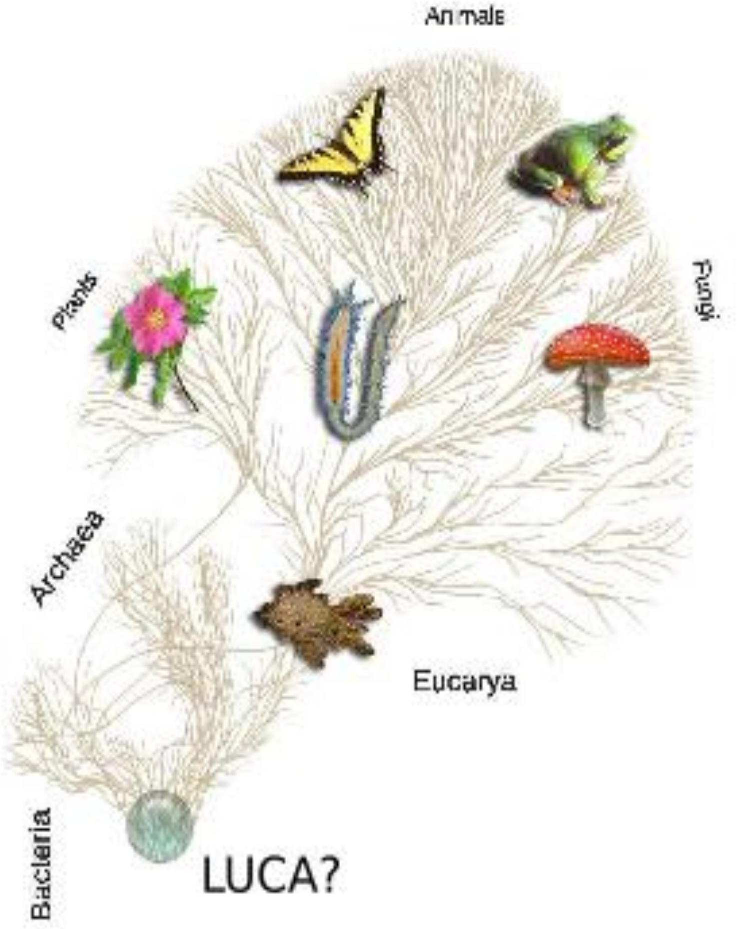
Tree of life

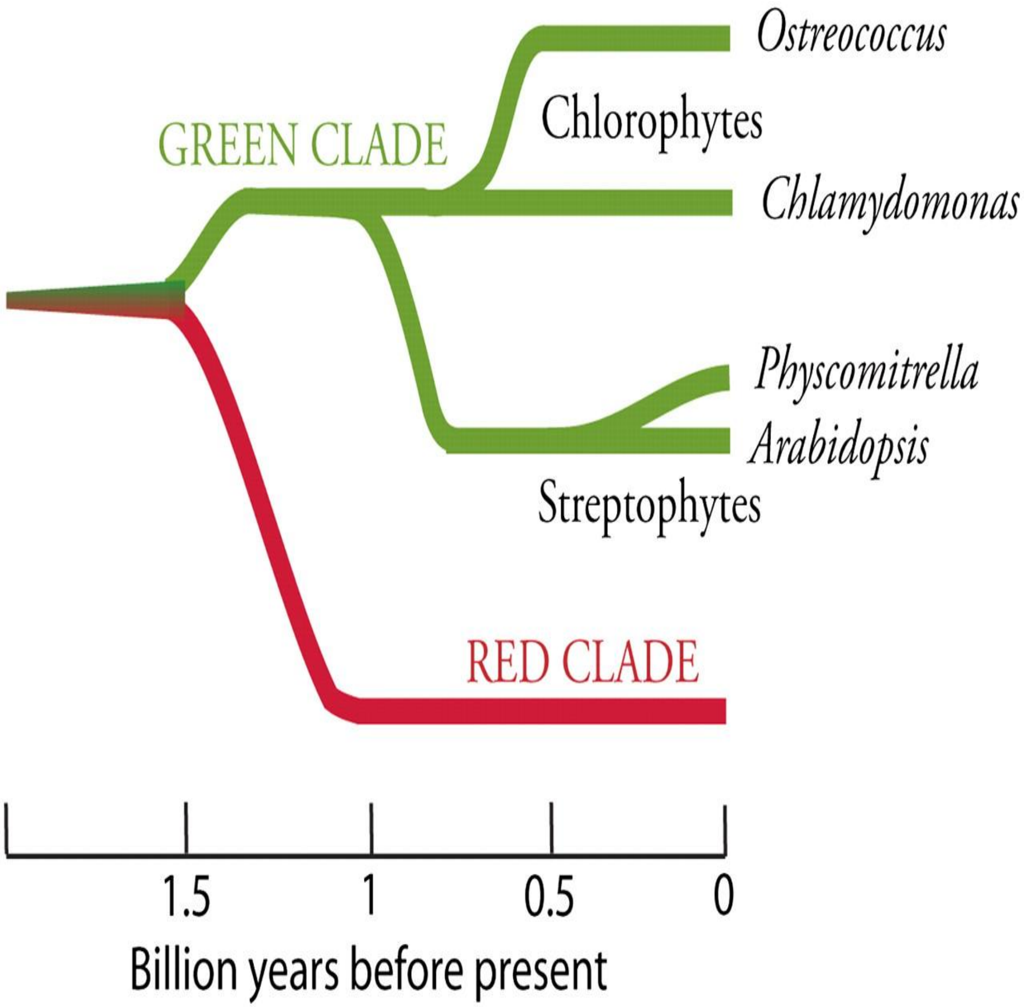
Inferring species phylogenies from entire proteomes

