Minimum Common Supergraphs and Haplotype Networks

Anthony Labarre[1]
alabarre@ulb.ac.be

Université libre de Bruxelles (U.L.B.)

December 21st, 2007

Future Directions in Phylogenetic Methods and Models

[1]Funded by the "Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture" (F.R.I.A.).

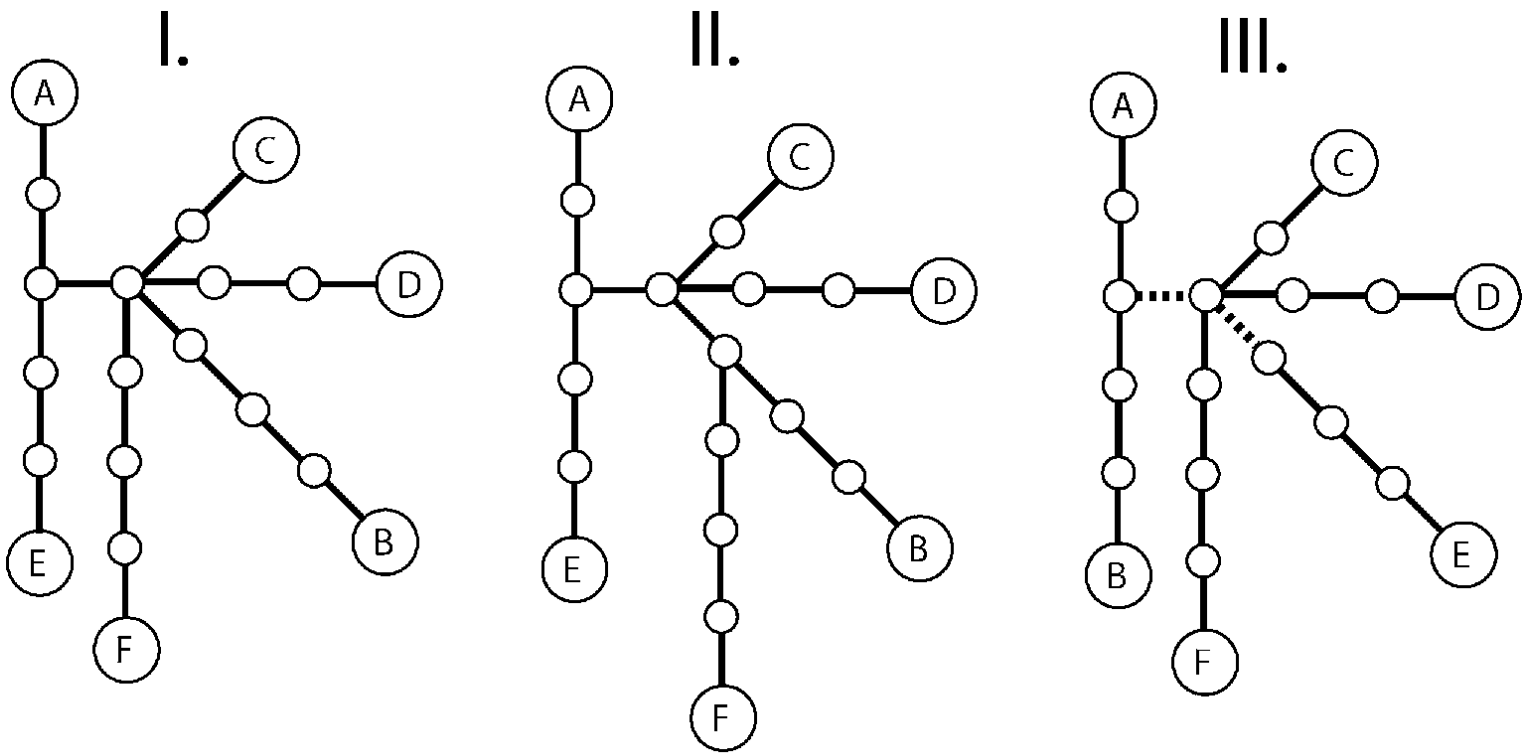
Motivation: haplotype networks

- Reconstruction of haplotype networks: Given a set of genes, represent all shortest evolutionary relations between those genes
- Existing methods build a graph, then try to reduce its size:
  - Minimum Spanning Network [Excoffier and Smouse, 1994]
  - Statistical Parsimony Network [Templeton et al., 1992]
  - Median Joining Network [Bandelt et al., 1999]
  - ...

