7th Course: Phylogeny and chordal graphs

Michel Habib
habib@liafa.jussieu.fr
http://www.liafa.jussieu.fr/~habib

Sophie Germain, octobre 2013

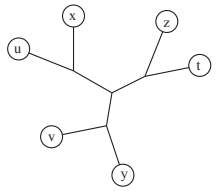
Schedule

Schedule

Schedule

Phylogenetics

Objective: reconstruct evolutionary history of species from biological data of present-day species.
Models: phylogenetic trees, networks (rooted or unrooted)



An unrooted phylogenetic tree on the set of species $\mathcal{L} = \{x, y, z, t, u, v\}$

Perfect Phylogeny and chordal graphs

The data is an incidence matrix from a set of species $S_1, \ldots, S_n$ to a set of attributes $F_1, \ldots, F_m$.
Attributes can have different values (integers or colors).
These species are modern species and we aim at reconstructing a plausible historical evolution tree T, finding ancestral species.
In T vertices are species, and its leaves correspond to $S_1, \ldots, S_n$ and every value of an attribute must define a connected subtree.

Let us call the value of an attribute : a character (N.B. this definition is not standard).
Using Darwin theory of evolution, to each character must correspond a unique subtree of the tree of life.

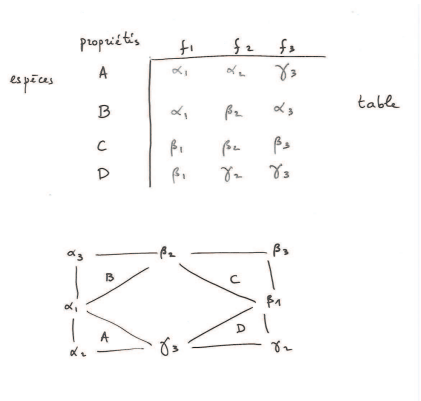
We define a graph G whose vertices are the value of the attributes (or characters). Two values are joined by an edge if they are both present in some species.
Therefore every species corresponds to a clique in G. This clique is maximal since in our formalism every species is defined by a unique set of characters and no species contains another one (as set of characters).
Of course if we do not know all the characters of ancient species, the graph is not chordal and we have to complete it as a chordal graph by adding edges (and therefore new maximal cliques and new species).
When G is chordal, it admits maximal clique trees. If there exists a regular clique tree T having $S_1, \ldots, S_n$ as set of leaves, such a situation is called a Perfect phylogeny.

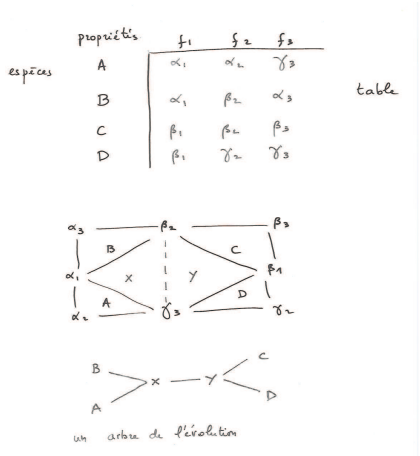
Many researches in this field to cope with uncertainity, bad data or partial data.
One model is using partial partitions of the species such as triplets, quadruplets . . . .

One major problem:

For which data there exists a unique phylogenetic tree ?

An example





Discussion

1. $X(\alpha_1, \beta_2, \gamma_3)$, $Y(\beta_1, \beta_2, \gamma_3)$ the new maximal cliques of $G + \beta_2\gamma_3$ are plausible ancestral species.
   In this solution to α_1 it corresponds the subtree : A, X, B and to β_1 the subtree C, Y, D.

Discussion

1. $X(\alpha_1, \beta_2, \gamma_3)$, $Y(\beta_1, \beta_2, \gamma_3)$ the new maximal cliques of $G + \beta_2\gamma_3$ are plausible ancestral species.
   In this solution to α_1 it corresponds the subtree : A, X, B and to β_1 the subtree C, Y, D.
2. G could have been also completed as a chordal graph, by adding the edge $\alpha_1\beta_1$, but then this would imply that F_1 is not properly defined since α_1, β_1 are values of F_1.

