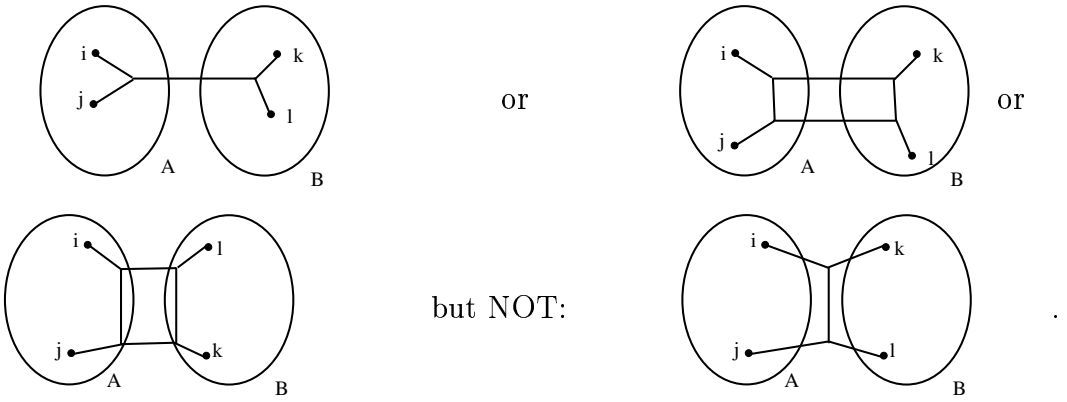
- First, given a set of taxa X and a distance matrix d. How do we compute a set of weakly compatible system of splits Σ and values d_S, such that $\sum_{S \in \Sigma(a,b)} d_S$ is a useful approximation of d_{ab}?
- Second, given a weakly compatible set of splits, how do we compute the corresponding splits graph?

6.10 Distance matrices and d-splits

Given a distance matrix d on X. We call a split $S = \{A, \bar{A}\}$ a *d-split*, if for all $i, j \in A$ and $k, l \in \bar{A}$ we have

$$d_{ij} + d_{kl} < \max(d_{ik} + d_{jl}, d_{il} + d_{jk}).$$

In other words, the metric induced by d on any four taxa $i, j \in A$, $k, l \in \bar{A}$, places i, j and k, l together as indicated here:

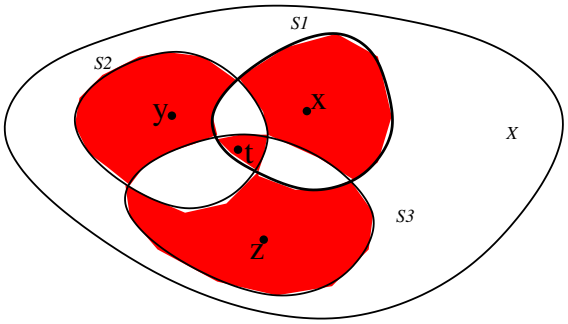


.

6.11 d-splits are weakly compatible

Lemma (Bandelt & Dress 1992) Let d be a distance matrix on X. Then the set of all d-splits is weakly compatible.

Proof Consider three d-splits $S_1 = \{A_1, \bar{A}_1\}$, $S_2 = \{A_2, \bar{A}_2\}$ and $S_3 = \{A_3, \bar{A}_3\}$ and *assume that they are not weakly-compatible*. Then there exist four taxa x, y, z, t contained in $A_1 \cap \bar{A}_2 \cap \bar{A}_3$, $\bar{A}_1 \cap A_2 \cap \bar{A}_3$, $\bar{A}_1 \cap \bar{A}_2 \cap A_3$ and $A_1 \cap A_2 \cap A_3$, respectively:



The definition of a d split implies the following three inequalities:

For S_1 : $d_{xt} + d_{yz} < \max(d_{xy} + d_{tz}, d_{xz} + d_{ty})$,
for S_2: $d_{yt} + d_{xz} < \max(d_{yx} + d_{tz}, d_{yz} + d_{tx})$, and
for S_3: $d_{zt} + d_{xy} < \max(d_{zx} + d_{ty}, d_{zy} + d_{tx})$.

