Detection of Orthologs

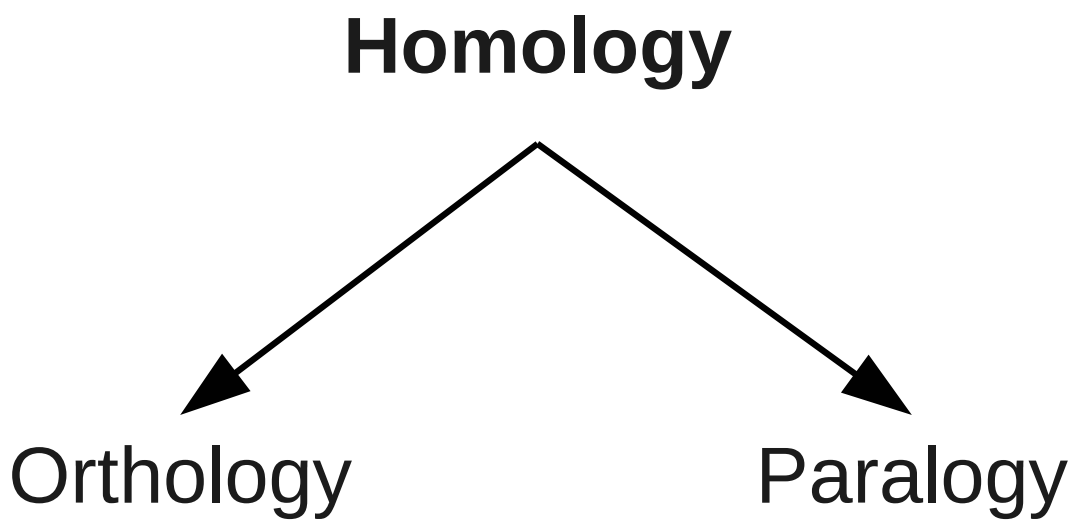
Marcus Lechner

University of Marburg

Bled, 2010-02-20

Table of contents

Definitions



Homologous genes

- have derived from a common ancestor

Definitions

Orthologous genes

- evolved by speciation
- thought to have a similar function

Definitions

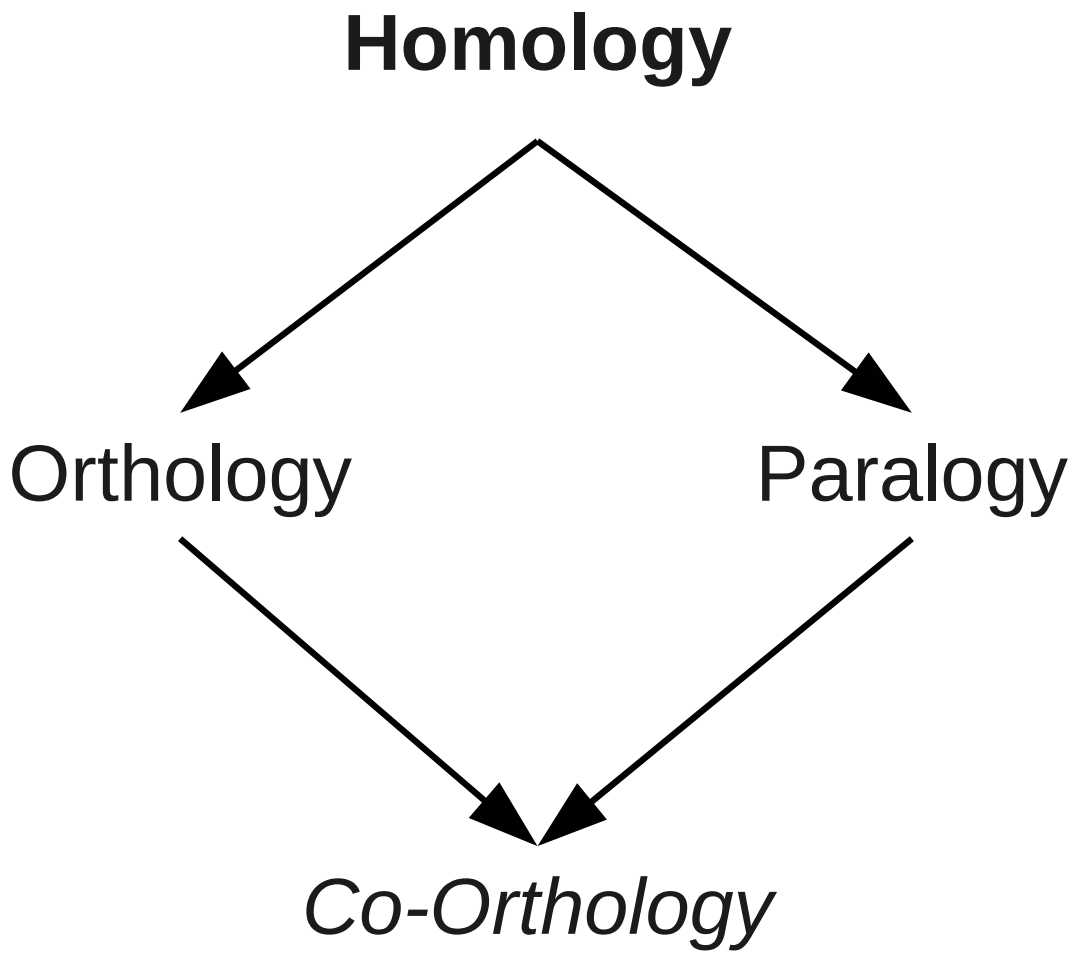
Orthologous genes

- evolved by speciation
- thought to have a similar function

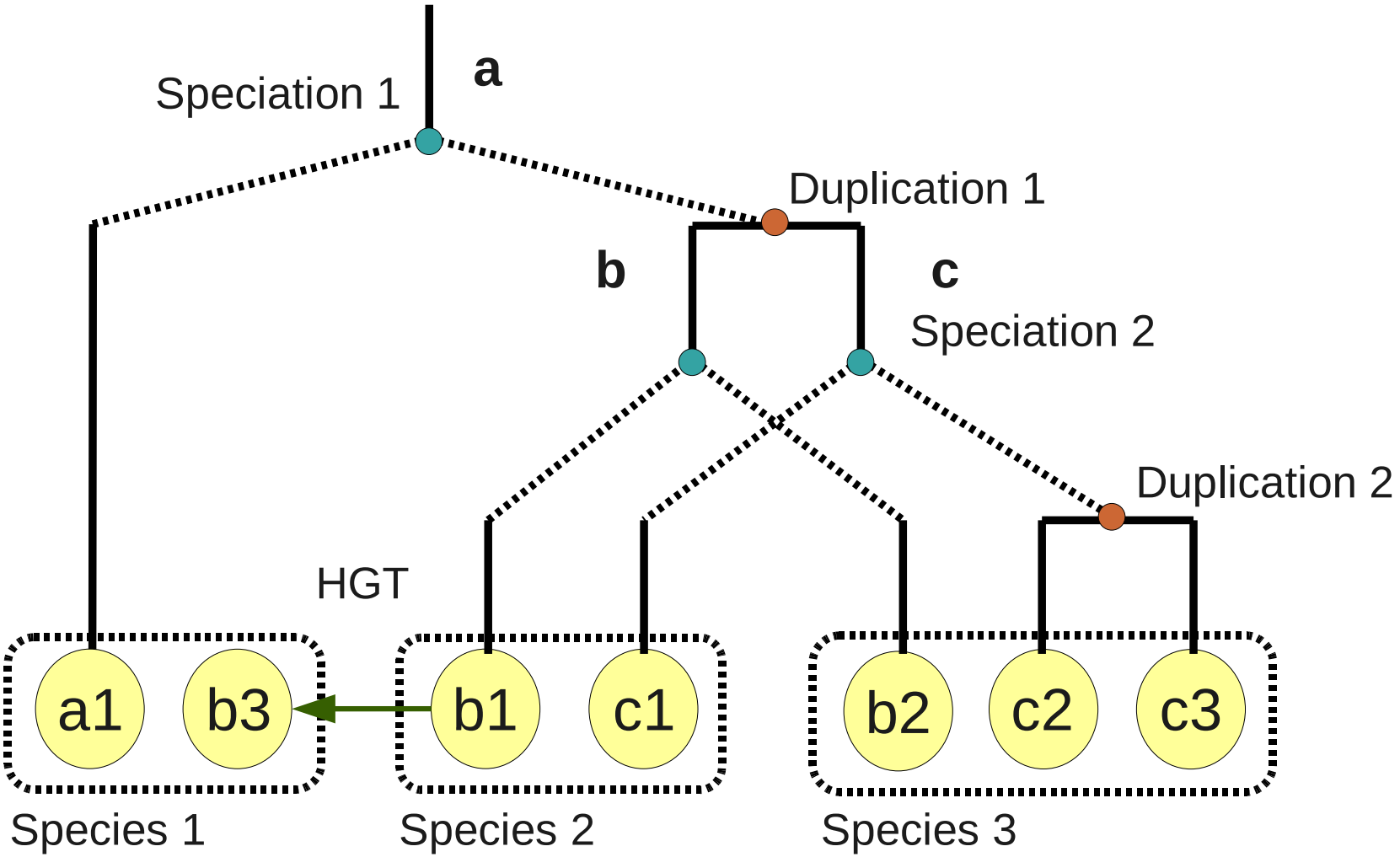
Paralogous genes

- homologous genes within the same species
- thought to have a related function (neo-/subfunctionalization)

Definitions



An example


Speciation 1
a
Duplication 1
b
c
Speciation 2
Duplication 2
HGT
a1
b3
b1
c1
b2
c2
c3
Species 1
Species 2
Species 3

Applications

Some examples

- gene function prediction
- biological pathways
- detection of functional site
- detection of putative drug targets
- phylogeny
- protein evolution

