

Groves – clustering phylogenetic datasets

Cécile Ané¹, Oliver Eulenstein², Raul Piaggio² and
Mike Sanderson³

¹University of Wisconsin - Madison

²Iowa State University

³University of California - Davis

MEP 2005



Clustering trees in a large database

Example: Database built in Driskell *et al.*, Science (2004).
All green plant sequences from **GenBank**

Goals:

- Determine whether a given set of genes is **worth combining**.
- Build **clusters** of genes for **future** combined analysis.
- Determine the **minimum number** of clusters required to **cover** GenBank.



Combining trees for Supertree construction



angiosperms,
pines, ferns



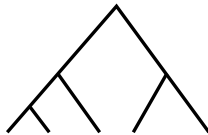
mosses, fungi,
animals



Combining trees for Supertree construction



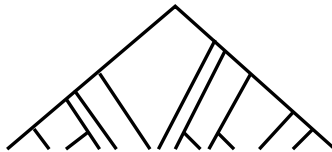
angiosperms,
pines, ferns



pines, ferns,
algae, mosses



mosses, fungi,
animals



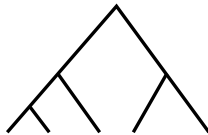
angiosperms, pines, ferns, algae, mosses, fungi, animals



Combining trees for Supertree construction



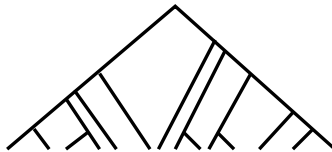
angiosperms,
pines, ferns



pines, ferns,
algae, mosses



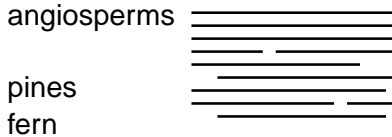
mosses, fungi,
animals



angiosperms, pines, ferns, algae, mosses, fungi, animals



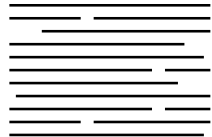
Combining matrices for Supermatrix analysis