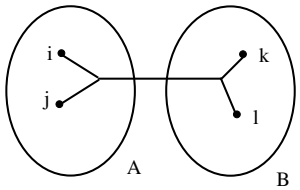
Note that these three inequalities cannot be fulfilled simutaneously, contradicting our assumptions and thus the three splits must be weakly-compatible. □.

6.12 The isolation index of a split

We give any d-split $S = \{A, \bar{A}\}$ a positive weight, namely the quantity

$$\alpha_{A,B} := \alpha^d_{A,B} := \frac{1}{2} \min_{i,j \in A,\, k,l \in B} \{\max(d_{ik} + d_{jl}, d_{il} + d_{jk}) - (d_{ij} + d_{kl})\},$$

called the *isolation index* of S.



We can easily modify this definition to apply to *any* split $S = \{A, \bar{A}\}$, whether d-split or not:

$$\alpha_{A,B} := \alpha^d_{A,B} := \frac{1}{2} \min_{i,j \in A,\, k,l \in B} \{\max(d_{ik} + d_{jl}, d_{il} + d_{jk}, d_{ij} + d_{kl}) - (d_{ij} + d_{kl}))\},$$

thus obtaining a value ≥ 0 that equals the previously defined isolation index, if S is a d-split, and 0, if not.

6.13 The split decomposition

For any split $S = \{A, \bar{A}\}$ of X, the *split metric* δ_S is given by

$$\delta_S(i,j) := \begin{cases} 0, & \text{if } i,j \in A \text{ or } i,j \in \bar{A}, \\ 1, & \text{else.} \end{cases}$$

Theorem (Bandelt & Dress) Any given distance matrix d on X possesses the following unique decomposition:

$$d_{ij} = \left(\sum_S \alpha_S \delta_S(i,j) \right) + d^0_{ij},$$

for all $i, j \in X$. Here, the sum runs over all possible splits S and the map $d^0 : X \times X \to \mathbb{R}^{\geq 0}$ is a (pseudo-)metric that does not admit any further splits with positive isolation index, i.e. there exist no d^0-splits.

Hence, we have $\sum_S \alpha_S \delta_S(i,j) \leq d_{ij}$ for any pair of taxa $i, j \in X$ and the Σ-distance α_S approximates d_{ij} from below.

One can prove that the number of d-splits is $\leq \binom{|X|}{2}$.

6.14 Computing the set of d-splits

Given a distance matrix d on X. The set of all d-splits can be computed iteratively in $O(n^6)$ steps:

Algorithm
Input: Distance matrix d, taxon set $X = \{x_1, x_2, \ldots, x_n\}$.
Output: Set $\Sigma = \Sigma_n$ of all d-splits

Initialization: $\Sigma_0 := \emptyset$

for each $k = 1, 2, \ldots, n$ do:
    Set $\Sigma_k := \emptyset$
    for each split $S = \{A, \bar{A}\} \in \Sigma_{k-1}$:
        if $\{A \cup \{x_k\}, \bar{A}\}$ has positive isolation index **then**
            Add $\{A \cup \{x_k\}, \bar{A}\}$ to Σ_k
        if $\{A, \bar{A} \cup \{x_k\}\}$ has positive isolation index **then**
            Add $\{A, \bar{A} \cup \{x_k\}\}$ to Σ_k
    If $\{\{x_1, x_2, \ldots, x_{k-1}\}, \{x_k\}\}$ has positive isolation index **then**
        Add $\{\{x_1, x_2, \ldots, x_{k-1}\}, \{x_k\}\}$ to Σ_k.
end

Lemma This algorithm computes all d-splits.

