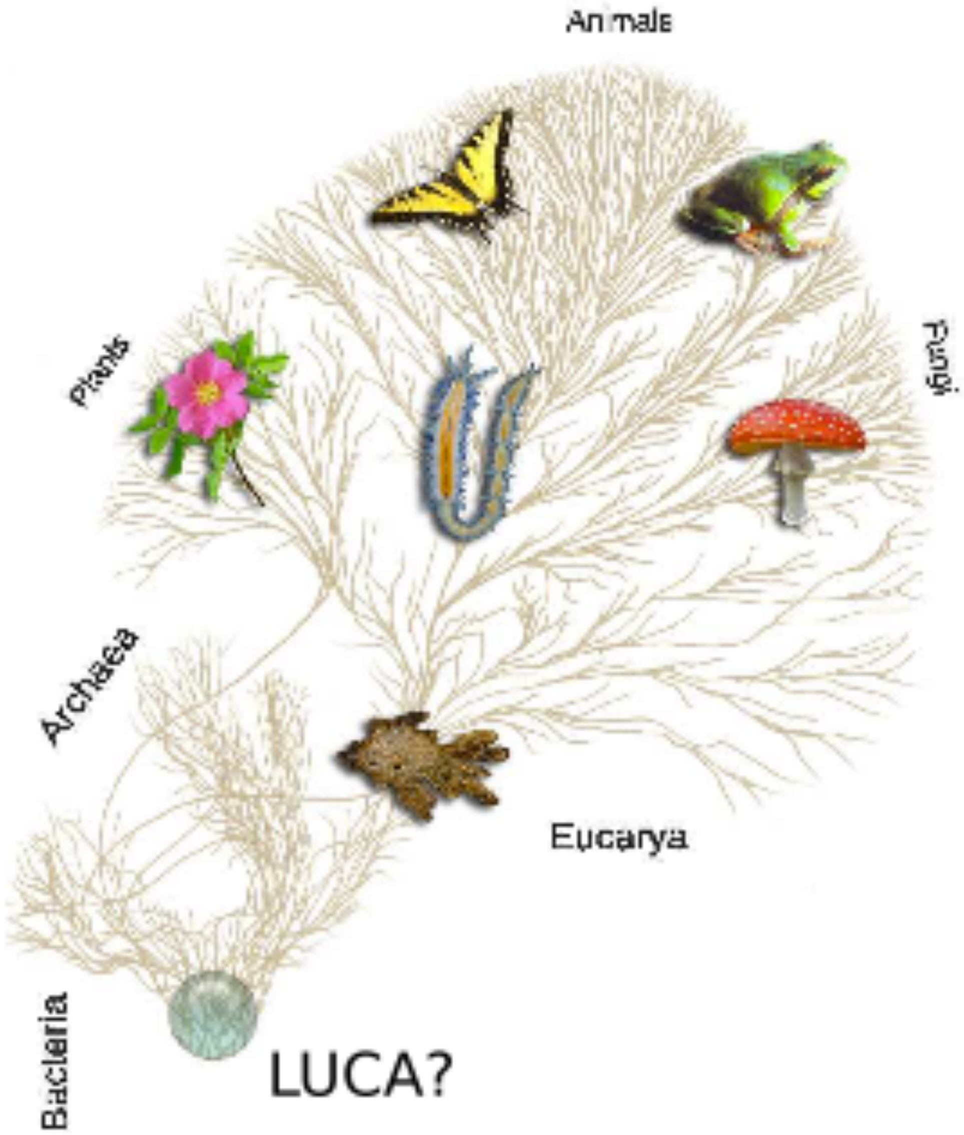
Solving biological problems that require math

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


Animals
Plants
Fungi
Archaea
Eucarya
Bacteria
LUCA?

Project's goals

Infer species tree(s) using entire proteomes of 7 plants

Compare our results to single-gene and «supertree» approaches from literature

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 get familiar with some bioinformatic tools : BLAST, databases (Uniprot)
- We write simple scripts in Linux: bash, Perl, R
- We read scientific papers and get familiar with domain-specific terminology
- We improve our math and writing skills

Some terminology

Homology: a hypothesis about a common ancestry of two genes (proteins) having sufficiently similar sequences