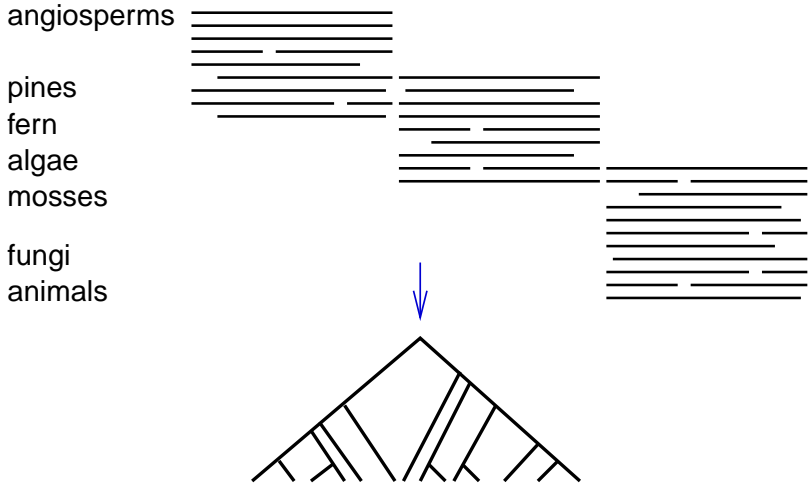
mosses

fungi

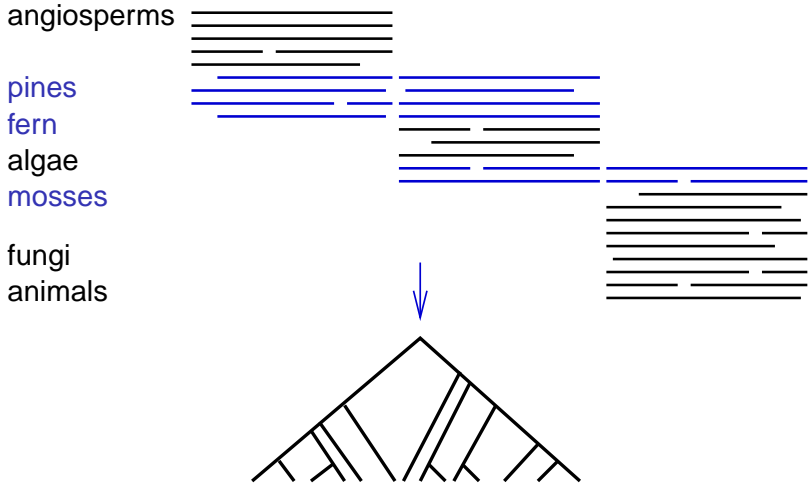
animals



Combining matrices for Supermatrix analysis

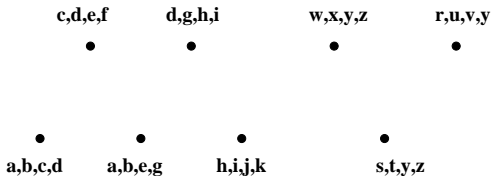


Combining matrices for Supermatrix analysis



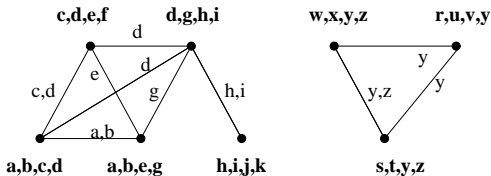
Clustering a database

Criterion: **Potential** for **new phylogenetic information** through combined analysis.



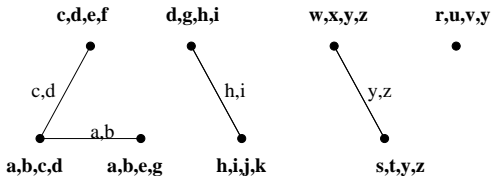
Clustering a database

Criterion: **Potential** for **new phylogenetic information** through combined analysis.



Clustering a database

Criterion: **Potential** for **new phylogenetic information** through combined analysis.

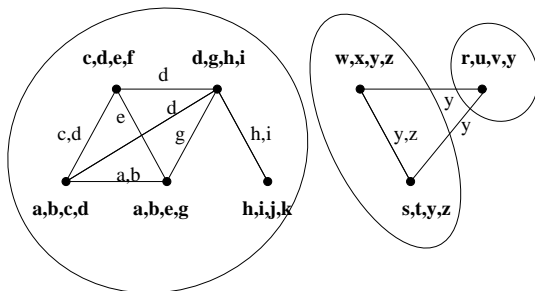


Overlap of 2 taxa necessary? (Sanderson *et. al*, 1998)



Clustering a database

Criterion: **Potential** for **new phylogenetic information** through combined analysis.

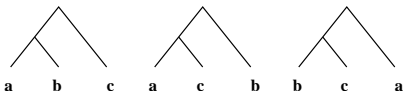


Overlap of 2 taxa necessary? (Sanderson *et. al*, 1998)



Basic information unit

- Rooted trees.
- Basic information unit = 3-taxon tree
- 3 possible trees over the triple {a,b,c}



Cross triples and tree assignments

b, c, d
•

a, b, c
•

- {**abc**} observed
- {**abd**} cross triple
- This collection of taxon sets is a **Grove**



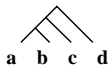
Cross triples and tree assignments



- **{abc}** observed
- **{abd}** cross triple
- This collection of taxon sets is a **Grove**



Cross triples and tree assignments

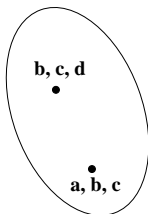


One single parent tree

- **{abc}** observed
- **{abd}** cross triple
resolved by some assignment of compatible trees
- This collection of taxon sets is a **Grove**



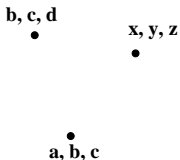
Cross triples and tree assignments



- **{abc}** observed
- **{abd}** cross triple
resolved by some assignment of compatible trees
- This collection of taxon sets is a **Grove**



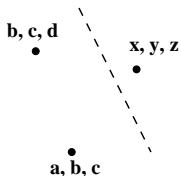
Cross triples and tree assignments



- Some information might be gained from combining all three. ex: {abd}. **Not enough!**
- Consider a **partition**, and cross triples not already resolved by any side of the partition ex: {**abz**}
- Can some such cross triple get resolved?
- **Not a Grove!**



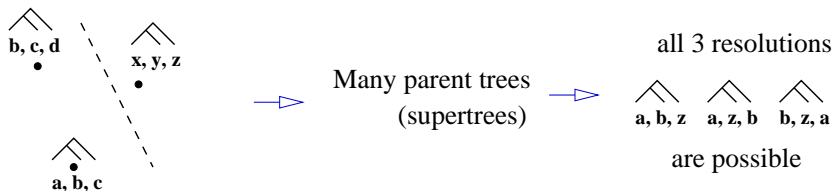
Cross triples and tree assignments



- Some information might be gained from combining all three. ex: {abd}. **Not enough!**
- Consider a **partition**, and cross triples not already resolved by any side of the partition ex: {**abz**}
- Can some such cross triple get resolved?
- **Not a Grove!**