Supervisors :
Arnold Kuzniar arnold.kuzniar@unil.ch
Hannes Schabauer hannes.schabauer@unil.ch

Students :
Didar Tolou didar.tolou@unil.ch
Marie Gallot Lavallée marie.gallotlavallee@unil.ch
Rachel Barman rachel.barman@unil.ch

27 mai 2011





(Peers *et al.*, 2008)

Project's goals

Infer species tree(s) using entire proteomes of 7 plants rather than only a few protein families

Compare our results to literature and taxonomy DB

Plants we are working on

Species	Taxonomy	Family	Proteome size	Uni/Multi cellular
Broad leef tree	Populus trichocarpa	Salicaceae Dicotyledon	58036	Multicellular
Grape	Vitis vinifera	Vitaceae Dicotyledon	54411	Multicellular
Moss	Physcomitrella patens	Funariaceae	36067	Multicellular
Rockcress	Arabidopsis thaliana	Brassicaceae Dicotyledon	32816	Multicellular
Rice	Oryza sativa	Poaceae Monocotyledon	27006	Multicellular
Green Alga	Chlamydomonas reinhardtii	Chlamydomonadaceae	14489	Unicellular
Other green alga	Ostreococcus lucimarinus	Prasinophyceae	7603	Unicellular

Why is this project interesting ?

- We got familiar with some bioinformatic tools
- We read scientific reviews and got familiar with domain-specific terminology
- We learned programming and improved our math
- We ploted trees and made clusters