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

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

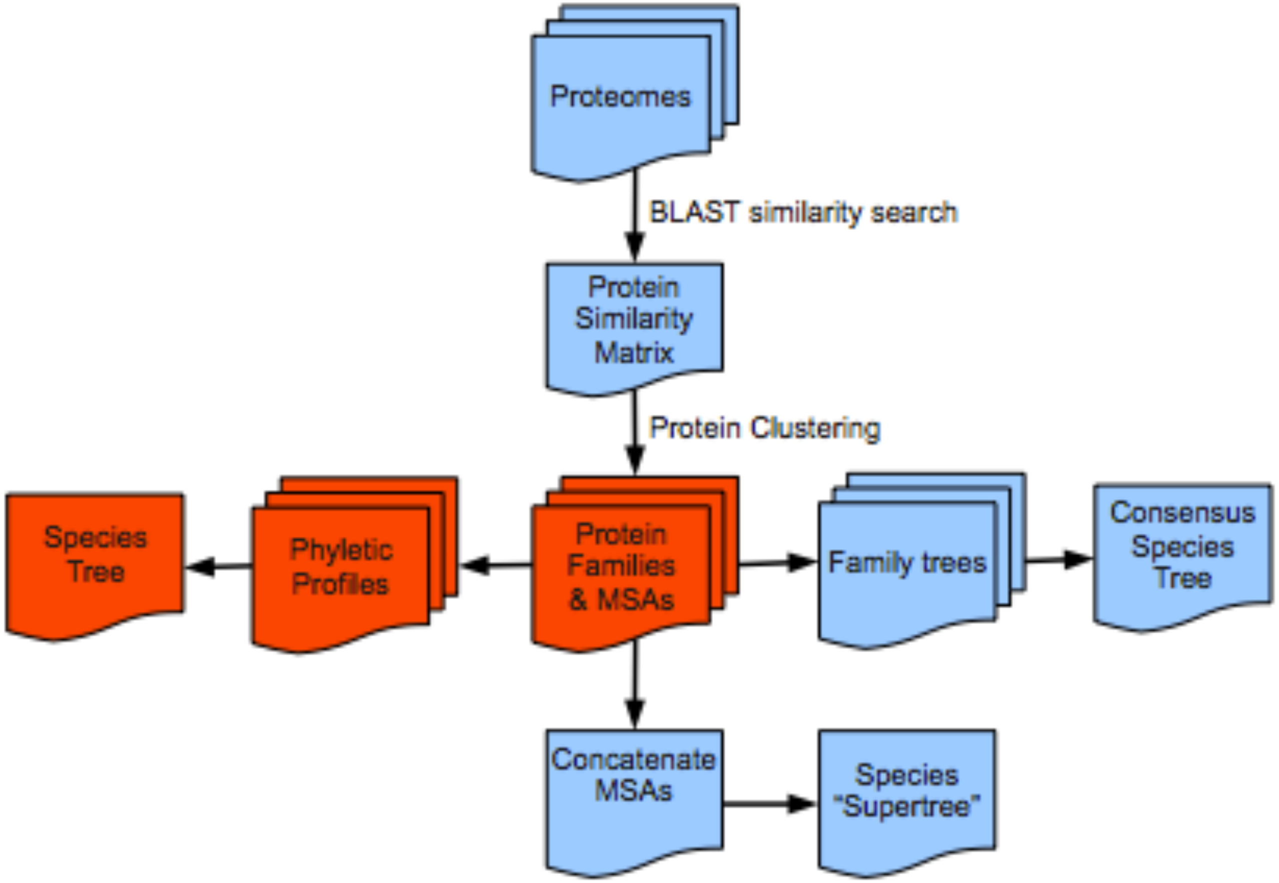
Methods

- We learn to use BLAST to detect homology between protein sequences.

- We use Perl for data manipulation.

- We will construct phylogenetic (species) trees based on distances between phyletic profiles.

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"

