Practical use of combinatorial methods for phylogenetic network reconstruction

Philippe Gambette
Outline

• Phylogenetic networks

• Motivations for the combinatorial reconstruction approach

• Combinatorial reconstruction methods

• Practical use

• Illustrations

• Perspectives
Outline

• Phylogenetic networks
  • Motivations for the combinatorial reconstruction approach
  • Combinatorial reconstruction methods
  • Practical use
  • Illustrations
  • Perspectives
Phylogenetic trees

Phylogenetic tree of a species set

species tree $S$

“tokogeny”
Genetic material transfers between coexisting species:

- horizontal gene transfer
- hybridization
Genetic material transfers between coexisting species:

- horizontal gene transfer
- hybridization
Genetic material transfers between coexisting species:

- horizontal gene transfer
- hybridization

- gene G1
- gene G2
- incompatible gene trees
Phylogenetic networks

**Phylogenetic network**: network representing evolution data

- **explicit** phylogenetic networks to **model** evolution

- **abstract** phylogenetic network to **classify, visualize** data

---

**Networks**:
- **Simplistic** level-2 network
- **Dendroscope** galled network
- **SplitsTree** split network
- **HorizStory** synthesis diagram
- **Network** median network
- **TCS** minimum covering network
Phylogenetic network software

Who is Who in Phylogenetic Networks, Articles, Authors & Programs

Based on BibAdmin by Sergiu Chelcea + tag clouds, date histogram, journal lists, co-author graphs, keyword definitions.

http://www.atgc-montpellier.fr/phylnet
Explicit phylogenetic networks

**Phylogenetic network**: network representing evolution data

- **explicit** phylogenetic networks

**evolution model**
Explicit phylogenetic networks

Rooted explicit phylogenetic network: tree-like parts + **blobs**.

vertices with more than one parent: **reticulations**
Plan

• Phylogenetic networks

• **Motivations for the combinatorial reconstruction approach**

• Combinatorial reconstruction methods

• Practical use

• Illustrations

• Perspectives
Phylogenetic network reconstruction

distance methods

parsimony methods
Hein 1990 - Kececioglu & Gusfield 1994 - Jin, Nakhleh, Snir, Tuller 2009

likelihood methods
Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009

gene sequences

espèce 1: AATTGCAG TAGCCAAAAT
espèce 2: ACCTGCAG TAGACCAAT
espèce 3: GCTTGCCG TAGACAAGAAT
espèce 4: ATTTGCAG AAGACCAAAT
espèce 5: TAGACCAAGAAT
espèce 6: ACTTGCA TAGCACAAT
espèce 7: ACCTGGTG TAAAAT

network N
Phylogenetic network reconstruction

Problem: usually slow, lots of sequences available.

Gene sequences:

- espèce 1: AATTGCAG TAGCCAAAAAT
- espèce 2: ACCTGCAG TAGACCAAT
- espèce 3: GCTTGCCG TAGACAAGAAT
- espèce 4: ATTTGCAG AAGACCAAAT
- espèce 5: TAGACAAAGAAT
- espèce 6: ACTTGCACT TAGACACAAAAT
- espèce 7: ACCTGGTG TAAAAT

Network $N$

distance methods

parsimony methods

likelihood methods
- Snir & Tuller 2009 - Jin, Nakhleh, Snir, Tuller 2009 - Velasco & Sober 2009
Phylogenetic network reconstruction

Reconstruction of one tree for each gene present in several species
Guindon & Gascuel, SB, 2003

{gene sequences}

HOGENOM database
Dufayard, Duret, Penel, Gouy, Rechenmann & Perrière, BioInf, 2005

Tree consensus or reconciliation

optimal network $N$
Phylogenetic network reconstruction

Reconstruction of one tree for each gene present in several species

Tree consensus or reconciliation

Problem: tree reconciliation is difficult even for 2 trees
(NP-complete for 2 trees with minimum reticulation number)
Problem:
Reconstructing the **supernetwork** of a set of trees is **hard**.