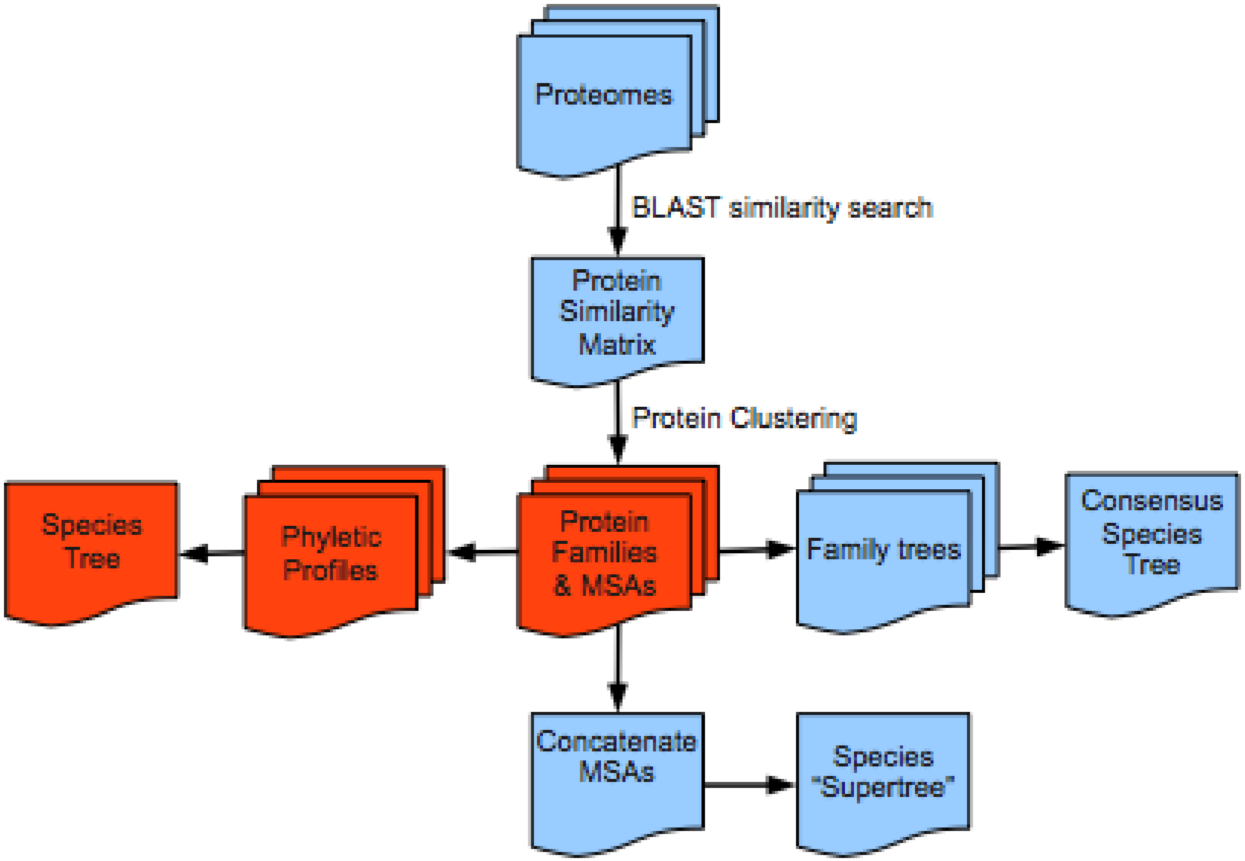
Sequence homology

Homology: two sequences sharing enough sequences similarity to infer the hypothesis that they descend from a single common ancestor.

Paralogy: homologous genes relating through gene duplication from one single ancestral gene.

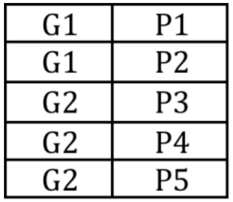
Orthology: homologous genes relating through speciation from one single ancestral gene.

Methods


Proteomes
BLAST similarity search
Protein Similarity Matrix
Protein Clustering
Species Tree
Phyletic Profiles
Protein Families & MSAs
Family trees
Consensus Species Tree
Concatenate MSAs
Species "Supertree"