BLAST alignment scores

- Amino acid substitution matrices (e.g. BLOSUM)
- Raw score

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

- Bits score

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

- E-value

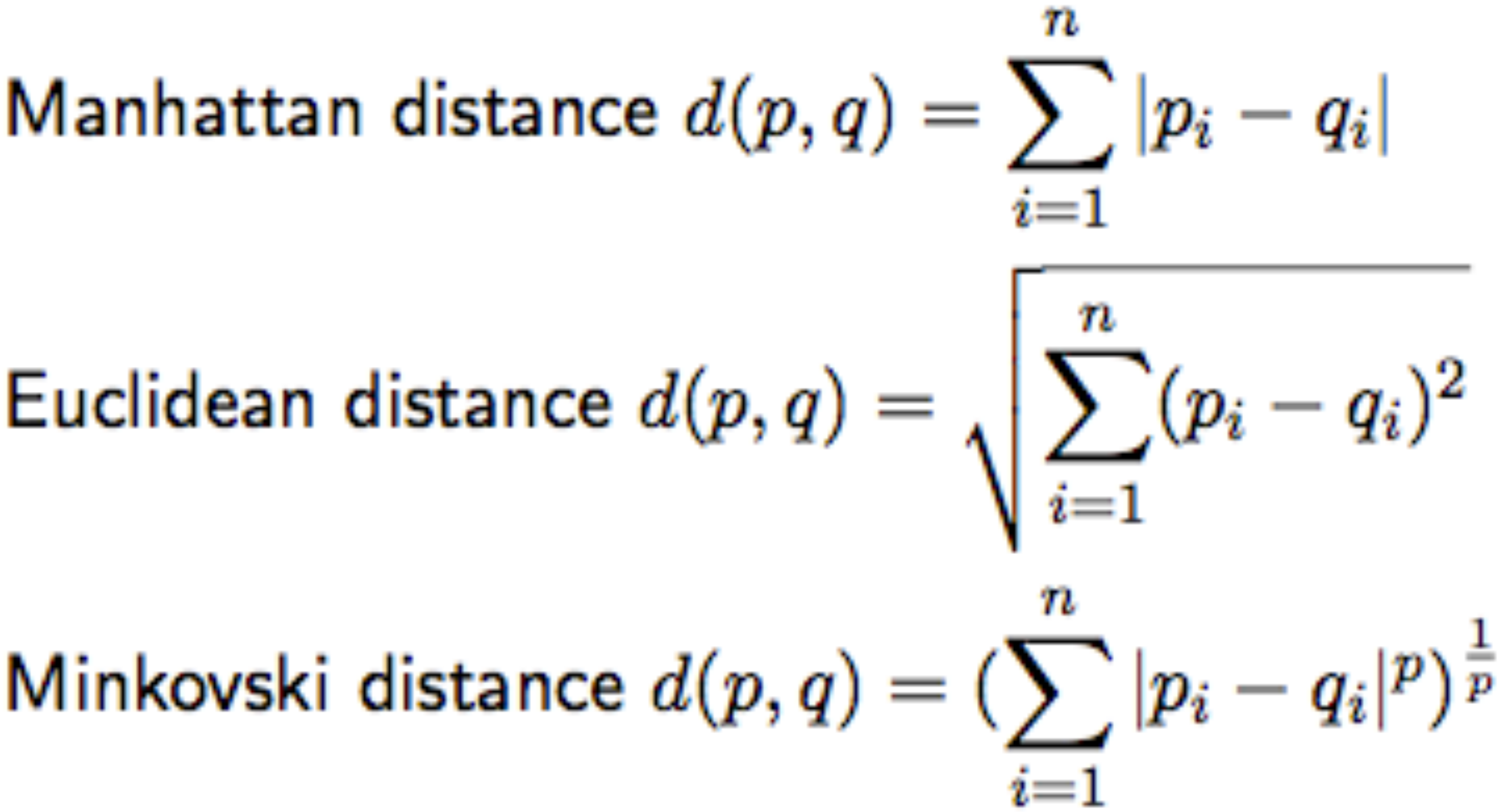
$$\begin{aligned} E &= mn \cdot Pval \\ &= Kmne^{-\lambda S} \end{aligned}$$

Example Perl script

```
#!/usr/bin/perl

use strict;
use warnings;

my %proteomes; #stores the species-to-proteinIDs key-value pairs.

my $file_name = $ARGV[0]; #store the input file name.

#Error handling is important: check if $file_name not empty

if($file_name) {#TRUE
   print "INPUT FILE: $file_name\n";
}else{#FALSE
   print "Usage: $0 [FILE]\n";
   exit 1; #exit with failure if $file_name is empty.
}

open(FILE,$file_name) or die "Cannot open $file_name file.\n";
```

```
while(<FILE>){
   my $line = $_;
   my ($protID,$species)=split(/\s+/,$line); #to tell Perl to split $line
on one or
   #more spaces

   push(@{$proteomes{$protID}}, $species);

}
close FILE;

#Print the content of the hash table onto STDOUT

foreach my $key(keys %proteomes) {
   my $value = $proteomes{$key}; # $value refers now to an
anonymous array of protein IDs NOT a single protein ID (scalar)!!!
   my $size = scalar(@$value); # scalar() returns the size of an array
   my $proteinIDs = join(", ", @$value); # join() concatenates the
array elements by a comma into a string
   print "$key\t$size\t$proteinIDs\n";
}
```

Distances between vectors



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

Challenges

- Programming !!!

- To understand the scoring of BLAST alignments

Suggestions

Give a dedicated course on programming (bash, Perl)

Find a room with a project

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