Proof First note that in the k-iteration of the algorithm the new *partial* trivial split $\{\{x_1, \ldots, x_{k-1}\}, \{x_k\}\}$ is evaluated and then added to the current set of splits, if $\alpha_{\{\{x_1,\ldots,x_{k-1}\},\{x_k\}\}} > 0$. Additionally, the algorithm attempts to extend all existing partial splits by adding x_k to the one side, or the other side, of them. By definition of the isolation index as the *minimum* of certain sums involving quartets of taxa, adding a taxon to either side of a partial split can only decrease the isolation index. Hence, any split of X is obtainable as a partial trivial split for some k, followed by successive addition of the remaining taxa to the split. □

6.15 Computing the splits graph

Given a compatible system of splits, it is easy to construct the corresponding tree and to compute coordinates for the tree.

The problem of computing a splits graph for a given set of weakly compatible splits is more difficult. In practice, one distinguishes between *circular* split systems, which correspond to planar split graphs, and non-circular ones. A nice algorithm exists for circular split systems that produces a planar graph.

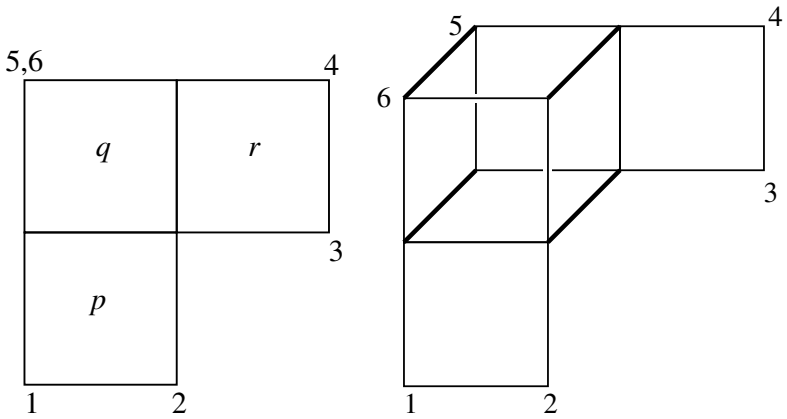
Here we discuss the *convex hull* approach that applies to any set of weakly compatible splits, whether circular or not. This method is easy to describe. Its main draw-back is that it usually produces redundant nodes and edges and so the resulting graph is not always *minimal* in the sense postulated above.

6.16 Convex-hull construction method

Given a splits graph G. For a given set of taxa $A \subset X$, let G_A denote the set of all nodes labeled by taxa in A. The *convex hull* $\overline{G_A}$ of G_A is obtained by first setting $\overline{G_A} = G_A$ and then repeatedly adding any node v to $\overline{G_A}$, if there exist two nodes a, b already in $\overline{G_A}$ such that $d_G(a, v) + d_G(v, b) \leq d_G(a, b)$.

Given a weakly compatible set of splits $\Sigma = \{S_1, S_2, \ldots, S_k\}$. The convex-hull construction method constructs a graph by adding one split at time. For each split, the convex hull for both sides of the split is computed. The intersection of the two is duplicated, one copy is connected to one side of the graph corresponding to one side of the split, the other to the other and then the two duplicated subgraphs are connected by a set of new edges that represent the new split.

Given the graph shown in (a). How do we add the split $S = \{\{1, 2, 6\}, \{3, 4, 5\}\}$ to it? The convex hull of $A = \{1, 2, 6\}$ consists of all nodes in squares p and q, and the convex hull of $\bar{A}$ of all nodes in q and r. The intersection H of both is q.



The graph (b) is obtained by duplicating H as described in the algorithm.

