Constraint Programming and Biology: Phylogenetic trees

Agostino Dovier

Dept. Math and Computer Science, Univ. of Udine, Italy

ACP Summer School in Constraint Programming
Wrocław, September 2012

Phylogenetic trees

Basics

- The starting point is a set L of elementary (taxonomic) units, known as taxa (e.g., $L = \{English, German, French, Spanish, Italian\}$ or $L = \{dog, cat, horse, chicken\}$)
- A set I of characters is assigned to each element of L (e.g., characters "hand" and "father", or characters "number of legs", "length of the tail", etc.)
- Characters can have values (e.g. {1 (hand), 2 (mano/main)} for "hand" and {1 (father/padre), 2 (vater/pere)} for "father") Each element in L is assigned a value.
- Let us focus on Boolean characters
- Reconstruction of phylogenies is the first step of reconstructing the evolutionary history of the set L of taxa.

Phylogenetic trees

Example (Grapewine Genome Characterization — Nature 449, Sept. 2007)

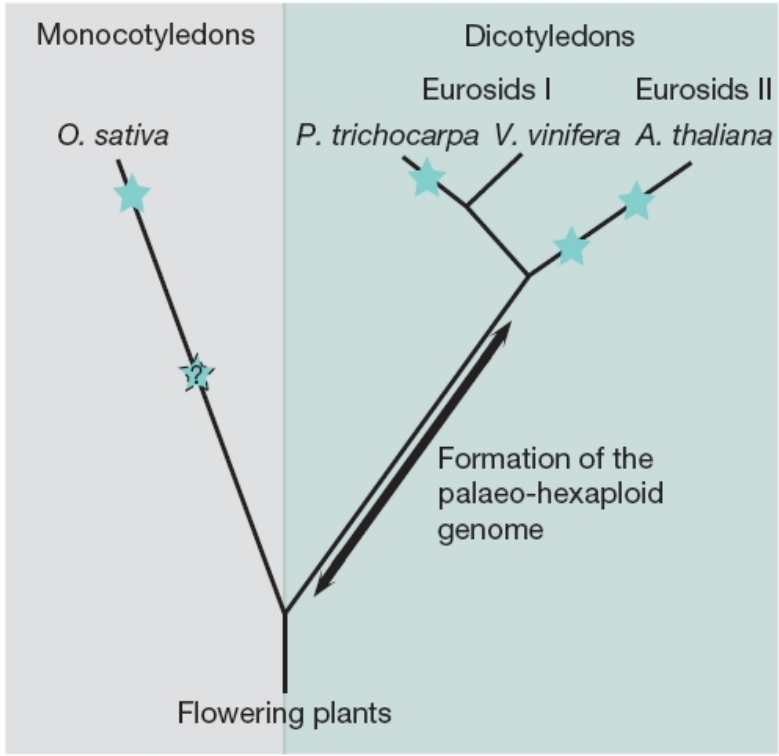
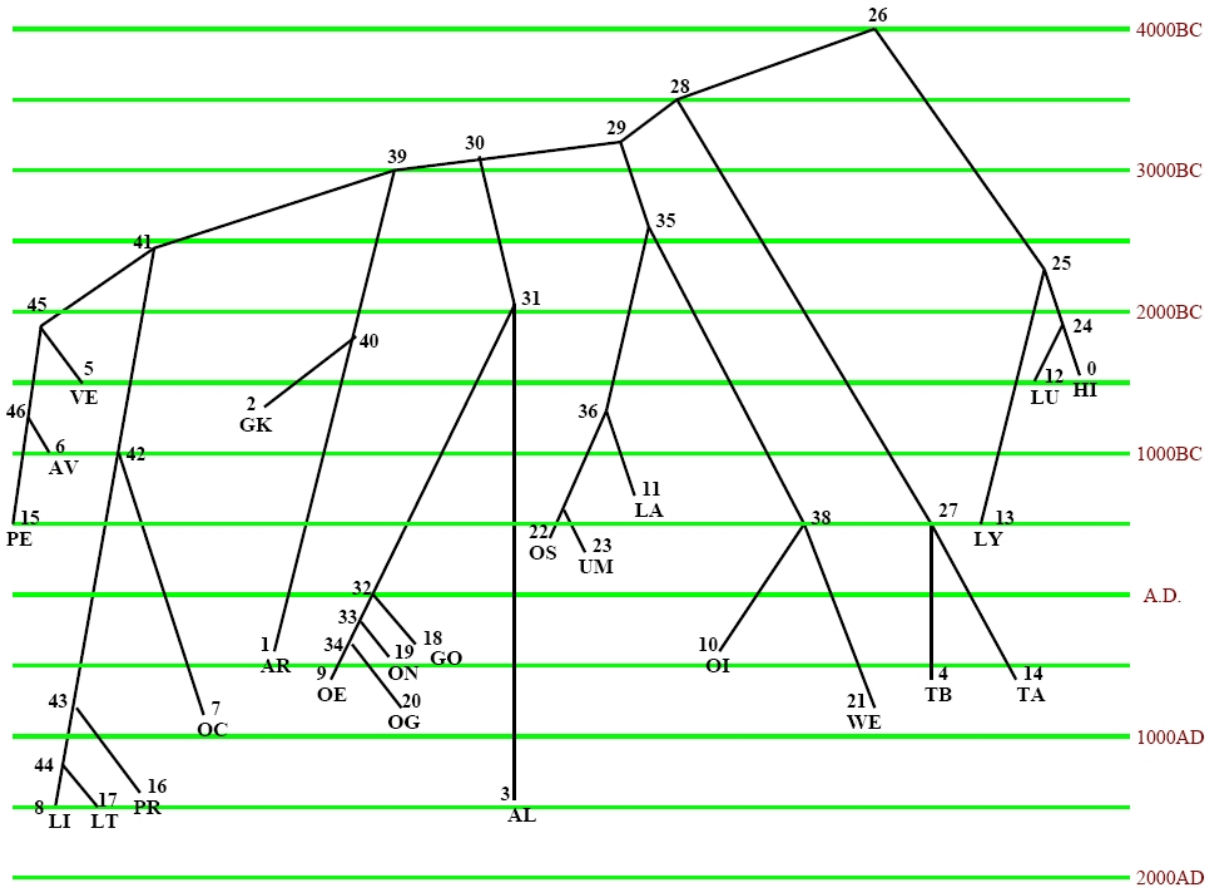


Figure 3 | Positions of the polyploidization events in the evolution of plants with a sequenced genome. Each star indicates a WGD (tetraploidization) event on that branch. The question mark indicates that ancient events are visible in the rice genome that would require other monocotyledon genome sequences to be resolved. The formation of the palaeo-hexaploid ancestral genome occurred after divergence from monocotyledons and before the radiation of the Eurosids.