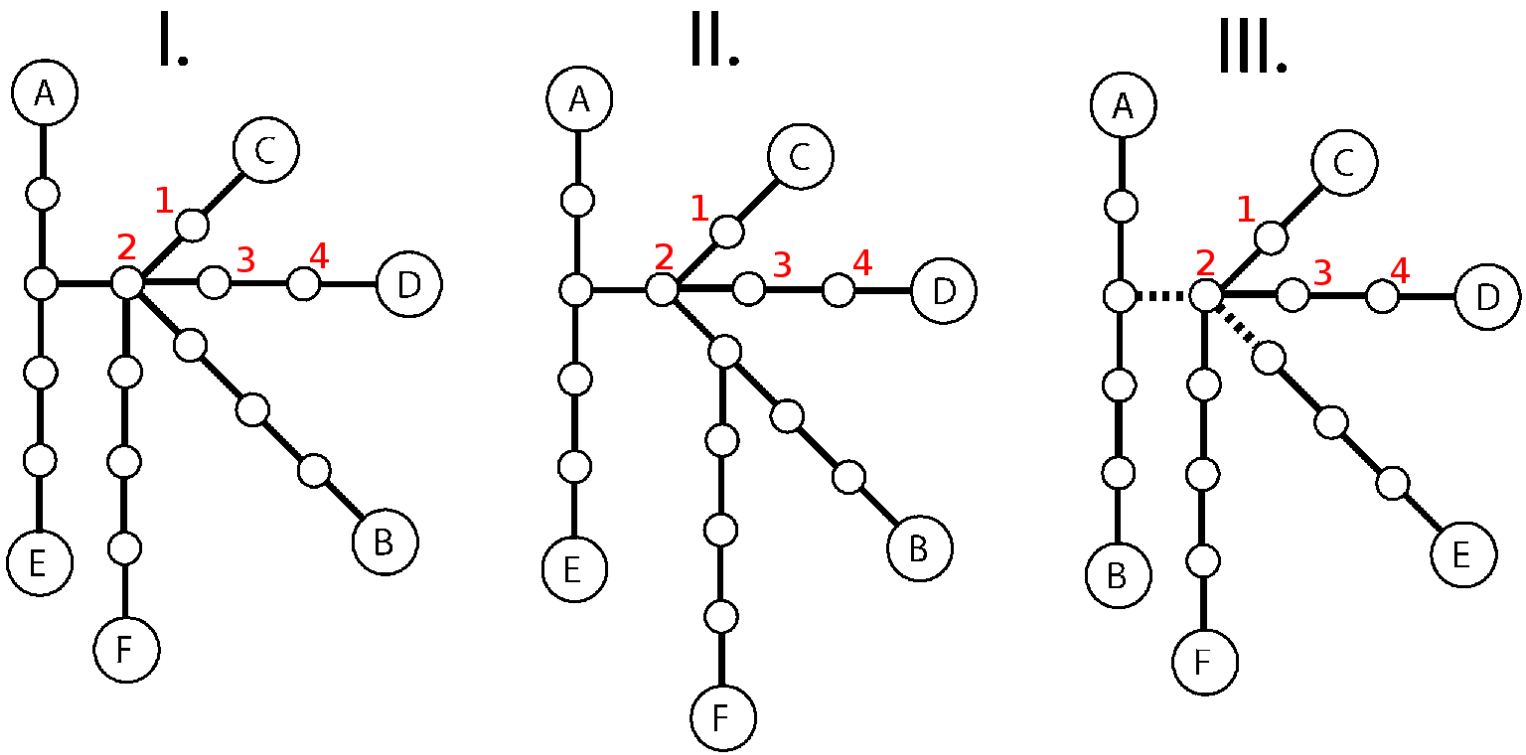
Cassens et al.'s method

- Method proposed by [Cassens et al., 2005];
- The method has two steps:
  1. generating all equally most parsimonious trees;
  2. "merging" those trees into a graph with as few vertices and edges as possible, by identifying common paths;
- We will concentrate on step 2;
- Remark: those trees are
  - undirected,
  - unrooted,
  - labels are not restricted to leaves,
  - and there is no constraint on the degree of internal vertices.