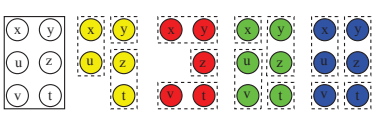
Discussion

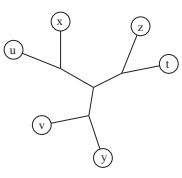
1. $X(\alpha_1, \beta_2, \gamma_3)$, $Y(\beta_1, \beta_2, \gamma_3)$ the new maximal cliques of $G + \beta_2\gamma_3$ are plausible ancestral species.
   In this solution to α_1 it corresponds the subtree : A, X, B and to β_1 the subtree C, Y, D.
2. G could have been also completed as a chordal graph, by adding the edge $\alpha_1\beta_1$, but then this would imply that F_1 is not properly defined since α_1, β_1 are values of F_1.
3. Therefore attributes must correspond to independent sets of G (or colors).

More precisely

Input:
A species set $\mathcal{L}$
A character set $\mathcal{C}$ on $\mathcal{L}$

Output:
A phylogenetic tree (ternary) on $\mathcal{L}$ on which all characters of $\mathcal{C}$ are convex



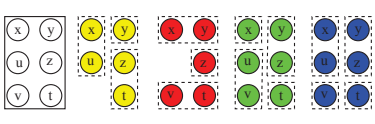


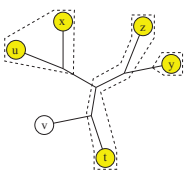
A: a 3-*state* character
C: a 4-*state full* character.

More precisely

Input:
A species set $\mathcal{L}$
A character set $\mathcal{C}$ on $\mathcal{L}$

Output:
A phylogenetic tree (ternary) on $\mathcal{L}$ on which all characters of $\mathcal{C}$ are convex





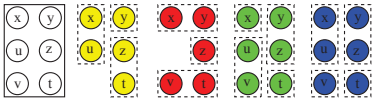
A: a 3-*state* character
C: a 4-*state full* character.

A is convex on T

More precisely

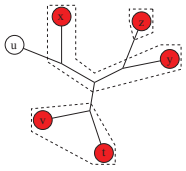
Input:
A species set $\mathcal{L}$
A character set $\mathcal{C}$ on $\mathcal{L}$

Output:
A phylogenetic tree (ternary) on $\mathcal{L}$ on which all characters of $\mathcal{C}$ are convex



$\mathcal{L}$ A B C D

A: a 3-*state* character
C: a 4-*state full* character.

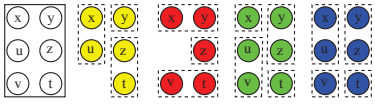


B is convex on T

More precisely

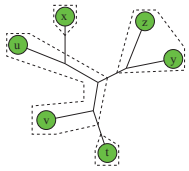
Input:
A species set $\mathcal{L}$
A character set $\mathcal{C}$ on $\mathcal{L}$

Output:
A phylogenetic tree (ternary) on $\mathcal{L}$ on which all characters of $\mathcal{C}$ are convex



$\mathcal{L}$ A B C D

A: a 3-*state* character
C: a 4-*state full* character.

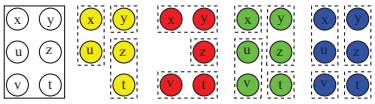


C is convex on T

More precisely

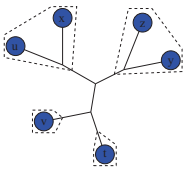
Input:
A species set $\mathcal{L}$
A character set $\mathcal{C}$ on $\mathcal{L}$

Output:
A phylogenetic tree (ternary) on $\mathcal{L}$ on which all characters of $\mathcal{C}$ are convex



$\mathcal{L}$ A B C D

A: a 3-*state* character
C: a 4-*state full* character.

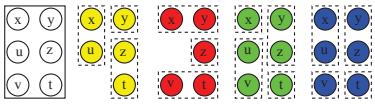


D is convex on T

More precisely

Input:
A species set $\mathcal{L}$
A character set $\mathcal{C}$ on $\mathcal{L}$

Output:
A phylogenetic tree (ternary) on $\mathcal{L}$ on which all characters of $\mathcal{C}$ are convex



$\mathcal{L}$ A B C D

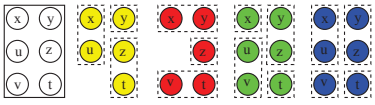
A: a 3-*state* character
C: a 4-*state full* character.