The Proteinortho approach

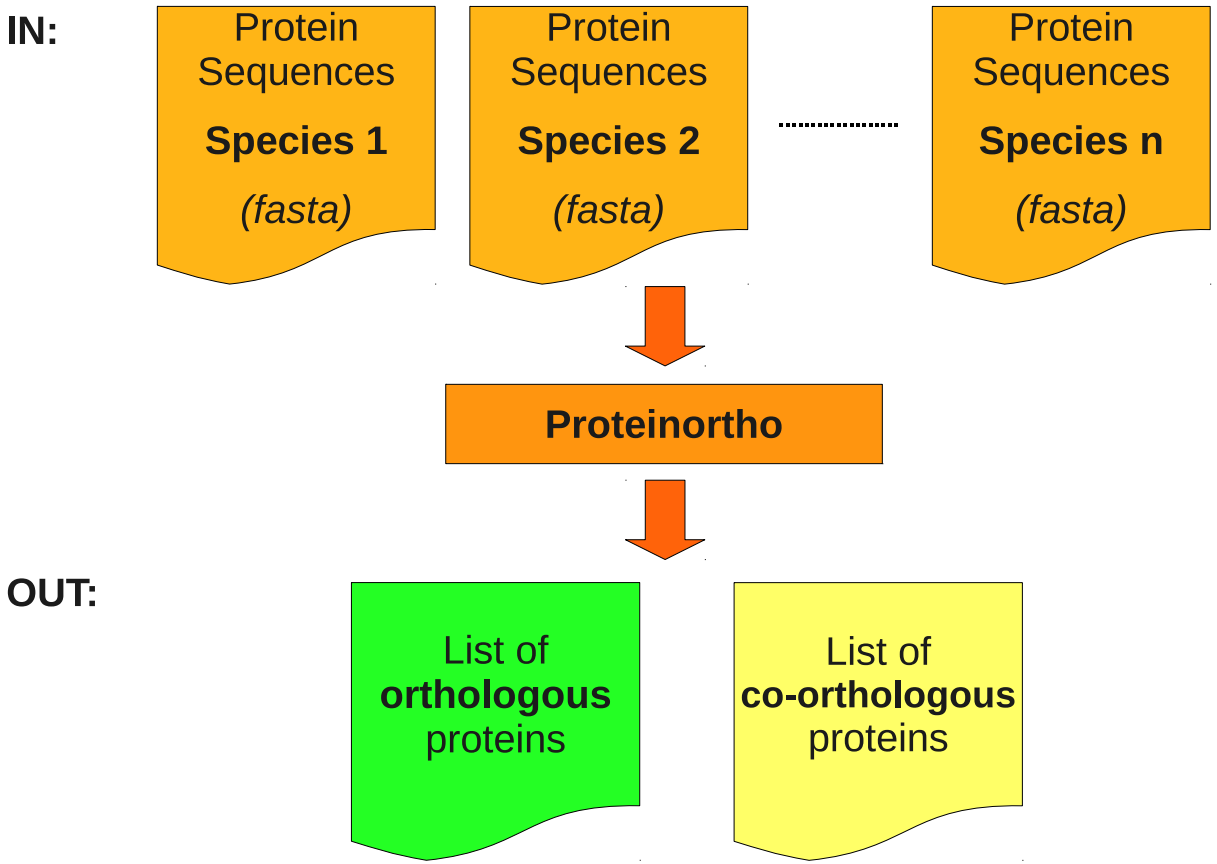
Background

- proteins arose from the same ancestor + similar function
  $\Rightarrow$ similar sequence
- look for similar sequences
  $\Rightarrow$ get isofunctional orthologs (this is at least the best guess)