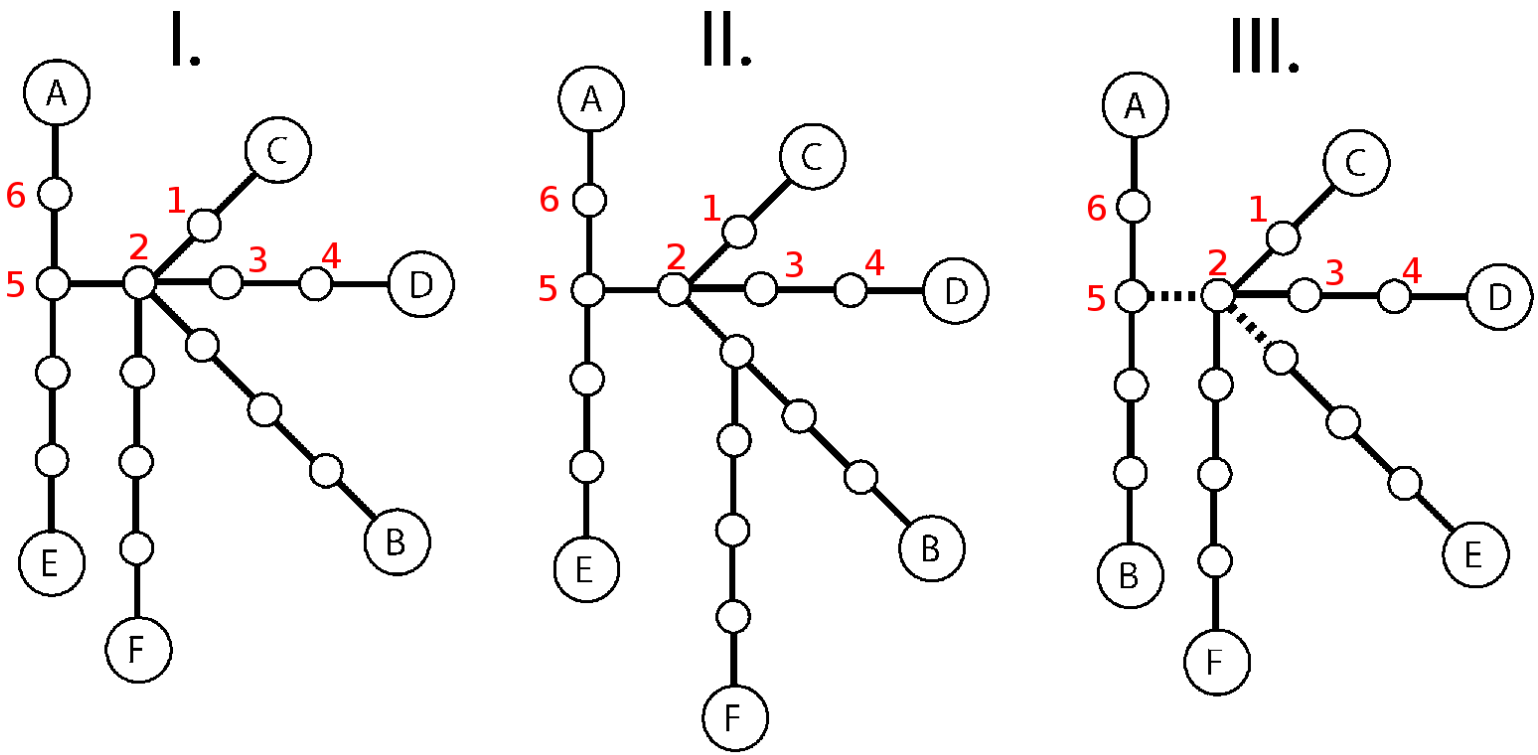
Cassens et al.'s method: example



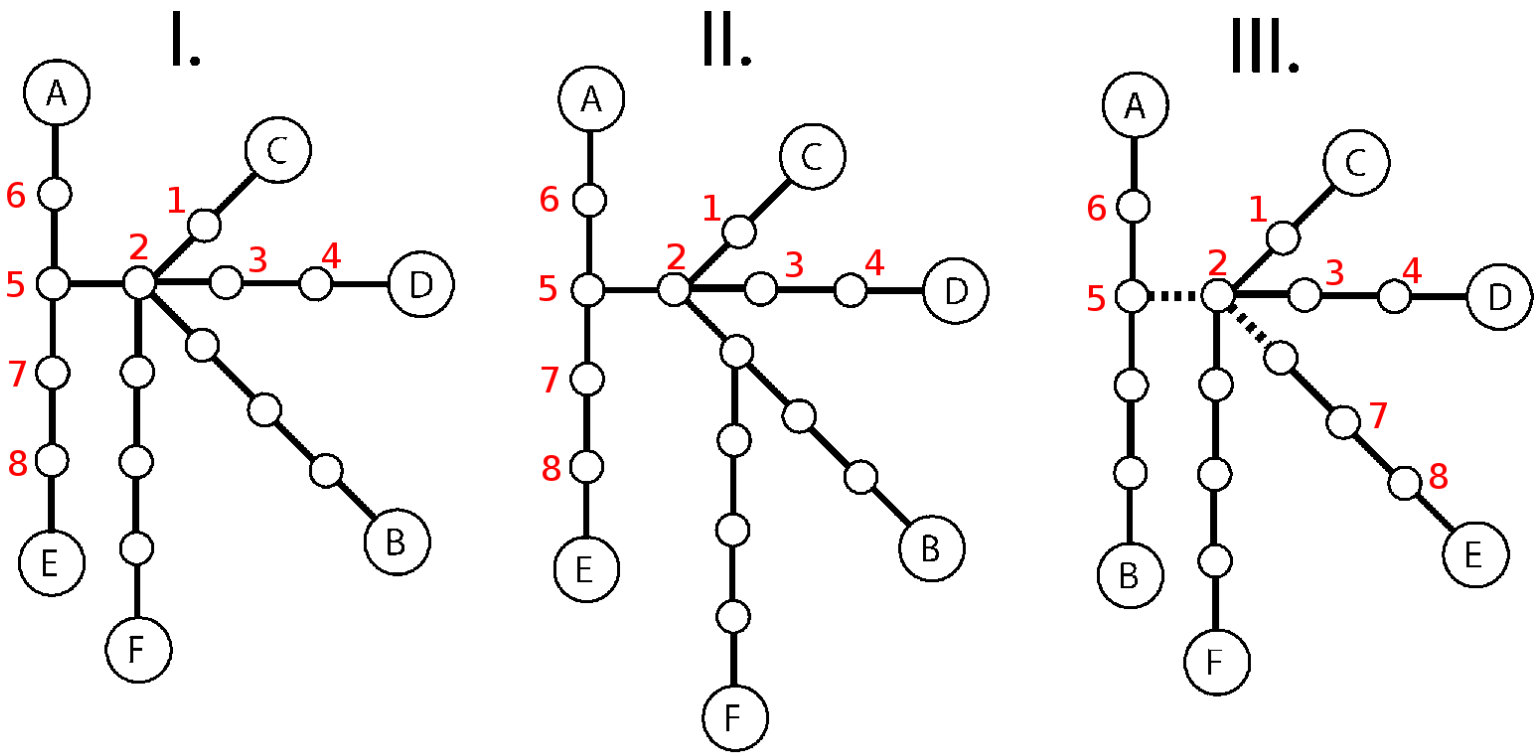
Cassens et al.'s method: example



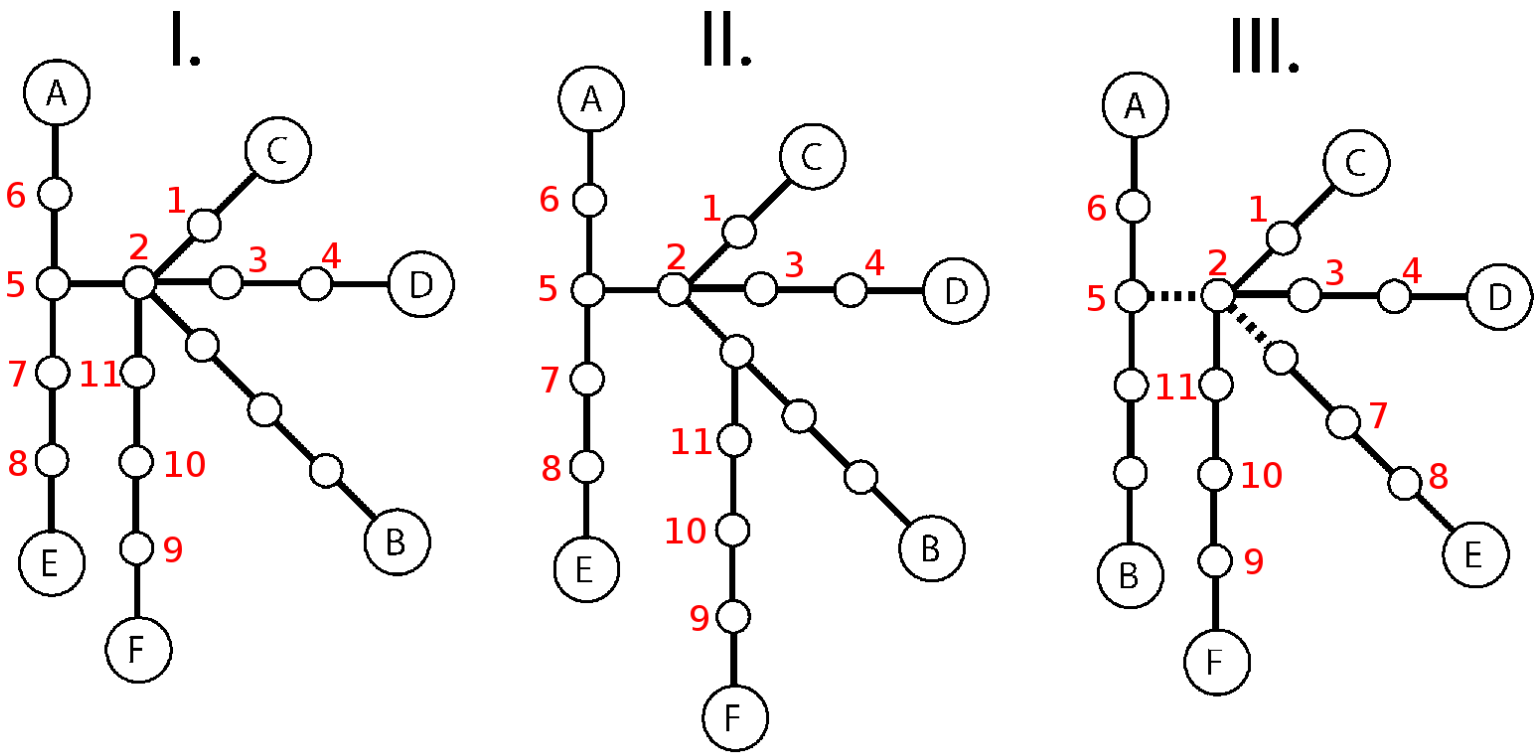
Cassens et al.'s method: example



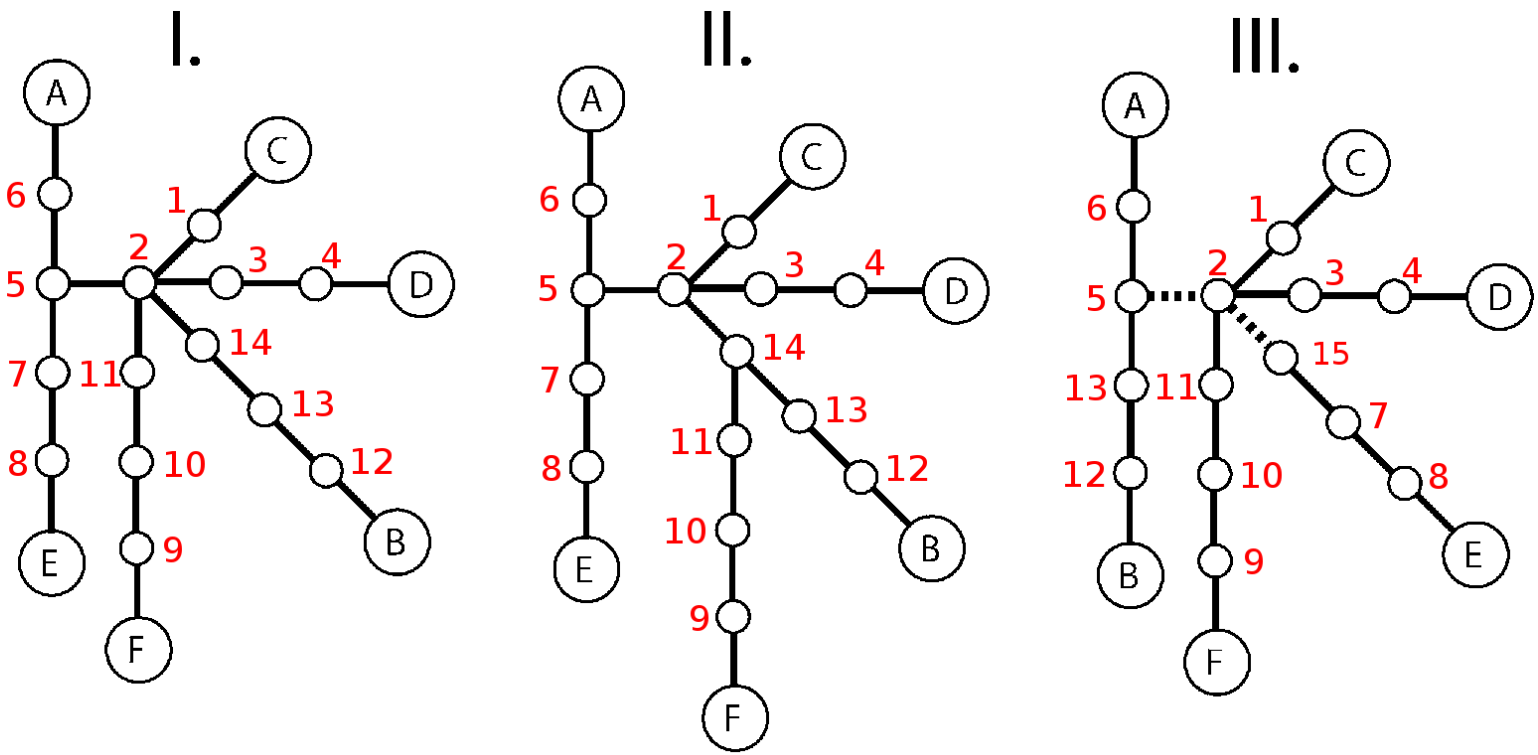