Cross triples and tree assignments



- Some information might be gained from combining all three. ex: $\{abd\}$. **Not enough!**
- Consider a **partition**, and cross triples not already resolved by any side of the partition ex: $\{abz\}$
- Can some such cross triple get resolved? No.
- **Not a Grove!**

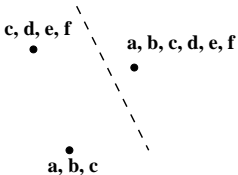


Groves: definition

Definition

A **grove** is a collection of taxon sets that satisfies the following.

- There is no cross triple, or
- For **any** partition of the collection, **some** assignment of compatible input trees resolves **some** cross triple that was not already resolved by any side of the partition.



Groves: definition

Definition

A **grove** is a collection of taxon sets that satisfies the following.

- There is no cross triple, or
- For **any** partition of the collection, **some** assignment of compatible input trees resolves **some** cross triple that was not already resolved by any side of the partition.

$\{a,b,c\}$ is a **cross triple** if no taxon set contains all three taxa **a,b,c**.



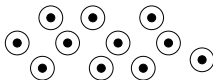
Groves: definition

Definition

A **grove** is a collection of taxon sets that satisfies the following.

- There is no cross triple, or
- For **any** partition of the collection, **some** assignment of compatible input trees resolves **some** cross triple that was not already resolved by any side of the partition.

$\{a,b,c\}$ is a **cross triple** if no taxon set contains all three taxa a,b,c .



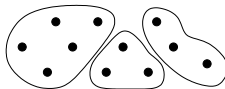
Groves: definition

Definition

A **grove** is a collection of taxon sets that satisfies the following.

- There is no cross triple, or
- For **any** partition of the collection, **some** assignment of compatible input trees resolves **some** cross triple that was not already resolved by any side of the partition.

$\{a,b,c\}$ is a **cross triple** if no taxon set contains all three taxa a,b,c .

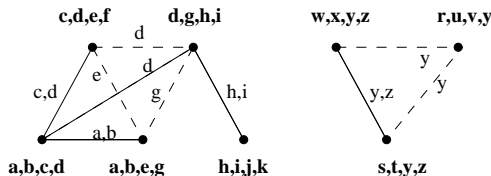


Overlap of 2 is sufficient

Theorem

If the 2-overlap graph is connected, then the database is a grove.

2-overlap graph: an edge connects 2 taxon sets if they overlap by 2 or more taxa.

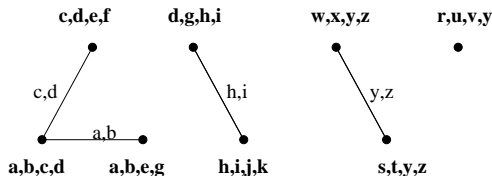


Overlap of 2 is sufficient

Theorem

If the 2-overlap graph is connected, then the database is a grove.

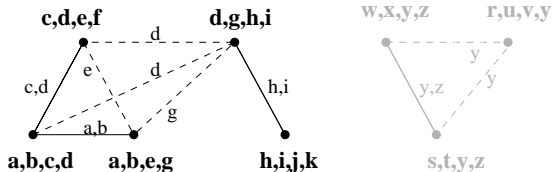
2-overlap graph: an edge connects 2 taxon sets if they overlap by 2 or more taxa.



Overlap of 2 not necessary

Theorem

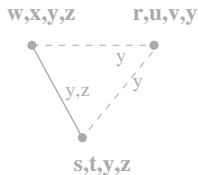
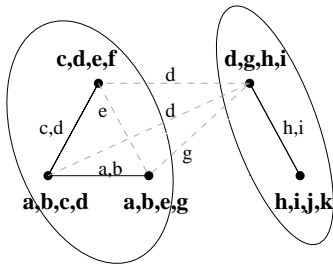
If the 2-overlap graph has 2 connected components, themselves sharing 2 or more taxa, then the database is a grove.



Overlap of 2 not necessary

Theorem

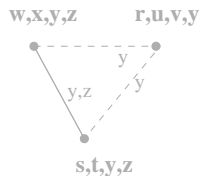
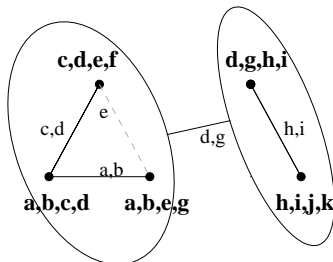
If the 2-overlap graph has 2 connected components, themselves sharing 2 or more taxa, then the database is a grove.