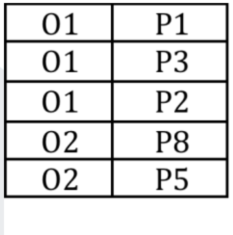
Phyletic profile

File 1



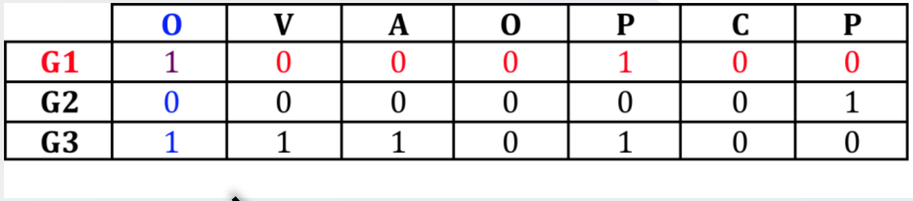
G1	P1
G1	P2
G2	P3
G2	P4
G2	P5

File 2

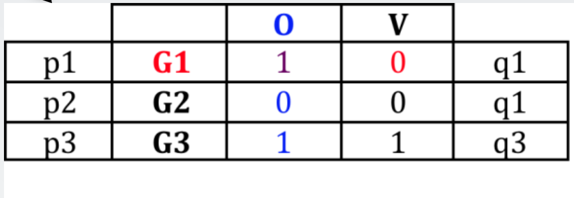


O1	P1
O1	P3
O1	P2
O2	P8
O2	P5

Phyletic profile

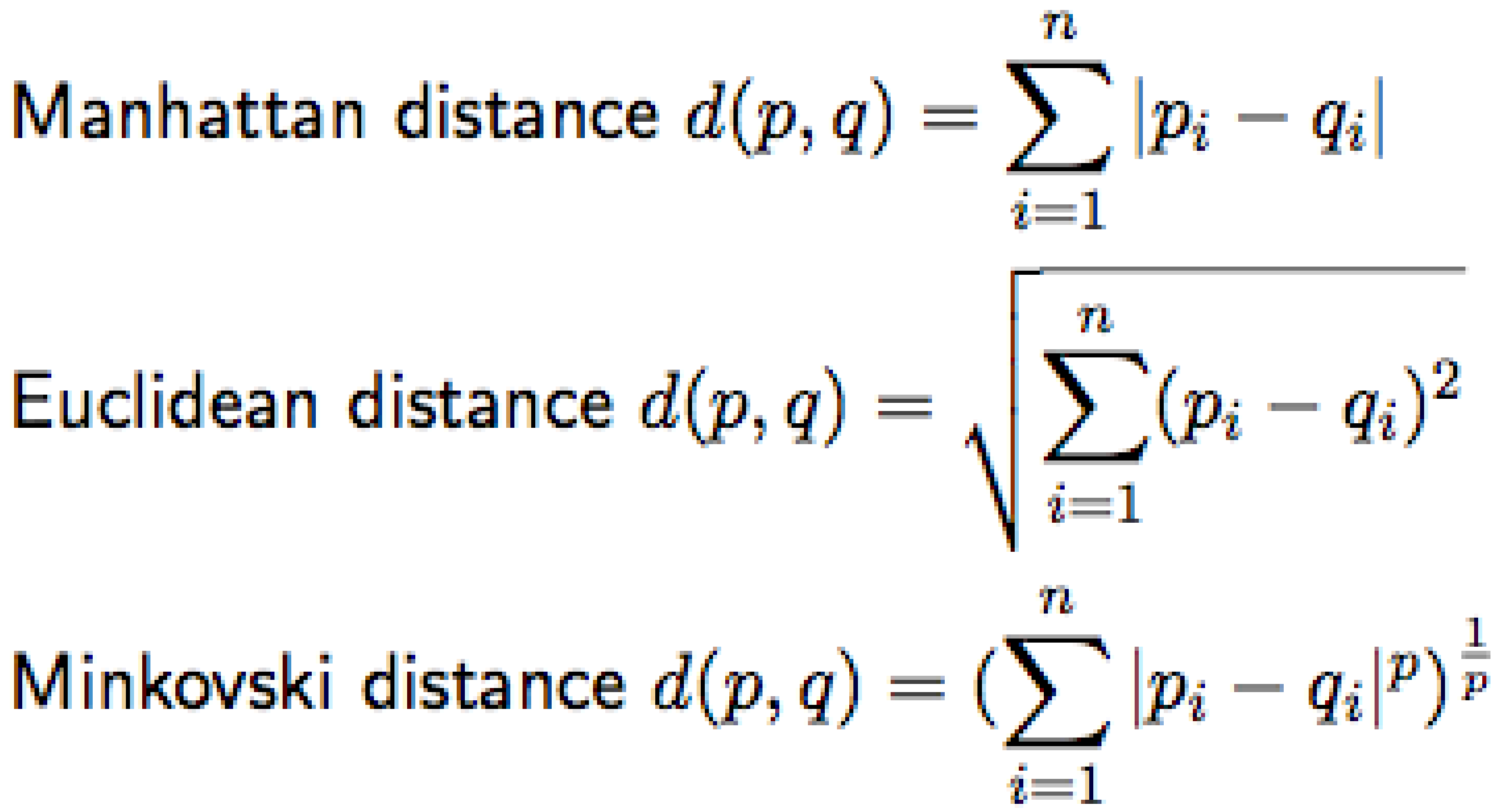


	O	V	A	O	P	C	P
G1	1	0	0	0	1	0	0
G2	0	0	0	0	0	0	1
G3	1	1	1	0	1	0	0



		O	V	
p1	G1	1	0	q1
p2	G2	0	0	q1
p3	G3	1	1	q3

Distance formulas



Manhattan distance $d(p,q) = \sum_{i=1}^{n} |p_i - q_i|$

Euclidean distance $d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$