Phylogenetic trees

Example (Indo-European languages phylogeny–Erdem et al. 2003)



Phylogenetic trees

Example (Indo-European languages phylogeny–Erdem et al. 2003)

Hittite (HI), Luvian (LU), Lycian (LY), Tocharian A (TA), Tocharian B (TB), Vedic (VE), Avestan (AV), Old Persian (PE), Classical Armenian (AR), Ancient Greek (GK), Latin (LA), Oscan (OS), Umbrian (UM), Gothic (GO), Old Norse (ON), Old English (OE), Old High German (OG), Old Irish (OI), Welsh (WE), Old Church Slavonic (OC), Old Prussian (PR), Lithuanian (LI), Latvian (LT), and Albanian (AL).

Phylogenetic trees

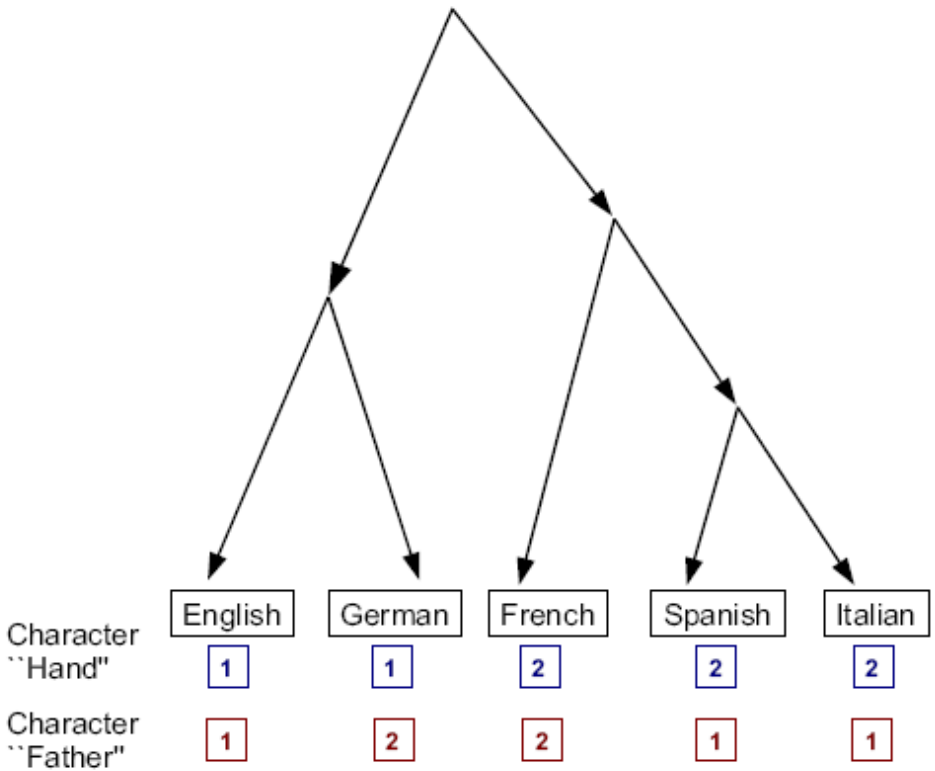
- A phylogeny

$$(V, E, L, I, S, f)$$

for a set L of taxa is a
  - finite binary tree (V, E) (rooted by some $\nu \in V$) with leaves $L \subseteq V$ (taxa=leaves, with a slight abuse of notation)
  - along with two finite sets I and S and a function $f : L \times I \longrightarrow S$.
- Intuitively, L are the taxonomic units, $V \setminus L$ describes their ancestral units and E genetic relationships between them.
- I is the set of characters, and S is the set of their values (also knows are states of these characters)
- f labels every leaf $v \in L$ by assigning a state to each character $i \in I$

Phylogenetic trees

Example (from Erdem 2011)



A phylogeny (V, E, L, I, S, f) where
$L = \{\text{English}, \text{German}, \text{French}, \text{Spanish}, \text{Italian}\}$ (taxa)
$I = \{\text{Hand}, \text{Father}\}$ (characters), $S = \{1, 2\}$ (states).

Phylogenetic trees

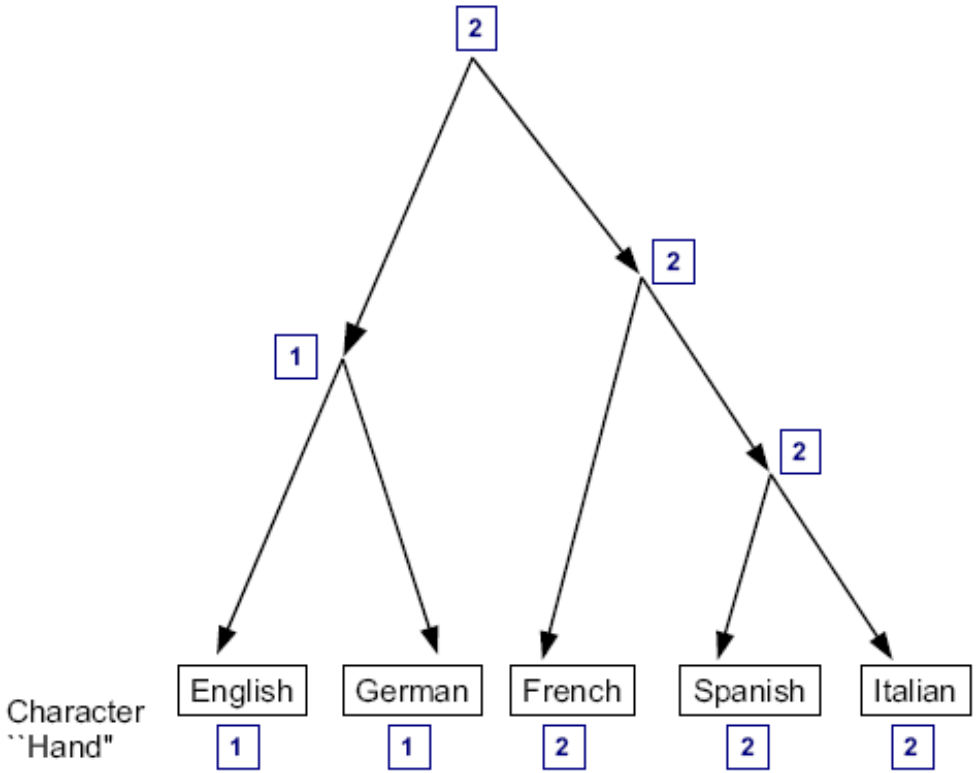
- A character $i \in I$ is compatible with a phylogeny (V, E, L, I, S, f) if there is a function $g : V \times \{i\} \longrightarrow S$ s.t.
  - ✓ $(\forall v \in L)(g(v, i) = f(v, i))$ (g extends f) and
  - ✓ Let us denote $V_{i,s} = \{x \in V : g(x, i) = s\}$.