Cassens et al.'s method: example



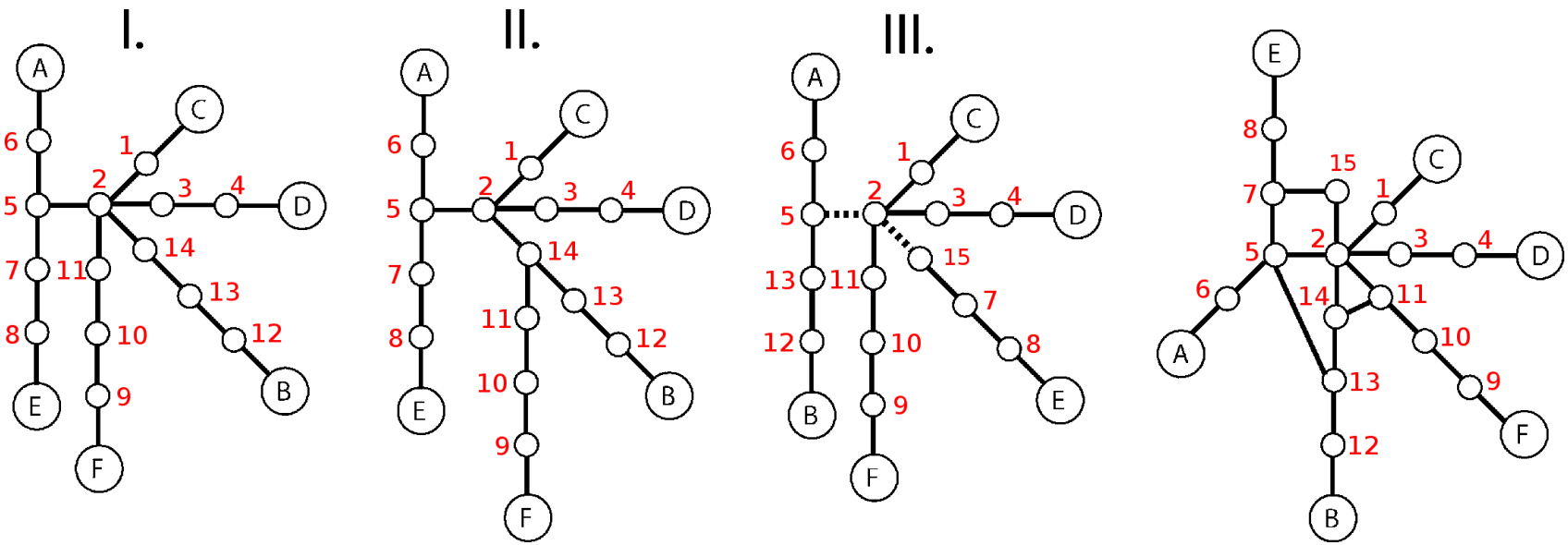
Cassens et al.'s method: example



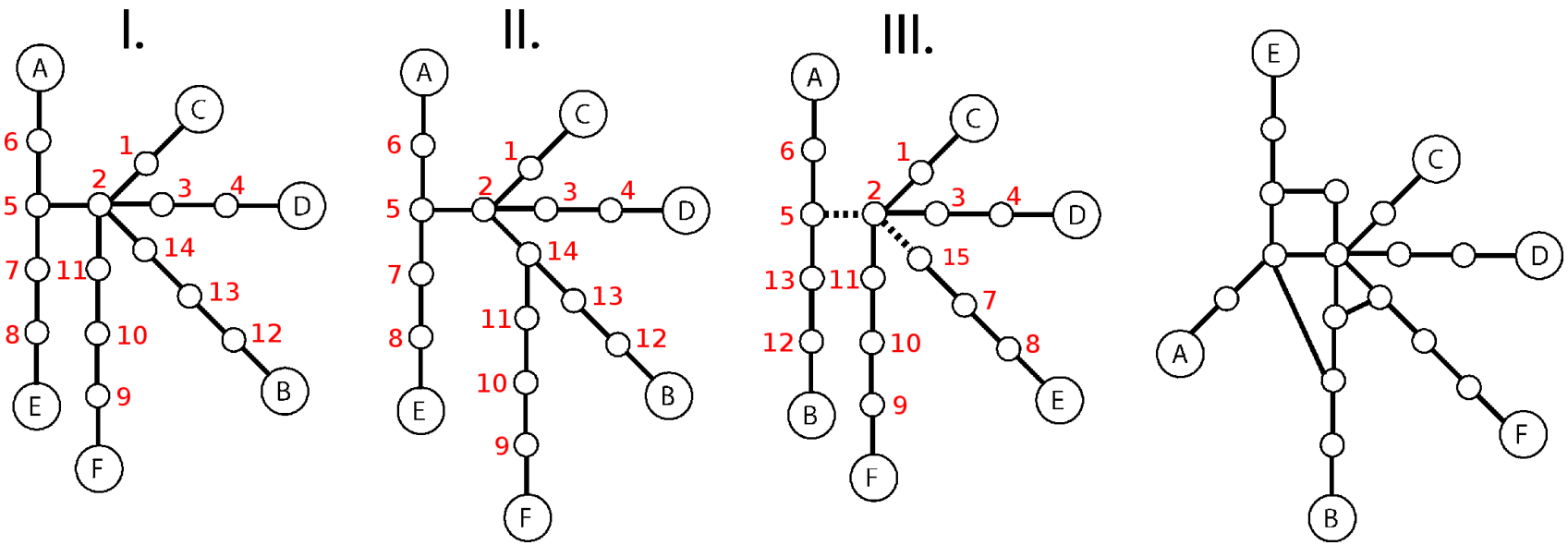
Cassens et al.'s method: example



Cassens et al.'s method: example



Cassens et al.'s method: example



Problems of the method

- Not guaranteed to be optimal, or to be an approximation within some constant factor;
- Order-dependent;
- Heuristic with an unclear objective function;

This work proposes and studies one possible model for their method.