Idea:
reconstruct a network containing all:

![Diagram of triplets and clusters](image)
Problem:
Reconstructing the *supernetwork* of a set of trees is **hard**.

Idea:
reconstruct a network containing all:

![Diagram of triplets and clusters](image)

of the input trees?
Softwired clusters

“softwired” cluster: cluster of a tree contained in the network

Tree-like model of gene transmission:
each gene comes from a single parent
Softwired clusters

“softwired” cluster: cluster of a tree contained in the network

Tree-like model of gene transmission:
each gene comes from a single parent
Idea:
change the type of data to process

\[ N = N' = N''? \]
Combinatorial phylogenetic network reconstruction

Idea:
change the type of data to process

\[ \{ \text{trees} \} \]

\[ \{ \text{clusters} \} \]

\[ \{ \text{triplets} \} \]

\[ \{ N \} \subseteq \{ N' \} \subseteq \{ N'' \} \]
Plan

• Phylogenetic networks
• Motivations for the combinatorial reconstruction approach
• Combinatorial reconstruction methods
• Practical use
• Illustrations
• Perspectives
Reconstruction from softwired clusters

Fast exact **galled network** reconstruction method from **softwired clusters**

Huson, Rupp, Berry, Gambette & Paul, ISMB 2009

Step 1- Solve cluster conflicts by **deleting taxa**, reconstruct tree on remaining taxa

MAXIMUM COMPATIBLE SUBSET

Step 2- Attach taxa involved in conflicts to the tree with the **minimum number of arcs**: MINIMUM ATTACHMENT

N’
galled network

http://www.dendroscope.org
Reconstruction from softwired clusters

Exact method for **level $k$ network** reconstruction from **softwired clusters**

Less reticulations, but slower for level $> 2$.

**level** = maximum number of reticulations per blob.

http://www.dendroscope.org
Reconstruction from triplets

Exact methods to reconstruct level-1 and 2 networks (if there exist any) from a dense triplet set

Jansson, Nguyen & Sung, SODA’05: $O(n^3)$ pour niveau 1,
van Iersel, Kelk & al, RECOMB’08: $O(n^8)$ pour niveau 2,
To & Habib, CPM’09: $O(n^{5k+4})$ pour niveau $k$

$T$ dense triplet set =
On any subset of 3 leaves, $T$ contains at least one triplet

Program Simplistic

http://homepages.cwi.nl/~kelk/simplistic.html

Yeast phylogenetic network -
Van Iersel et al.: Constructing level-2 phylogenetic networks from triplets.
RECOMB 2008
Reconstruction from triplets

{trees} → Fast heuristic method to reconstruct a **level-1 network** containing most of the input triplets

Program Lev1athan

{triplets} →  

$N'$ level-1 network

Phylogenetic network built from triplets extracted from 2 trees of HIV-1 strains  
Huber, van Iersel, Kelk & Suchecki  
A practical algorithm for reconstructing level-1 phylogenetic networks  
TCBB, 2011

Huber, van Iersel, Kelk & Suchecki, TCBB, 2011

http://homepages.cwi.nl/~kelk/lev1athan/
Outline

• Phylogenetic networks
• Motivations for the combinatorial reconstruction approach
• Combinatorial reconstruction methods
• Practical use
• Illustrations
• Perspectives
Ambiguity of the results even with complete and correct data

Many distinct minimal networks have exactly the same set of contained trees, triplets, and clusters.

Characterization for level-1 networks:
The only ambiguous cases have such blobs (< 5 vertices)

Gambette & Huber, 2011
Ambiguity of the results even with complete and correct data.

Many distinct minimal networks have exactly the same set of contained trees, triplets, and clusters.

Characterization for level-1 networks:
The only ambiguous cases have such blobs (< 5 vertices).
Ambiguity of the results even with complete and correct data.

Many distinct minimal networks have exactly the same set of contained trees, triplets, and clusters.

Characterization for level-1 networks:
The only ambiguous cases have such blobs (< 5 vertices).
Solution ambiguity

Ambiguity of the results even with complete and correct data

Many distinct minimal networks have exactly the same set of contained trees, triplets, and clusters.