Phylogenetic trees

- A character $i \in I$ is compatible with a phylogeny (V, E, L, I, S, f) if there is a function $g : V \times \{i\} \longrightarrow S$ s.t.
  - ✓ $(\forall v \in L)(g(v, i) = f(v, i))$ (g extends f) and
  - ✓ Let us denote $V_{i,s} = \{x \in V : g(x, i) = s\}$.
    Then $(\forall s \in S)(V_{i,s} \neq \emptyset \Rightarrow$
    the subgraph $(V_{i,s}, E \cap V_{i,s}^2)$ of (V, E) is a rooted tree.)

Phylogenetic trees

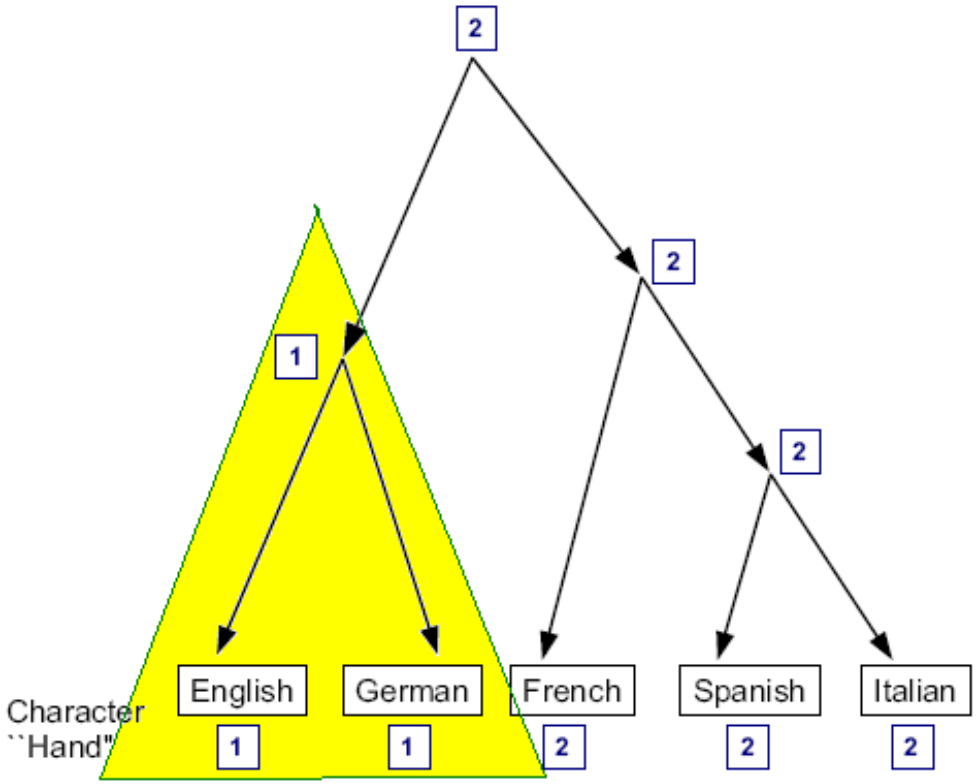
Example (from Erdem 2011)



Character Hand is compatible with the above tree

Phylogenetic trees

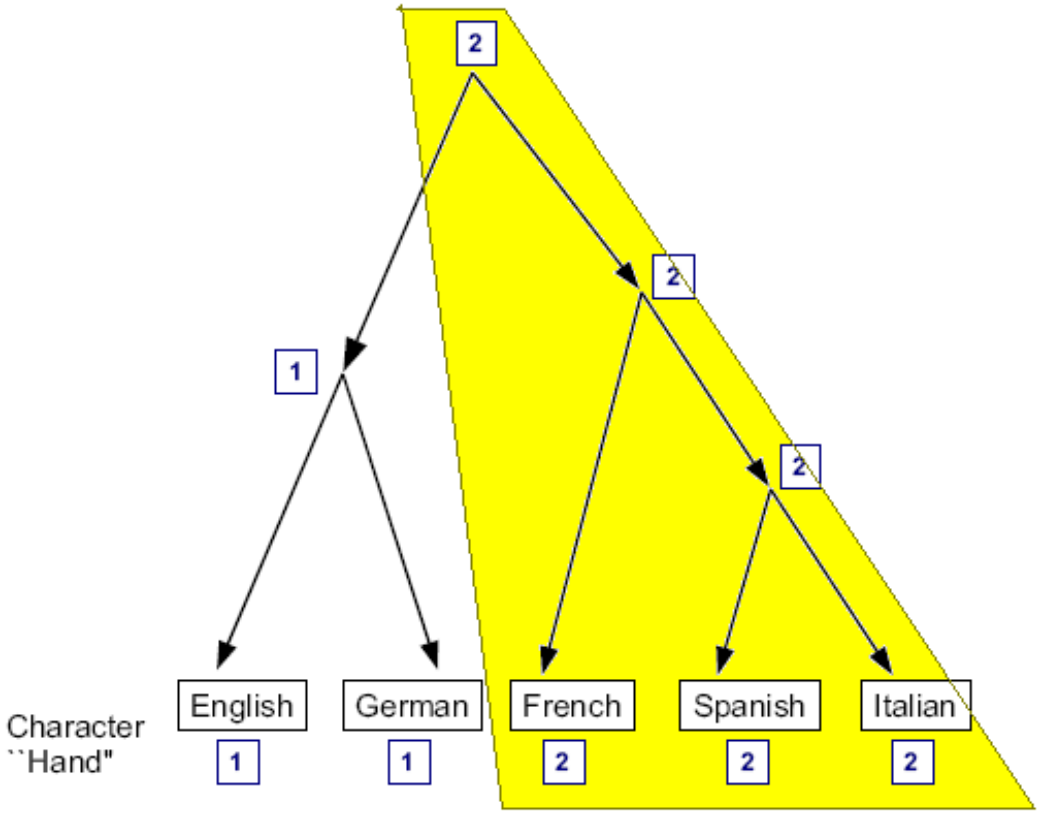
Example (from Erdem 2011)