Definitions

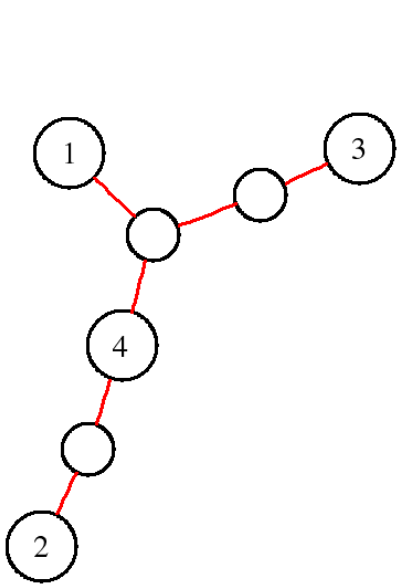
- (n, k)-*graph* G: graph on n vertices, k of which are labelled
  - $V_l(G) =$ labelled vertices, which includes all degree 1 vertices;
  - $V_u(G) =$ unlabelled vertices.
  - the label set is $\{1, 2, \ldots, k\}$;

Definitions

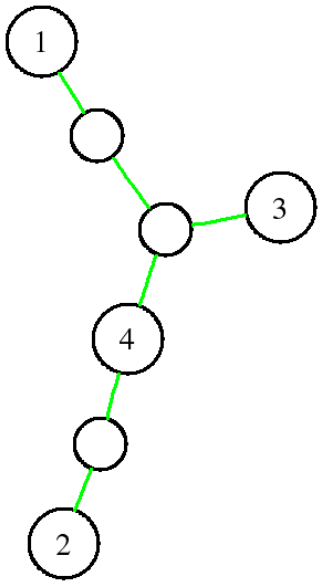
- (n, k)-*graph* G: graph on n vertices, k of which are labelled
  - $V_l(G)$ = labelled vertices, which includes all degree 1 vertices;
  - $V_u(G)$ = unlabelled vertices.
  - the label set is $\{1, 2, \ldots, k\}$;
- *Common supergraph* of (n, k)-graphs $G_1, G_2, \ldots, G_t$: (n, k)-graph G such that each G_i can be reconstructed by removing edges from $E(G)$. It is *minimum* if there is no other graph G' with $|E(G')| < |E(G)|$ that shares this property.

(n, k)-graphs and common supergraphs

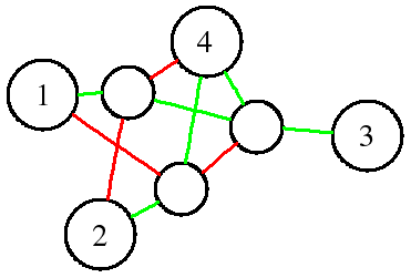
Example



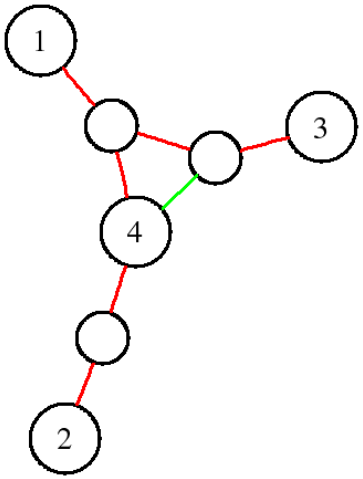
T_1



T_2



G_1 (10 edges)



G_2 (7 edges)

Formal statement of the problem

MINIMUM COMMON SUPERGRAPH OF PARTIALLY LABELLED TREES (MCS-PLT):

Input: (n, k)-trees T_1, T_2, ..., T_t on the same label set.

Problem: find a minimum common supergraph of T_1, T_2, ..., T_t.

Preliminary results

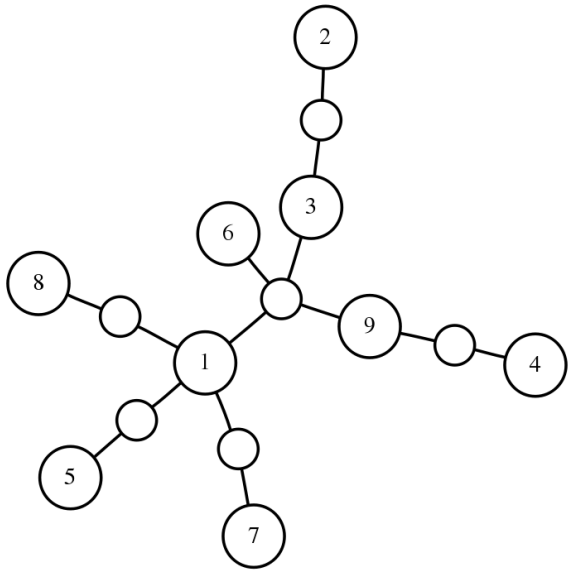
- Polynomial-time algorithm for two "restricted" trees;
- Remains polynomial if only one tree is "restricted";
- Exact exponential-time algorithm for two arbitrary trees;
- Those results extend directly to two *graphs*;

Restricted graphs

Definition

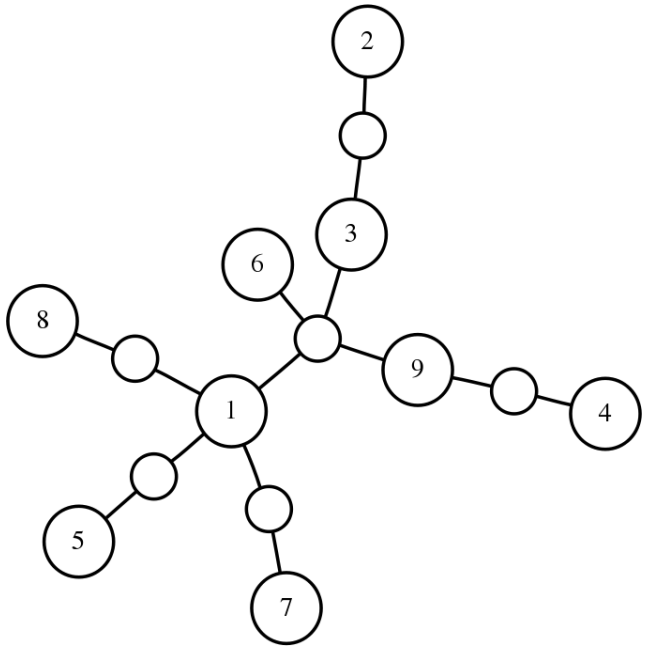
An (n, k)-graph is *restricted* if every unlabelled vertex it contains has labelled neighbours only.