$\Rightarrow A, B, C, D$ are compatible

More precisely

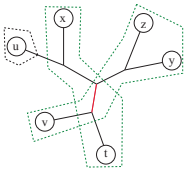
Input:
A species set $\mathcal{L}$
A character set $\mathcal{C}$ on $\mathcal{L}$

Output:
A phylogenetic tree (ternary) on $\mathcal{L}$ on which all characters of $\mathcal{C}$ are convex



$\mathcal{L}$ A B C D

A: a 3-*state* character
C: a 4-*state full* character.



A character not convex on T

Previous work

Theorem (Fitch 75, Meacham 83)

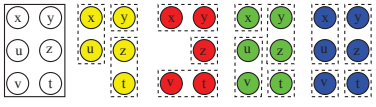
For any $r \geq 2$, there exists a set of r-state full characters in which every $r-1$ characters are compatible but the whole set is not compatible.

Theorem (Meacham 83, Lam and Gusfield 09)

For $r = 2, 3$, a set of r-state full characters is compatible iff every r characters are compatible.

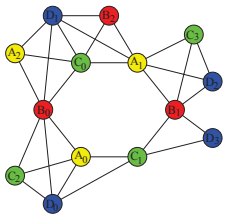
Question: $\forall r \geq 2$, a set of r-state full characters is compatible iff every r characters are compatible? [Lam and Gusfield 09]

Intersection partition graph



$\mathcal{L}$ A B C D

$\mathcal{C} = \{A, B, C, D\}$
$A = \{x, u\}|\{z, t\}|\{y\} = A_0|A_1|A_2$
$B = \{x, y\}|\{t, v\}|\{z\} = B_0|B_1|B_2$
$C = \{y, z\}|\{u, v\}|\{t\}|\{v\} = C_0|C_1|C_2|C_3$
$D = \{x, u\}|\{y, z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$



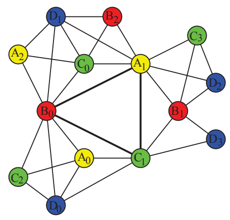
Intersection partition graph of $\mathcal{C}$

Proper chordal completion

A graph is **chordal** iff every chordless cycle is a triangle, i.e. has length 3.
A **chordal completion** of a graph $G = (V, E)$ is a chordal graph $G' = (V, E')$ such that $E \subseteq E'$.
A **proper chordal completion** of a vertex-coloured graph G is a chordal completion of G without connecting any pair of vertices of the same colour.

Theorem (Meacham 83, Steel 92)

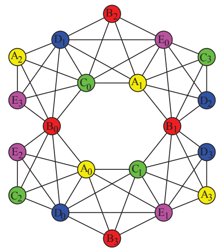
$\mathcal{C}$ is compatible $\Leftrightarrow$ the intersection partition graph of $\mathcal{C}$ has a proper chordal completion



Intersection partition graph

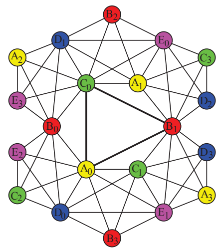
4-State Full Characters

$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



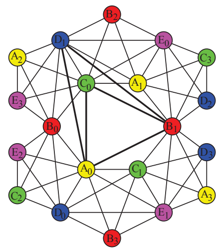
4-State Full Characters

$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



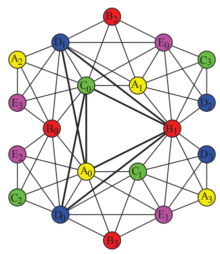
4-State Full Characters

$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



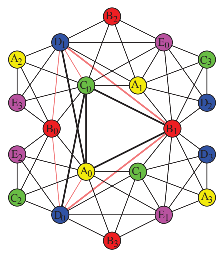
4-State Full Characters

$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



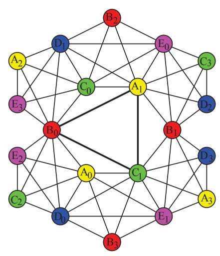
4-State Full Characters

$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



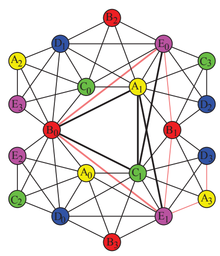
4-State Full Characters

$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



4-State Full Characters

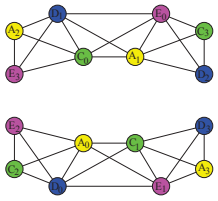
$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