Character Hand is compatible with the above tree

Phylogenetic trees

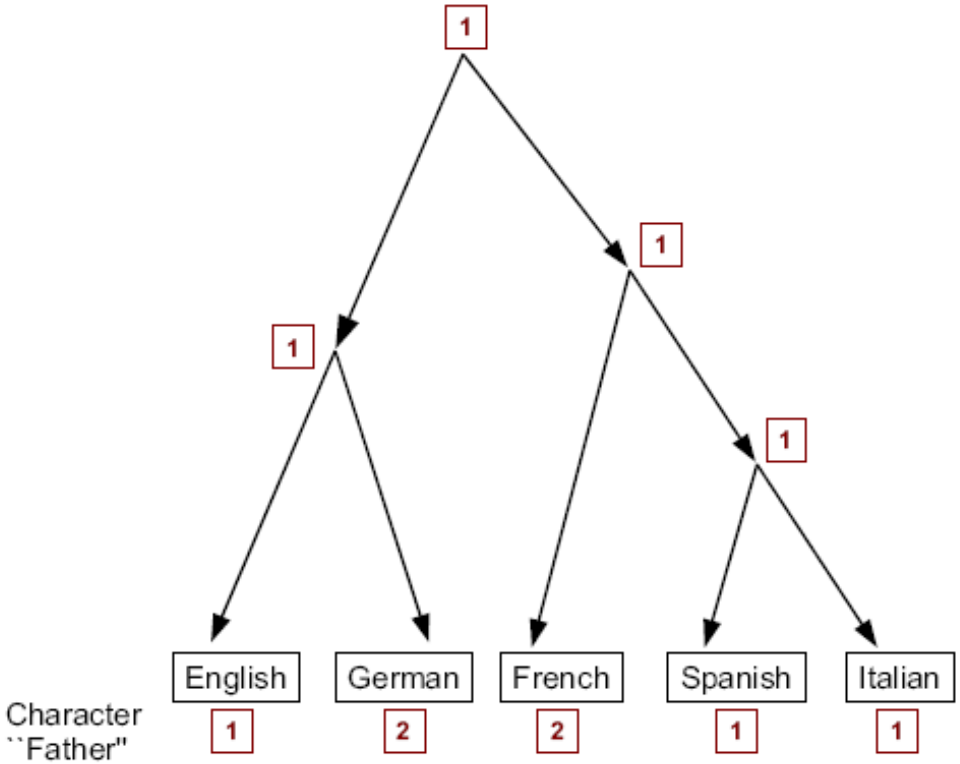
Example (from Erdem 2011)



Character Hand is compatible with the above tree

Phylogenetic trees

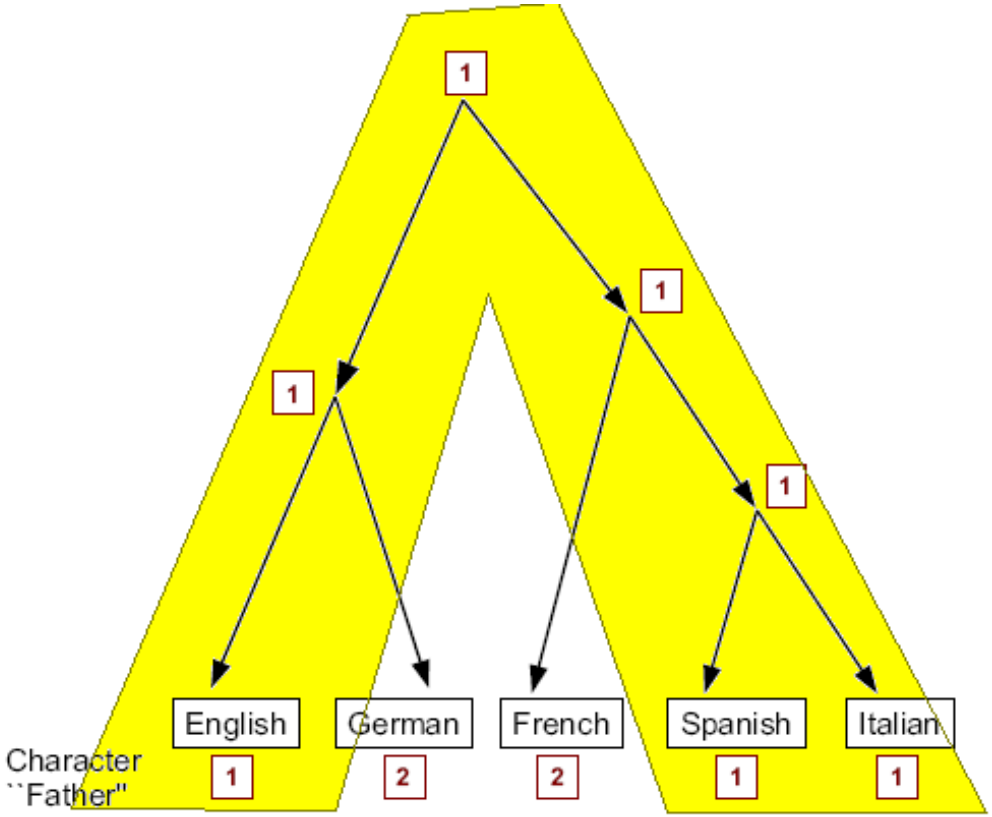
Example (from Erdem 2011)



Character Father is incompatible with the above tree

Phylogenetic trees

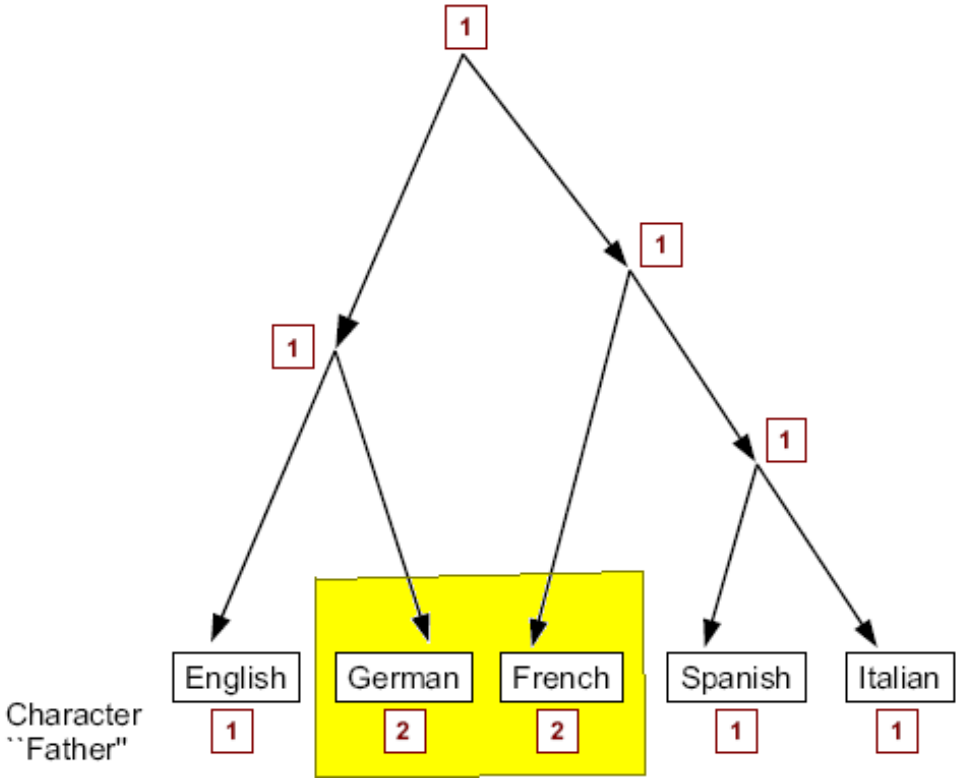
Example (from Erdem 2011)