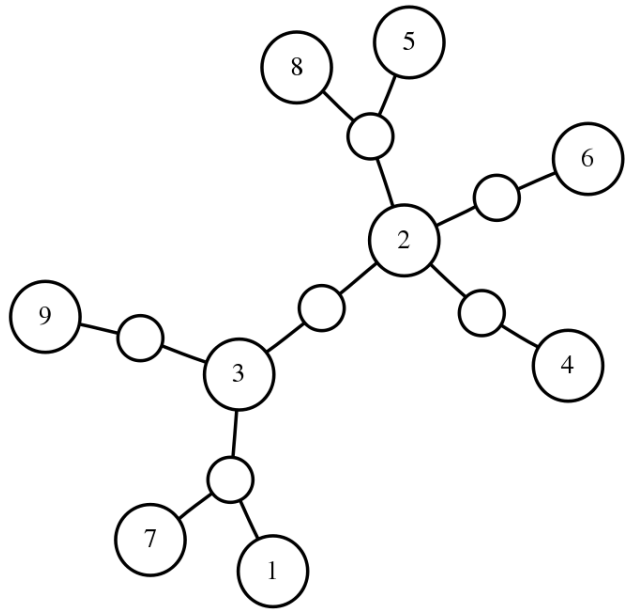
Example



Merging restricted (n, k)-trees

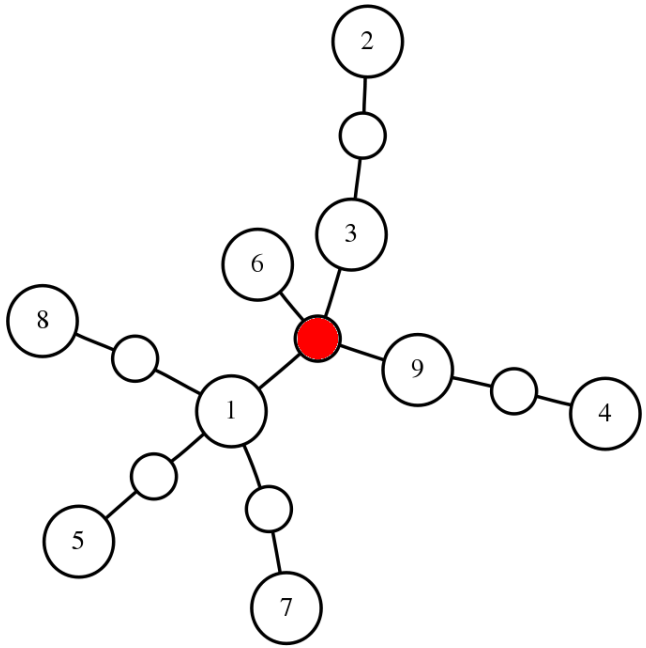
Example

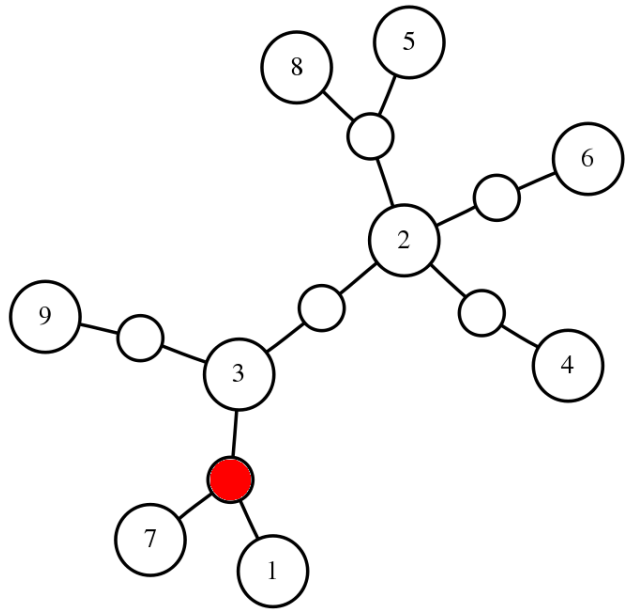




Merging restricted (n, k)-trees

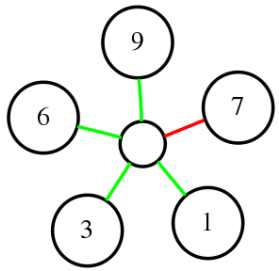
Example





Merging restricted (n, k)-trees

Example



- When pairing up $a \in V_u(T_1)$ and $b \in V_u(T_2)$, the number of edges in the corresponding subgraph of the resulting common supergraph is exactly

$$f(a, b) = \left| labels(N_I^{T_1}(a)) \cup labels(N_I^{T_2}(b)) \right|$$

Merging restricted (n, k)-trees

- Those trees can be optimally merged using the linear assignement problem:
  1. Build the complete bipartite graph B with vertex classes $V_u(T_1)$, $V_u(T_2)$
  2. Assign weights according to f
  3. Find a perfect matching of minimum weight of B
- This is well-known to be solvable in $O((n-k)^3)$ time [Schrijver, 2003]

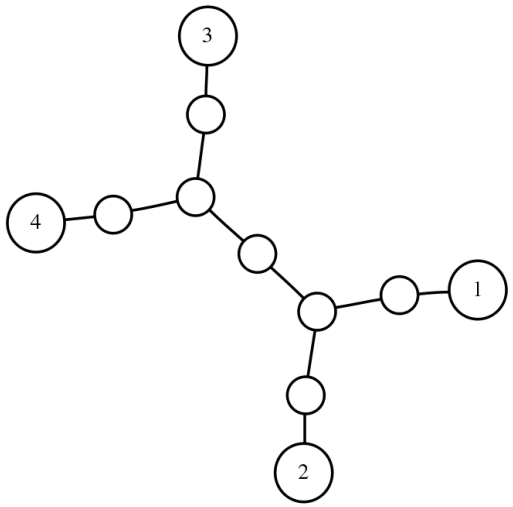
What if only one tree is restricted?