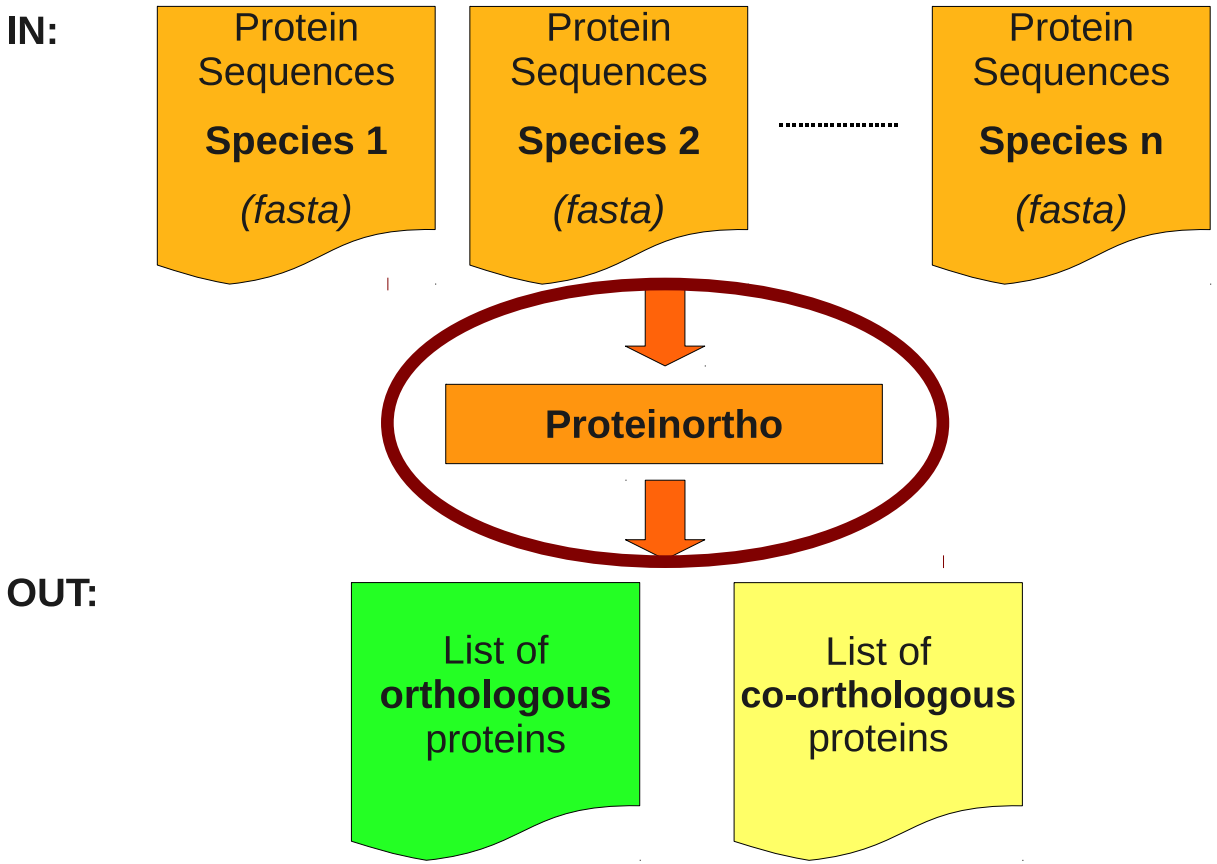
Methods

- adaptive reciprocal best blast hits
- filtering steps