Character Father is incompatible with the above tree

Phylogenetic trees

Example (from Erdem 2011)



Character Father is incompatible with the above tree

Phylogenetic trees

- A character $i \in I$ is compatible with a phylogeny (V, E, L, I, S, f) if there is a function $g : V \times \{i\} \longrightarrow S$ s.t.
  - ✓ $(\forall v \in L)(g(v, i) = f(v, i))$ (g extends f) and
  - ✓ Let us denote $V_{i,s} = \{x \in V : g(x, i) = s\}$.
    Then $(\forall s \in S)(V_{i,s} \neq \emptyset \Rightarrow$
    the subgraph $(V_{i,s}, E \cap V_{i,s}^2)$ of (V, E) is a rooted tree.)
- Otherwise it is incompatible
- The above (sub-tree) requirement implicitly states that when a character changes (in the evolution) it never go back to the previous value (Camin-Sokal). Moreover, that the "change" occurs in a unique place (Dollo).
- cladistic characters are those that have a temporal order (ancestral/derived), otherwise they are qualitative

Phylogenetic trees

One problem

k-INCOMPATIBILITY PROBLEM

Given sets L (taxa/leaves), I (characters), and S (states), a function $f : L \times I \longrightarrow S$, and $k \in \mathbb{N}$, decide the existence of a phylogeny (V, E, L, I, S, f) with at most k incompatible characters.

Phylogenetic trees

One problem

k-INCOMPATIBILITY PROBLEM

Given sets L (taxa/leaves), I (characters), and S (states), a function $f : L \times I \longrightarrow S$, and $k \in \mathbb{N}$, decide the existence of a phylogeny (V, E, L, I, S, f) with at most k incompatible characters.

This problem is NP-complete (Day, Sankoff 1986). We'll see a sketch of the proof later.

Phylogenetic trees

One problem

k-INCOMPATIBILITY PROBLEM

Given sets L (taxa/leaves), I (characters), and S (states), a function $f : L \times I \longrightarrow S$, and $k \in \mathbb{N}$, decide the existence of a phylogeny (V, E, L, I, S, f) with at most k incompatible characters.

This problem is NP-complete (Day, Sankoff 1986). We'll see a sketch of the proof later.

Search space (exercise): 1) How many binary trees are there with n leaves? 2) Now, leaves can be labeled by the n taxa: multiply by $n!$ 3) For phylogenies symmetries must be removed (left/right son is the same). How many trees remains? Is there a compact formula?

Phylogenetic trees

Alternative definitions

- Given the set $L = \{a_1, \ldots, a_n\}$ of taxa, and characters $I = \{i_1, \ldots, i_m\}$, the values of the function f can be compactly expressed by a matrix X:

$I \setminus L$	a_1	a_2	$\ldots$	a_n
i_1	$f(a_1, i_1)$	$f(a_2, i_1)$	$\ldots$	$f(a_n, i_1)$
i_2	$f(a_1, i_2)$	$f(a_2, i_2)$	$\ldots$	$f(a_n, i_2)$
$\vdots$		$\ddots$		
i_m	$f(a_1, i_m)$	$f(a_2, i_m)$	$\ldots$	$f(a_n, i_m)$

- In the previous example:

	English	German	French	Spanish	Italian
Hand	1	1	2	2	2
Father	1	2	2	1	1

Phylogenetic trees

Alternative definitions

- Focusing on Boolean values, X is a $m \times n$ $\{0,1\}$ matrix.
- Each column is a vector of $\{0,1\}^m$
- Given two vectors in $\{0,1\}^m$ they are connected if they differ in exactly one element (their Hamming distance d_H is 1—e.g., $(0,0,1,0,1,1,0)$ and $(0,0,1,1,1,1,0)$)
- $\{0,1\}^m$ with the above notion generates an hypercube graph. There is always a path between two vectors.

Phylogenetic trees

Alternative definitions

- Focusing on Boolean values, X is a $m \times n$ $\{0,1\}$ matrix.
- Each column is a vector of $\{0,1\}^m$
- Given two vectors in $\{0,1\}^m$ they are connected if they differ in exactly one element (their Hamming distance d_H is 1—e.g., $(0,0,1,0,1,1,0)$ and $(0,0,1,1,1,1,0)$)
- $\{0,1\}^m$ with the above notion generates an hypercube graph. There is always a path between two vectors.
- $V \subseteq \{0,1\}^m$ identifies a graph (V, E_V), where $E_V = \{\{x,y\} \subseteq V \mid d_H(x,y) = 1\}$.
- If (V, E_V) is acyclic, we can easily build a rooted tree.

Phylogenetic trees

Alternative definitions

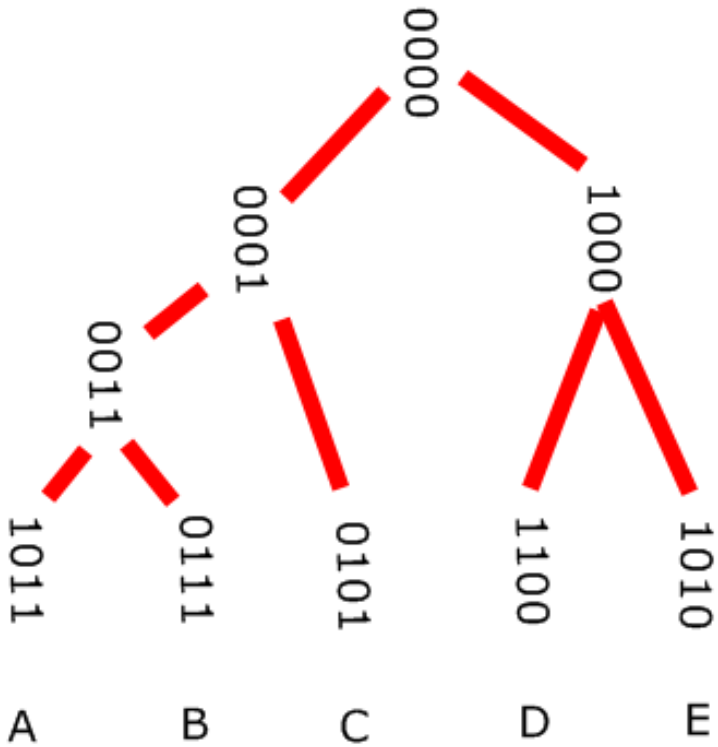
- Focusing on Boolean values, X is a $m \times n$ $\{0,1\}$ matrix.
- Each column is a vector of $\{0,1\}^m$
- Given two vectors in $\{0,1\}^m$ they are connected if they differ in exactly one element (their Hamming distance d_H is 1—e.g., $(0,0,1,0,1,1,0)$ and $(0,0,1,1,1,1,0)$)
- $\{0,1\}^m$ with the above notion generates an hypercube graph. There is always a path between two vectors.
- $V \subseteq \{0,1\}^m$ identifies a graph (V, E_V), where $E_V = \{\{x,y\} \subseteq V \mid d_H(x,y) = 1\}$.
- If (V, E_V) is acyclic, we can easily build a rooted tree.
- The tree is a phylogeny if, starting from the root, (cladistic) character changes $1 \mapsto 0$ are forbidden (Camin and Sokal 1965)
- Furthermore, one can require that the change of a character occurs in a unique edge (Dollo)