2 level-2 networks with exactly the same triplet set

Gambette & Huber, 2011
Ambiguity of the results even with complete and correct data

Many distinct minimal networks have exactly the same set of contained trees, triplets, and clusters.

2 level-2 networks with exactly the same triplet set

Even with complete and correct data, impossible to choose among the ambiguous configurations!
### Practical use

<table>
<thead>
<tr>
<th>Conditions for use</th>
<th>Available data</th>
<th>Possible processings</th>
</tr>
</thead>
<tbody>
<tr>
<td>rooted trees</td>
<td>unrooted trees</td>
<td>Rooting with a reference species tree or topological constraints</td>
</tr>
<tr>
<td>Single-copy gene trees</td>
<td>MUL-trees (with duplicated genes)</td>
<td>MUL-tree processing</td>
</tr>
<tr>
<td>Correct clusters and triplets</td>
<td>noisy data</td>
<td>Tree cleaning</td>
</tr>
<tr>
<td>Complete data (density for triplet sets, complete clusters)</td>
<td>partial data, deleted genes</td>
<td>Data filtering (clusters with high bootstrap value, present in &gt;x% of the trees)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data editing : solution containing most of the input data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selection of a large number of trees with a large number of common species</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selection of the maximal number of taxa with triplet density</td>
</tr>
</tbody>
</table>

**Notes:**
- **PhySIC_IST, 2008**
- **Scornavacca, Berry & Ranwez, 2009**
- **NP-complete problems**
- **existing methods still work to do!**
Outline

- Phylogenetic networks
- Motivations for the combinatorial reconstruction approach
- Combinatorial reconstruction methods
- Practical use
- Illustrations
- Perspectives
16 trees on 47 taxa from the HOGENOM database (proteobacteria)
24 Enterobacteriales
2 Pasteurellales
1 Aeromonadales
9 Alteromonadales
1 Oceanospirillales
6 Rhodobacterales
4 Rhizobiales

Networks containing triplets, softwired clusters, present in at least 20% of the trees

Simplistic (triplets, level-7 network) 63 sec.

Dendroscope (clusters, galled network) <1 sec.
**16 trees on 47 taxa** from the HOGENOM database (proteobacteria)
- 24 Enterobacteriales
- 2 Pasteurellales
- 1 Aeromonadales
- 9 Alteromonadales
- 1 Oceanospirillales
- 6 Rhodobacterales
- 4 Rhizobiales

Networks containing triplets, softwired clusters, present in at least 20% of the trees
9 trees on 279 procaryote species
Clusters in at least 2 trees

Auch, Steigele, Huson & Henz, 2009
Illustrations

23 trees, 45 species from the 3 domains of life clusters with 99% bootstrap confidence

23 trees, 45 species from the 3 domains of life clusters with 80% bootstrap confidence present in at least 2 trees

Illustrations

23 trees, 45 species from the 3 domains of life clusters with 80% bootstrap confidence present in at least 2 trees

• Phylogenetic networks
• Motivations for the combinatorial reconstruction approach
• Combinatorial reconstruction methods
• Practical use
• Illustrations
• Perspectives
**Perspectives**

**Combinatorics:**
- Better knowledge of small level networks
- Update of a network with new data
- Unrooted explicit phylogenetic network reconstruction

**Bioinformatics:**
- Function of transferred genes ("transfer highways")
- Integration of combinatorial methods in a statistical framework

---

**Diagram:**

1. Sequence data
2. Construction of combinatorial data
3. Combinatorial reconstruction of a set of candidates
4. Choice among the candidates by statistical methods
5. Phylogenetic network
Thank you!

Coauthors:
- Vincent Berry, Christophe Paul (LIRMM)
- Daniel Huson, Regula Rupp (Tübingen)
- Katharina Huber (East Anglia)

Brown et al.'s data provided by Sophie Abby

Réticulogramme des 25 mots les plus fréquents de ma thèse, construit par TreeCloud, SplitsTree et T-Rex
Coloration: rouge au début, bleu à la fin

http://www.treecloud.org