- detection of connected components

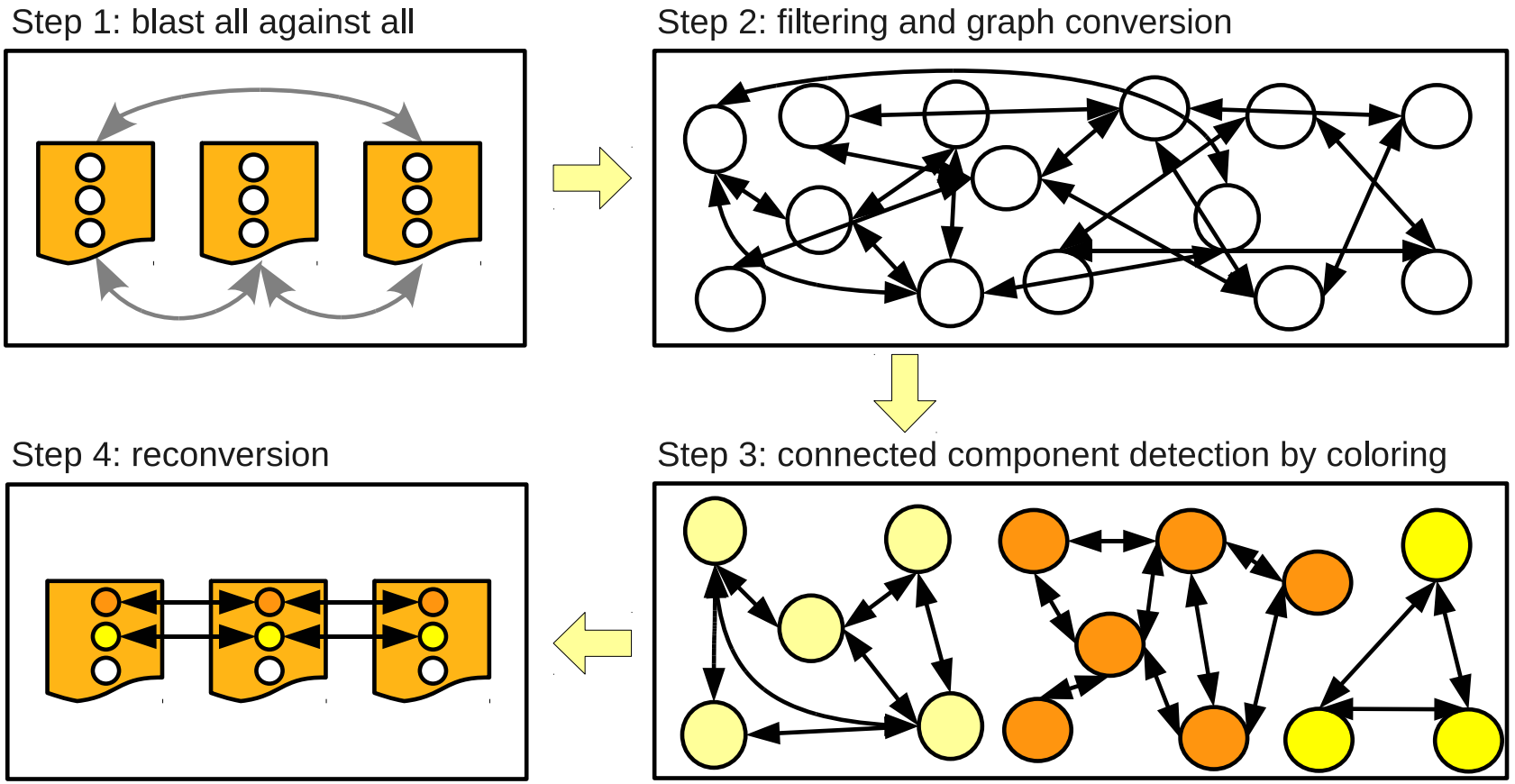
What do we need? What do we get?