Phylogenetic trees

Example (Camin and Sokal)

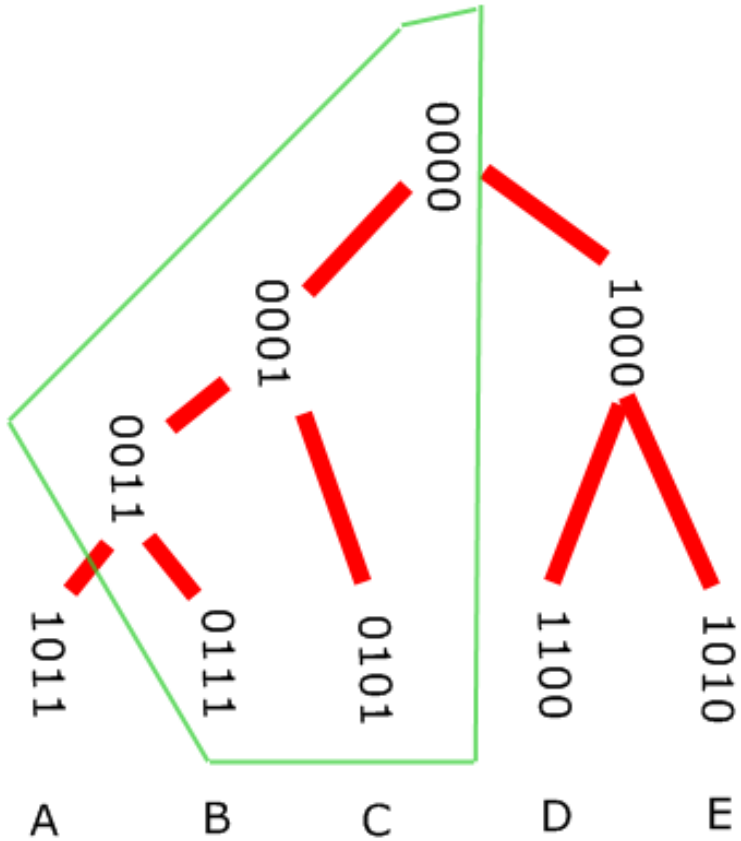
Phylogenetic trees

Example (Camin and Sokal)



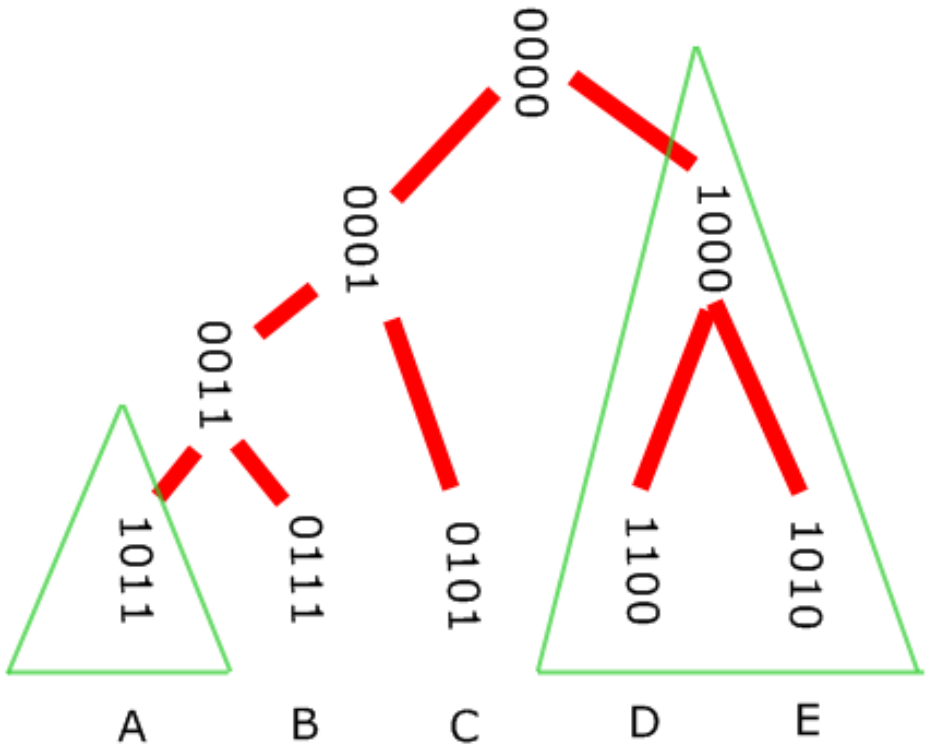
Phylogenetic trees

Example (Camin and Sokal)



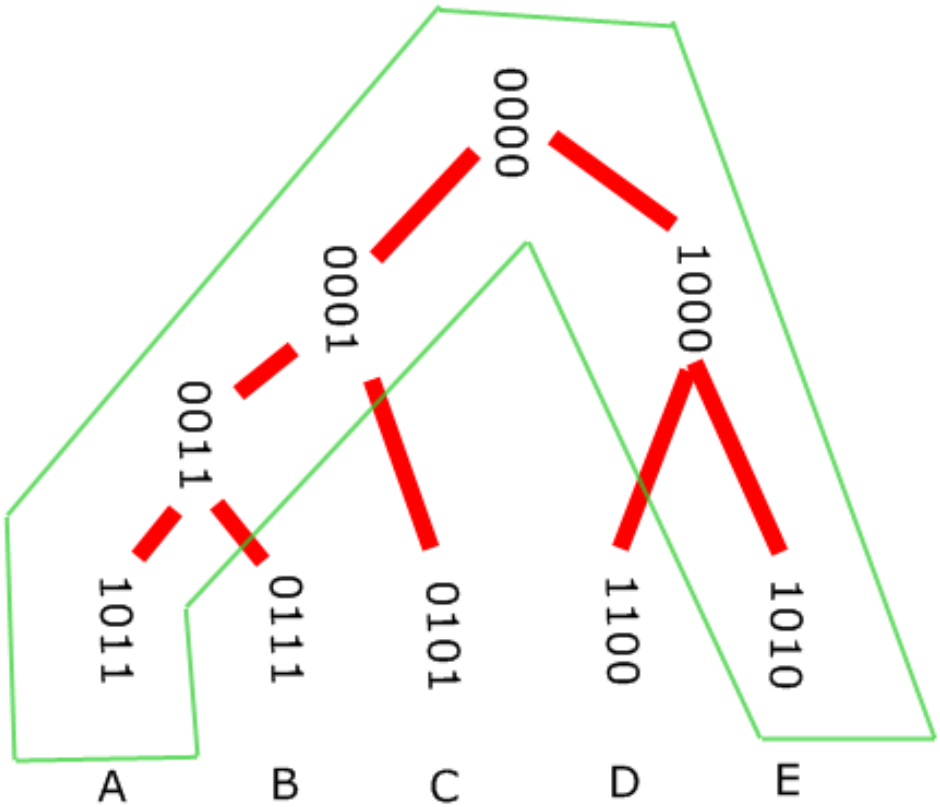
Phylogenetic trees

Example (Camin and Sokal)