- If T_1 is restricted and T_2 is not, then pairing up $a \in V_u(T_1)$ and $b \in V_u(T_2)$ will result in a subgraph with $f(a, b) + |N_u^{T_2}(b)|$ edges:
- In other words, we "have no power" over edges connecting unlabelled vertices
- Hence the $O((n-k)^3)$ LAP approach still works in that case

Restricting (n, k)-trees

- Any (n, k)-tree can be transformed into a restricted $(n, k + p)$-tree by arbitrarily labelling p vertices:

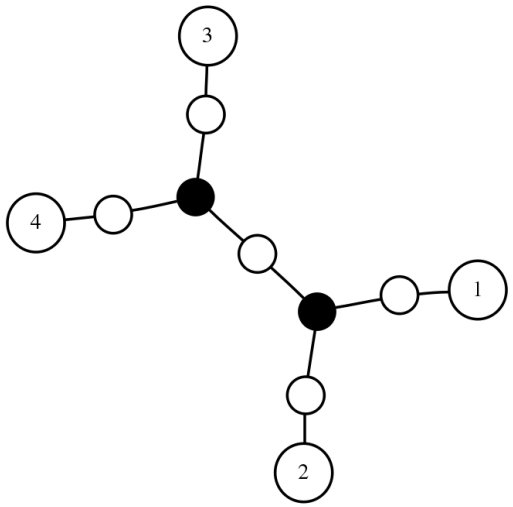
Example



Restricting (n, k)-trees

- Any (n, k)-tree can be transformed into a restricted $(n, k + p)$-tree by arbitrarily labelling p vertices:

Example



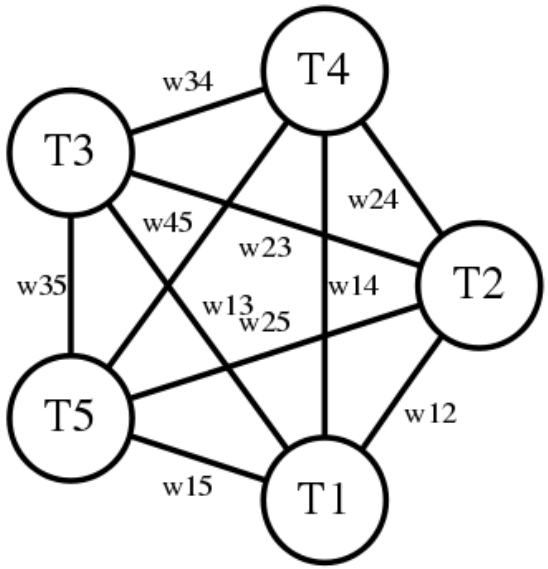
- This is equivalent to finding a minimum vertex cover on a forest, which is achievable in linear time;

Merging two arbitrary trees

- Sketch of the algorithm:
  1. "restrict" one of the trees (say T_1) by artificially labelling p vertices;
  2. try all possible labellings of p vertices in T_2;
  3. keep the best solution over all those pairs of trees;
- Every pair of trees is optimally merged in polynomial time, since at least one tree is restricted;
- The running time is $O((n-k)^{p+3})$, and can be improved through branch-and-bound and preprocessing.

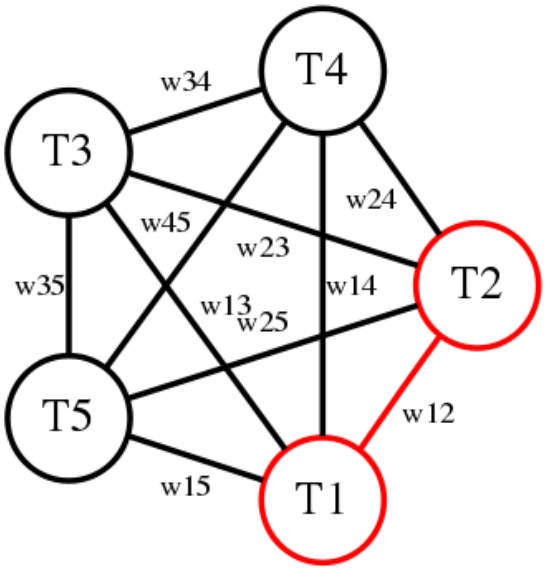
More than two trees

- What if we have more than two trees (which is usually the case)?
- Possible approach: merge trees in a "pairwise fashion"



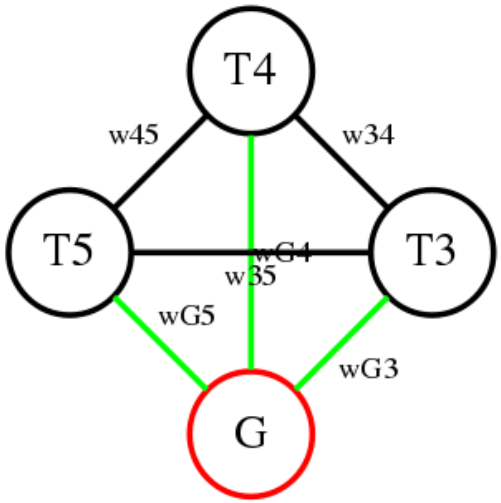
More than two trees

- What if we have more than two trees (which is usually the case)?
- Possible approach: merge trees in a "pairwise fashion"



More than two trees

- What if we have more than two trees (which is usually the case)?
- Possible approach: merge trees in a "pairwise fashion"



- Algorithms extend directly to *graphs*, so we can merge graphs and trees
- Order guaranteeing a satisfying solution?

Thank you!


Bandelt, H.-J., Forster, P., and Rohl, A. (1999).
Median-joining networks for inferring intraspecific phylogenies.
Molecular Biology and Evolution, 16(1):37–48.


Cassens, I., Mardulyn, P., and Milinkovitch, M. C. (2005).
Evaluating intraspecific "network" construction methods using simulated sequence data: Do existing algorithms outperform the global maximum parsimony approach?
Systematic Biology, 54(3):363–372.


Excoffier, L. and Smouse, P. E. (1994).
Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony.
Genetics, 136:343–359.


Schrijver, A. (2003).
Combinatorial optimization. Polyhedra and efficiency. Vol. A, volume 24 of *Algorithms and Combinatorics*, chapter 17, pages 285–300.
Springer-Verlag, Berlin.


Templeton, A. R., Crandall, K. A., and Sing, C. F. (1992).
A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data.
Genetics, 132:619–633.