Overlap of 2 not necessary

Theorem

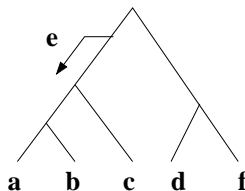
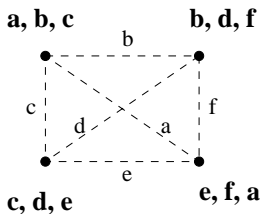
If the 2-overlap graph has 2 connected components, themselves sharing 2 or more taxa, then the database is a grove.



Overlap of 1 sometimes sufficient

Proposition

This is a grove!



Similar example in the unrooted case.

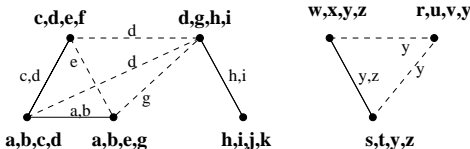


Bounds on the grove coverage number

Theorem

$$CC(\mathcal{G}_1) + b \leq G \leq CC(\mathcal{G}_2) - m$$

- G : *minimum # of groves* required to *cover* the database
- $CC(\mathcal{G}_1)$: # connected components of the 1-overlap graph
- $CC(\mathcal{G}_2)$: # connected components of the 2-overlap graph

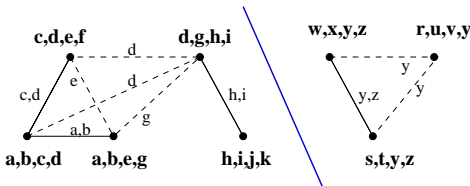


Bounds on the grove coverage number

Theorem

$$CC(\mathcal{G}_1) + b \leq G \leq CC(\mathcal{G}_2) - m$$

- G : *minimum # of groves* required to *cover* the database
- $CC(\mathcal{G}_1)$: # connected components of the 1-overlap graph
- $CC(\mathcal{G}_2)$: # connected components of the 2-overlap graph

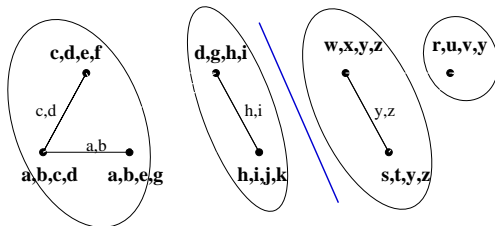


Bounds on the grove coverage number

Theorem

$$CC(\mathcal{G}_1) + b \leq G \leq CC(\mathcal{G}_2) - m$$

- G : *minimum # of groves* required to *cover* the database
- $CC(\mathcal{G}_1)$: # connected components of the 1-overlap graph
- $CC(\mathcal{G}_2)$: # connected components of the 2-overlap graph

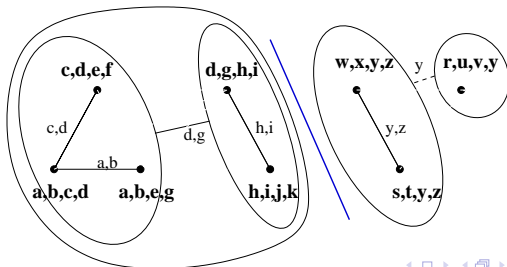


Bounds on the grove coverage number

Theorem

$$CC(\mathcal{G}_1) + b \leq G \leq CC(\mathcal{G}_2) - m$$

- G : **minimum # of groves** required to **cover** the database
- $CC(\mathcal{G}_1)$: # connected components of the 1-overlap graph
- $CC(\mathcal{G}_2)$: # connected components of the 2-overlap graph



Clustering in a real database

Database:

- From Driskell *et al.* (2004).
- **853 alignments** (genes) , over 39,000 sequences and 14,500 taxa.

Results:

- \mathcal{G}_1 has 8 connected components,
- \mathcal{G}_2 has 32 connected components,

$$24 \leq G \leq 31$$



Open questions

- Do maximal groves partition the database?
- Is the union of 2 groves a grove itself?



Acknowledgments

- co-authors: Oliver Eulenstein, Raul Piaggio-Talice and Mike Sanderson,
- Shelley McMahon and Amy Driskell
- NSF

Thank you!