A, B, C, D, E are not compatible

4-State Full Characters

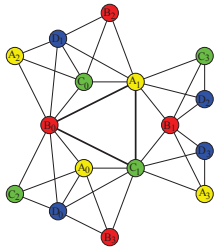
$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



A, C, D, E are compatible

4-State Full Characters

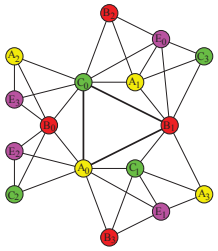
$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



A, B, C, D are compatible

4-State Full Characters

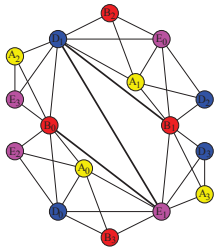
$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



A, B, C, E are compatible

4-State Full Characters

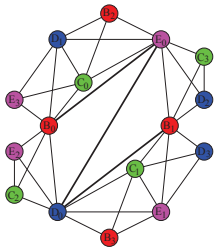
$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



A, B, D, E are compatible

4-State Full Characters

$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$



B, C, D, E are compatible

4-State Full Characters

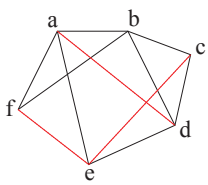
$A = \{x,u\}|\{z,t\}|\{y\}|\{v\} = A_0|A_1|A_2|A_3$
$B = \{x,y\}|\{t,v\}|\{z\}|\{u\} = B_0|B_1|B_2|B_3$
$C = \{y,z\}|\{u,v\}|\{x\}|\{t\} = C_0|C_1|C_2|C_3$
$D = \{x,u\}|\{y,z\}|\{t\}|\{v\} = D_0|D_1|D_2|D_3$
$E = \{z,t\}|\{u,v\}|\{x\}|\{y\} = E_0|E_1|E_2|E_3$

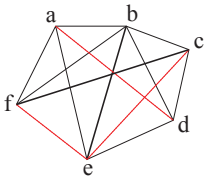
To determine the compatibility of set of a 4-state characters, it is not sufficient to test the compatibility of every 4 characters

Notations

$P = [V, E, F]$ is a **sandwich problem** of the graph $G = (V, E)$, where $F \subseteq V \times V \setminus E$.

$G_S = (V, E_S)$ is called a **Π-sandwich graph** of P if $E \subseteq E_S \subseteq V \times V \setminus F$ and G_S satisfies property Π.

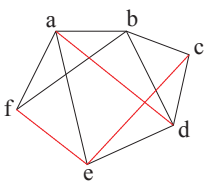


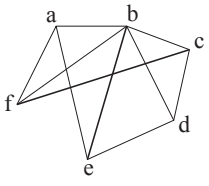


Notations

$P = [V, E, F]$ is a **sandwich problem** of the graph $G = (V, E)$, where $F \subseteq V \times V \setminus E$.

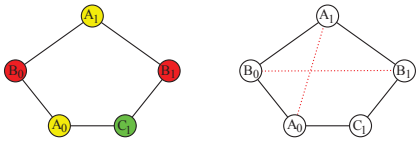
$G_S = (V, E_S)$ is called a **Π-sandwich graph** of P if $E \subseteq E_S \subseteq V \times V \setminus F$ and G_S satisfies property Π.





Chordal Sandwich Graph

By taking $F = \{(u, v)|u, v \text{ have the same colour}\}$, the problem of chordal completion of vertex-coloured graph is transformed to chordal sandwich graph.



Consequences

- Perfect phylogeny is NP-complete [Bodlaender et al. 92, Steel 92], since chordal sandwich problem is NP-hard.

Consequences

- Perfect phylogeny is NP-complete [Bodlaender et al. 92, Steel 92], since chordal sandwich problem is NP-hard.
- Unique Perfect phylogeny is NP-hard [Habib, Stacho 11] also unique minimal chordal sandwich problem is NP-hard.