Minkovski distance $d(p,q) = (\sum_{i=1}^{n} |p_i - q_i|^p)^{\frac{1}{p}}$

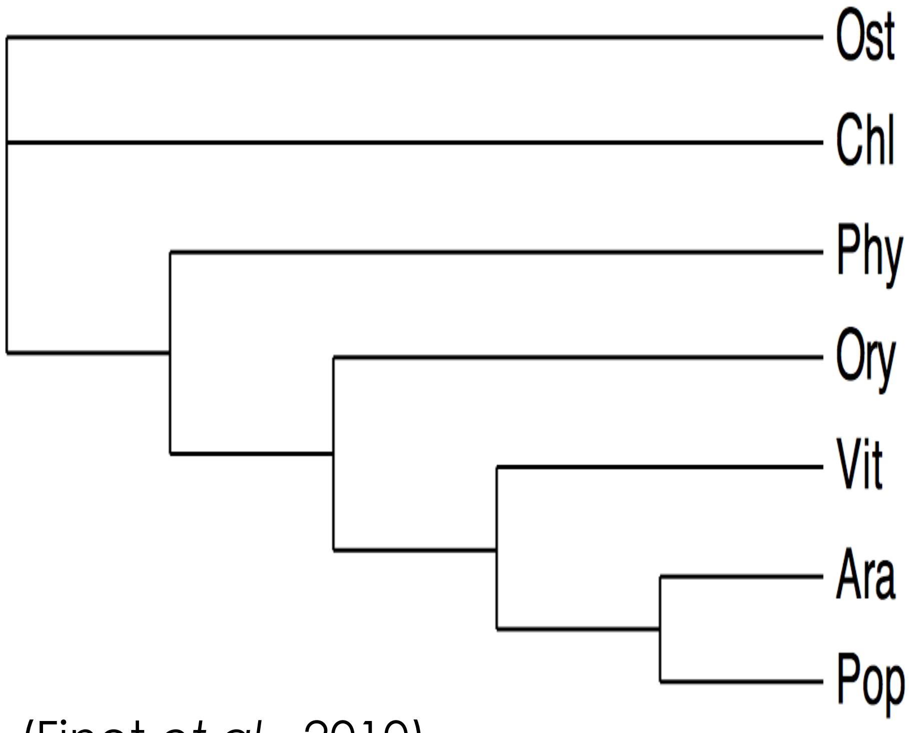
Final R commands

```
>table=scan("phyl3",list(col1="",col2="",col3="",col4=""),sep="\t")

>phyl3s=strsplit(table$col3,split="")

>bitvector=as.integer(unlist(phyl3s))

>m=matrix(bitvector,nrow=7,byrow=FALSE)

>rownames(m)<-c("Arabidopsis thaliana","Chlamydomonas
    reinhardtii","Oryza sativa","Ostreococcus lucimarinus","Physcomitrella
    patens","Populus trichocarpa","Vitis vinifera")

>dmin1=dist(m,method="minkowski",p=1)

>plot(hclust(dmin1,method="average"))
```