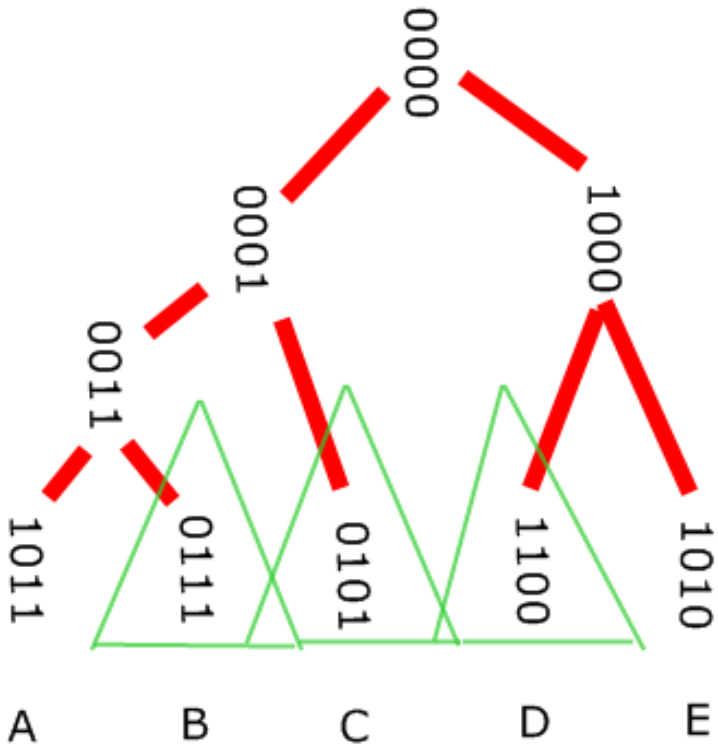
Phylogenetic trees

Example (Camin and Sokal)



Phylogenetic trees

Example (Camin and Sokal)



Phylogenetic trees

Another Problem

QCS/QDO (Qualitative Camin-Sokal/Dollo)

Given a set L of taxa, a set I of characters, and a number k, is there V s.t. $|V| = k$ and $L \subseteq V \subseteq \{0,1\}^n$ such that (V, E_V) is acyclic and connected (there is a path between each pair of nodes) and the induced tree is a phylogeny?

Phylogenetic trees

Another Problem

QCS/QDO (Qualitative Camin-Sokal/Dollo)

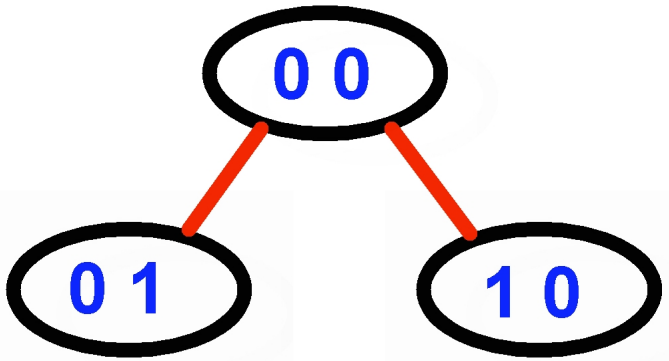
Given a set L of taxa, a set I of characters, and a number k, is there V s.t. $|V| = k$ and $L \subseteq V \subseteq \{0,1\}^n$ such that (V, E_V) is acyclic and connected (there is a path between each pair of nodes) and the induced tree is a phylogeny?

This problem (s) is NP-complete (Day, Johnson, Sankoff 1986). Reduction from vertex cover (see also the Haplotype Inference proof).

NP completeness of the k-Incompatibility Problem

An incompatible gadget

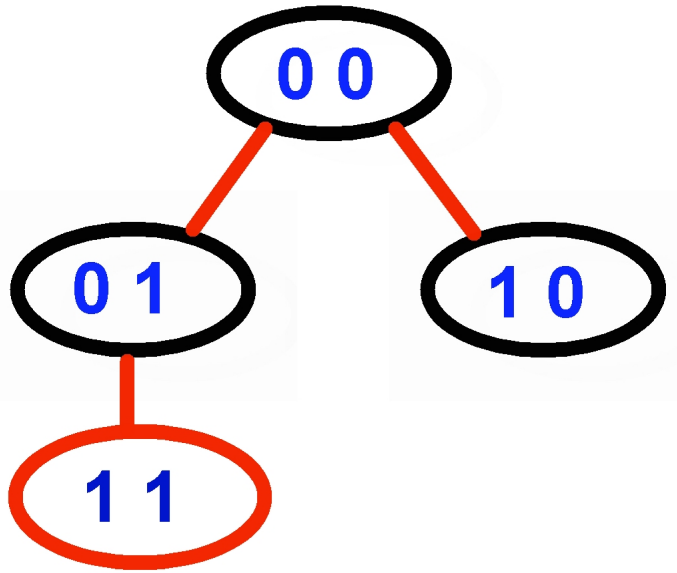
Consider two taxa and three/four characters.



NP completeness of the k-Incompatibility Problem

An incompatible gadget

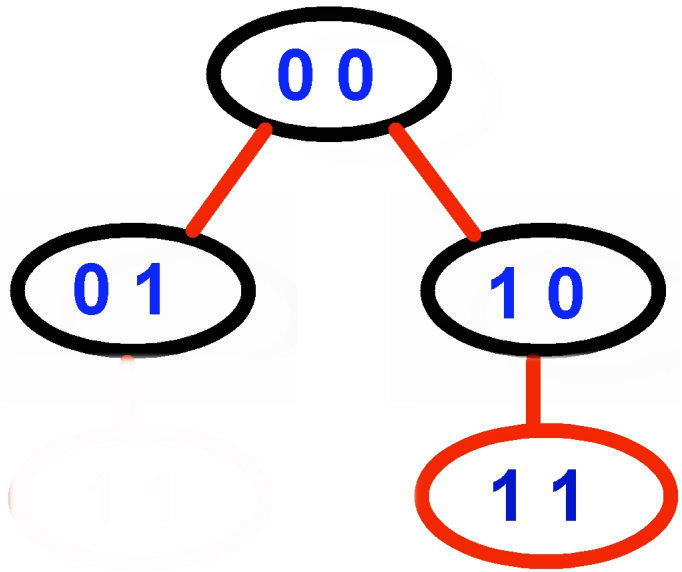
Consider two taxa and three/four characters.