Assume that G is the graph constructed for splits $S_1 \ldots, S_i$. To add the next split $S_{i+1} = \{A, \bar{A}\}$:

- Determine the two convex hulls $\overline{G_A}$ and $\overline{G_{\bar{A}}}$.
- Let $H := \overline{G_A} \cap \overline{G_{\bar{A}}}$ denote their intersection.
- For each node $v \in H$, produce two new nodes v^+ and v^- and connect them by an edge labeled S_{i+1}.
- If $v \in H$ is labeled by a taxon $x \in A$, or $x \in \bar{A}$, then attach this label to node v^+, or v^-, respectively.
- Connect any two nodes v^+ and w^+, and v^- and w^-, respectively, by an edge, if v and w are connected by an edge in G.
- If $v \in H$ is connected to some node $w \in \overline{G_A} \setminus \overline{G_{\bar{A}}}$, then connect v^+ and w by an edge.
- If $v \in H$ is connected to some node $w \in \overline{G_{\bar{A}}} \setminus \overline{G_A}$, then connect v^- and w by an edge.
- Delete H.

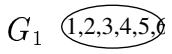
6.17 Computing the splits graph

Given the following set Σ of splits:

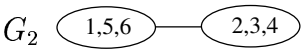
$$
\begin{array}{ll}
S_1 = & \{\{1,5,6\} \quad ,\{2,3,4\} \quad \} \\
S_2 = & \{\{1,2,3\} \quad ,\{4,5,6\} \quad \} \\
S_3 = & \{\{1,2,5,6\},\{3,4\} \quad \} \\
S_4 = & \{\{1,2\} \quad ,\{3,4,5,6\}\} \\
S_5 = & \{\{1,6\} \quad ,\{3,4,5,6\}\}
\end{array}
$$

We will demonstrate how to generate G_Σ.

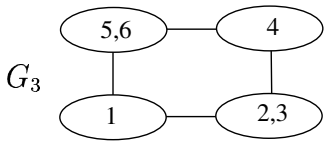
Initially, start with a single node labeled by all of $X = \{1, 2, 3, 4, 5, 6\}$:



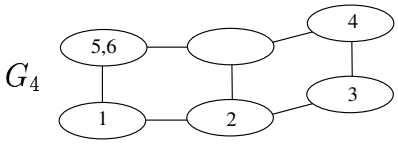
Then add the first split S_1. Note that H consists of the single node present in G_1:



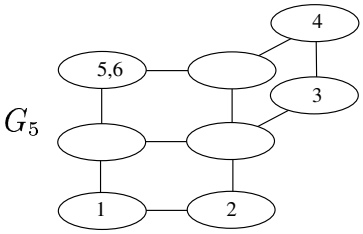
Add the second split $S_2 = \{\{1, 2, 3\}, \{4, 5, 6\}\}$. Note that H consists of both nodes in G_2:



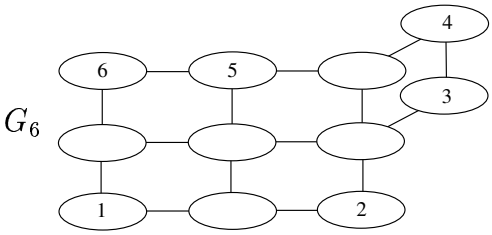
Add the third split $S_3 = \{\{1, 2, 5, 6\}, \{3, 4\}\}$. Note that H consists of the two nodes labeled $2, 3$ and 4 in G_3:



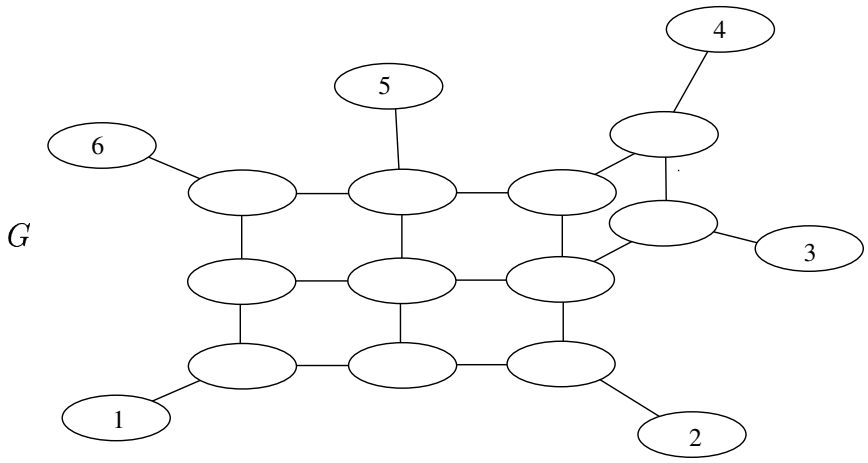
Add the fourth split $S_4 = \{\{1, 2\}, \{3, 4, 5, 6\}\}$. Note that H consists of the two nodes labeled 1 and 2 in G_4:



Add the fifth split $S_5 = \{\{1, 6\}, \{2, 3, 4, 5\}\}$. Note that H consists of the two nodes labeled 1 and $5, 6$, plus the node lying between these two in G_5:

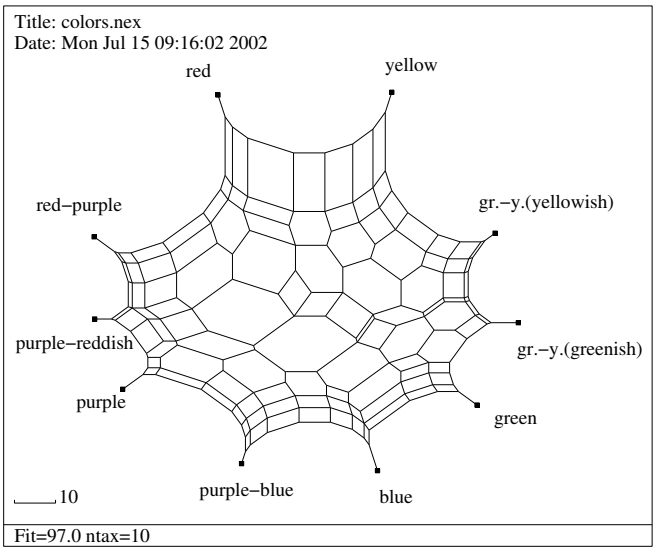


Finally, add all trivial splits to G_6 to obtain the final graph G:



6.18 Example of splits graph

The distance matrix for the following example was produced in a psychology experiment in which people where asked to estimate the distance between different colors:



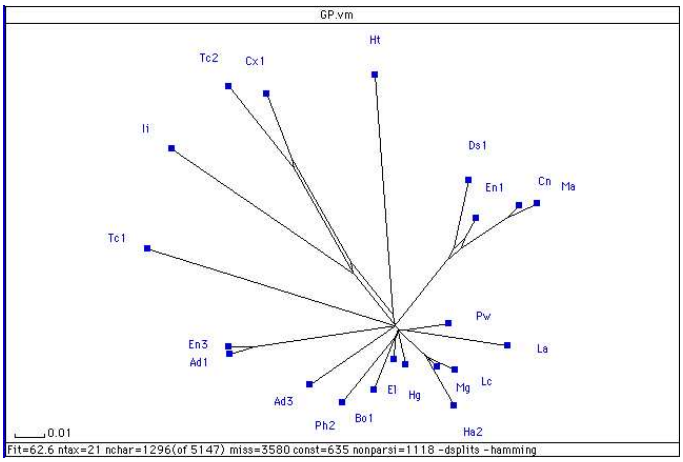
6.19 Manuscript analysis

The Canterbury Tales text was left unfinished in 1400 and survives in some 88 versions dating before 1500. Many questions remain open, including *which* version is closest to the original one written by Geoffrey Chaucer?

To address such questions, the spelling of a part of the manuscripts was transcribed and then compared, giving rise to "sequence data" that can be analyzed using methods such as split decomposition.

The main use of such programs in this context is seen as a way to suggest possible rela-

tionships that can then be further investigated, confirmed or rejected, using conventional tools.



For more details of the application of phylogenetic methods to the Canterbury Tales, see `http://www.cta.dmu.ac.uk/projects/ctp/desc2.html`.