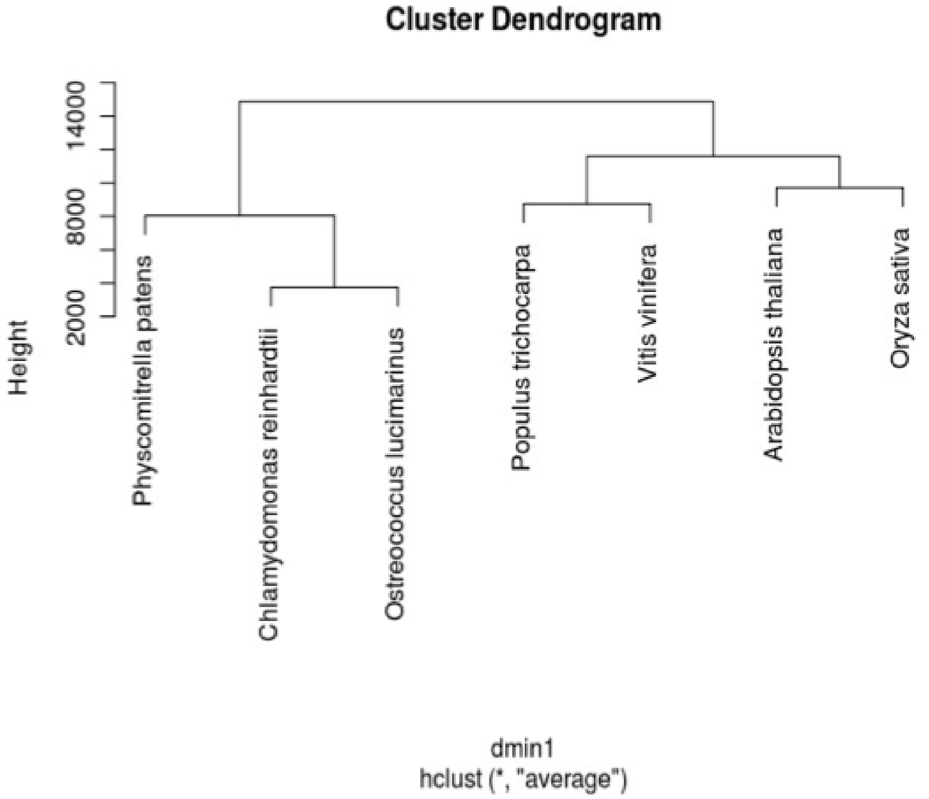
Conclusions

- Tree comparisons
- What does our tree reflect?
- Is there a tree better than another ?
- What could we change in our method?
- Perspectives