NP completeness of the k-Incompatibility Problem

An incompatible gadget

Consider two taxa and three/four characters.

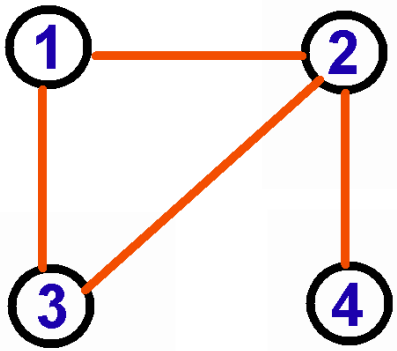


It can be proved that if the three characters contain pairs $(0,1),(1,0),(1,1)$ then the characters are not compatible.

NP completeness of the k-Incompatibility Problem

Reduction (sketch)

The property is exploited in the reduction of k-clique to k-Incompatibility.



$$|L| = 3^{\frac{|V|(|V|-1)}{2}}$$

$$|I| = |V|$$

Vertexes of a non-edge cannot stay both in a clique.

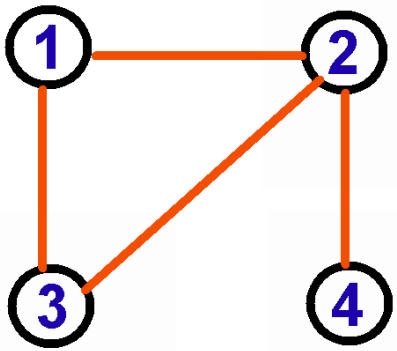
Add the incompatible gadget to all non-edges.

	(1,2)			(1,3)			(1,4)			(2,3)			(2,4)			(3,4)		
1							1	1	0									
2																		
3																1	1	0
4							0	1	1							0	1	1

NP completeness of the k-Incompatibility Problem

Reduction (sketch)

The property is exploited in the reduction of k-clique to k-Incompatibility.



$|L| = 3^{\frac{|V|(|V|-1)}{2}}$

$|I| = |V|$

Vertexes of a non-edge cannot stay both in a clique.

Add the incompatible gadget to all non-edges.

Set all other cells to 0

	(1,2)			(1,3)			(1,4)			(2,3)			(2,4)			(3,4)		
1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
4	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1

Encoding

Binary tree representation

- Input vector L of n elements (taxa) each of them characterized by a m-tuple of Boolean (character) values. E.g. $m = 3, n = 4$:

$$L = [[0, 1, 1], [1, 0, 0], [1, 1, 0], [1, 0, 1]]$$

- The Tree can be represented by a FD vector of t elements valued in $1, \ldots, t + 1$. Tree$[i] = j$ means that node i is a son of node j. For the root r, Tree$[r] = t + 1$.
- To force the tree to be binary, for $i = 1, \ldots, n$ we state

$$\text{count}(i, \mathit{Tree}, \leq, 2)$$

- Taxa are the leaves of the tree. To avoid symmetries, wlog,
  - ✓ leaves are nodes 1–n
  - ✓ Tree$[1] = n + 1$
  - ✓ Tree$[t] = t + 1$ (t is the root)
  - ✓ For $i, j \in \{1, \ldots, t\}$: $i < j \rightarrow$ Tree$[i] \leq$ Tree$[j]$

Encoding

Hypercube tree

- Each node of the tree is assigned a m-tuple of Boolean Values. This is stored in a vector Chars.
- Chars[1]–Chars[n] are assigned using the input L. Values for internal nodes must be computed.
- For $i < j$, if Tree[i] $= j$, the Hamming difference of the corresponding tuples is 1. Precisely:

$$\text{Tree}[i] = j \rightarrow \left(\sum_{\ell=1}^{m} |\text{Chars}[i][\ell] - \text{Chars}[j][\ell]| \right) = 1$$

Encoding

Hypercube tree

- Each node of the tree is assigned a m-tuple of Boolean Values. This is stored in a vector Chars.
- Chars[1]–Chars[n] are assigned using the input L. Values for internal nodes must be computed.
- For $i < j$, if Tree[i] $= j$, the Hamming difference of the corresponding tuples is 1. Precisely:

$$\text{Tree}[i] = j \rightarrow \left(\sum_{\ell=1}^{m} |\text{Chars}[i][\ell] - \text{Chars}[j][\ell]| \right) = 1$$

- Actually, we can either relax the above constraint to ≤ 1 (see e.g. hand/father example, italian and spanish) or (alternatively)
- Add the redundant constraint

$$\text{AllDifferentTuples}(\text{Chars})$$

Encoding

k-incompatibility

- We need to state that a character changes (actually, increases) in at most one node. This makes the tree compatible with that character.
- Let Comp be a vector of m elements (one per character).
- For $i < j$, let $F_{i,j} = 1$ if Tree$[i] = j$, $F_{i,j} = 0$ otherwise.
- Then, for $\ell = 1, \ldots, m$ (and $i, j = 1, \ldots, n$:

$$\text{Comp}[\ell] = \sum_{i<j} F_{i,j}(\text{Chars}[i][\ell] - \text{Chars}[j][\ell])$$

- Basically, after variable instantiation, Comp$[\ell]$ will contain the number of changes of character ℓ in the tree.
- The number of values different from 1 and 0 in Comp is forced to be less than or equal to k.
- See the program `phylo.pl`. Call the goal as: `:-phylo(List,0),phylo_decision(List,1,9,Tree).`

(Some) References

- Day, W.H.E., Johnson D.S., Sankoff, D. The Computational complexity of Inferring Rooted Phylogenies by Parsimony. Math. Biosciences 81:33–42, 1986.
- Day, W.H.E., Sankoff, D. Computational complexity of Inferring Phylogenies by Compatibility. *Systematic Zoology* 35(2):224–229, 1986.
- Erdem E., Lifschitz V., Nakhleh L., Ringe D. Reconstructing the Evolutionary History of Indo-European Languages Using Answer Set Programming. PADL 2003: 160-176.
- Thomas Schiex et al. Papers on *complex pedigree reconstructions* using weighted constraint satisfaction. In WCB 05, WCB 06, WCB 07.
- Moore N.C.A., and Prosser P. The Ultrametric Constraint and its Application to Phylogenetics. *J. Artif. Intell. Res.* 32:901–938, 2008 (also in WCB 06)
- Erdem E. Applications of Answer Set Programming in Phylogenetic Systematics MG65, LNCS 6565, 2011.
- Le Tiep, Nguyen Hieu, Pontelli Enrico, and Cao Son Tran. ASP at Work: An ASP Implementation of PhyloWS. ICLP 2012, LIPICS vol 17. (also in WCB 12)