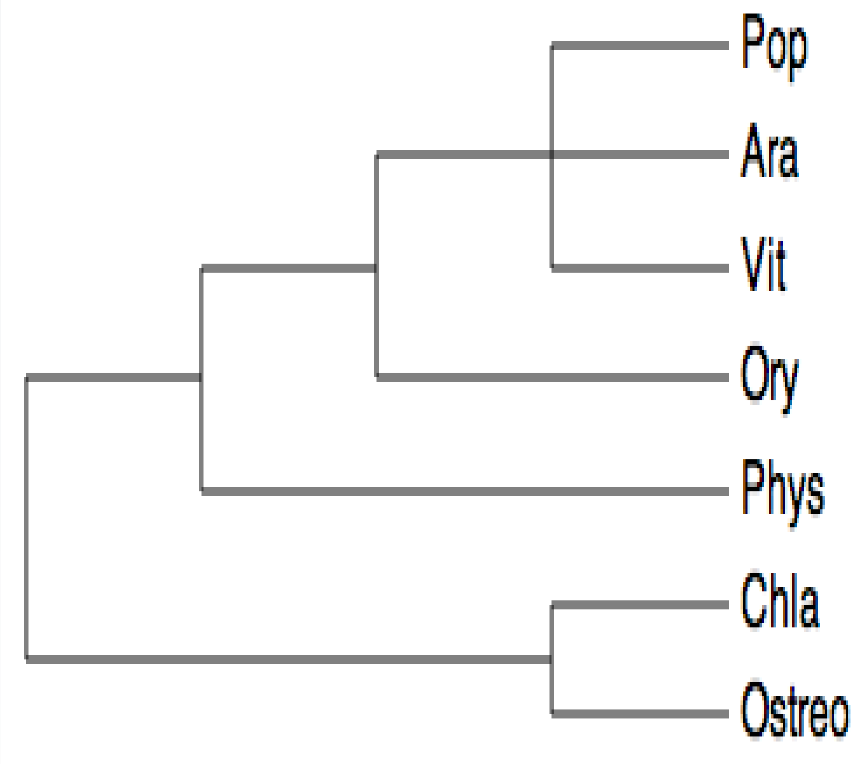
Literature / our tree

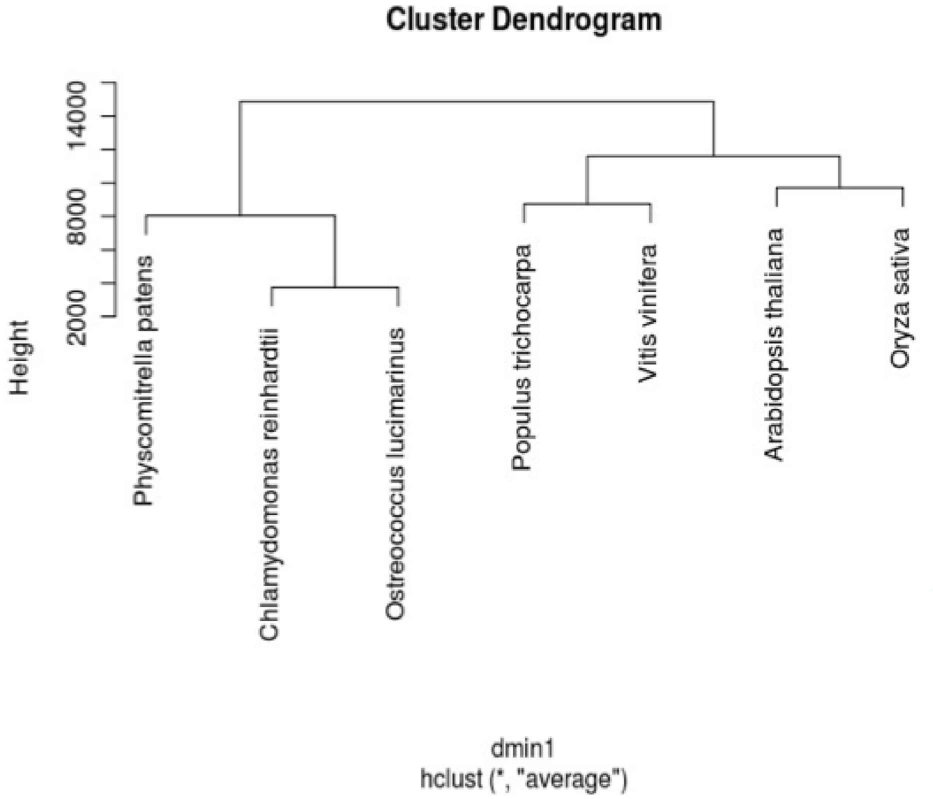


(Finet *et al.*, 2010)



Taxonomy DB / our tree





Conclusions

- Tree comparisons
- What does our tree reflect?
- Is there a tree better than another ?
- What could we change in our method?
- Perspectives

Conclusions

- Tree comparisons
- What does our tree reflect?
- Is there a tree better than another ?
- What could we change in our method?
- Perspectives

Thank you for your attention

Supervisors :
Arnold Kuzniar arnold.kuzniar@unil.ch
Hannes Schabauer hannes.schabauer@unil.ch

Students :
Didar Tolou didar.tolou@unil.ch
Marie Gallot Lavallée marie.gallotlavallee@unil.ch
Rachel Barman rachel.barman@unil.ch