What do we need? What do we get?



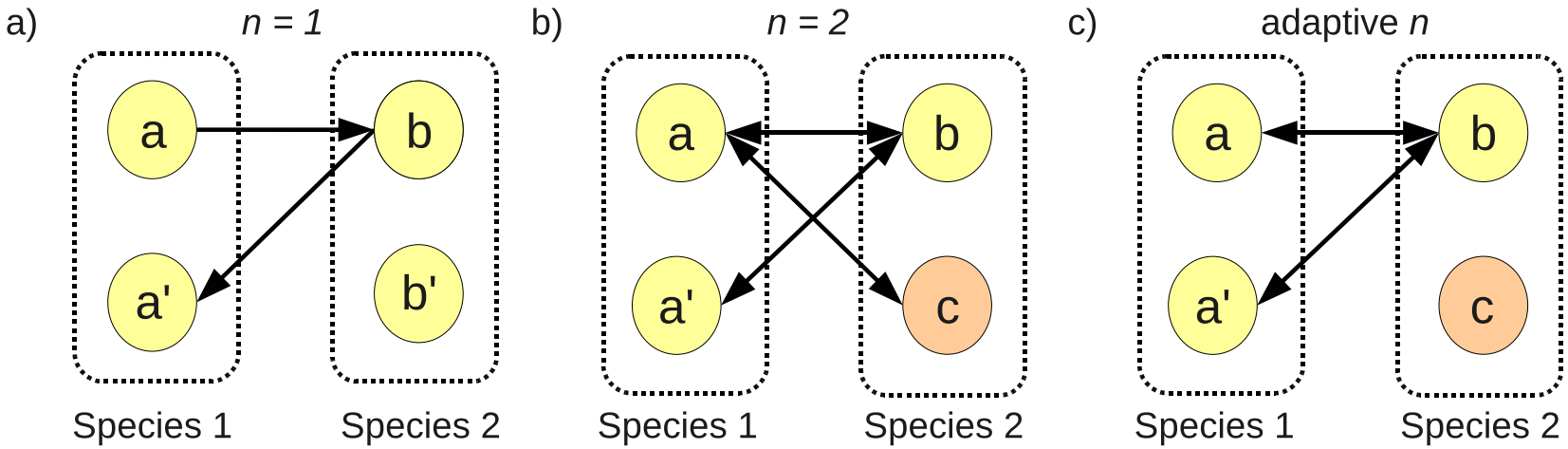
Workflow



1) Reciprocal blasts 2) Transformation into graph representation
3) Coloring and decomposition
4) Reconversion and mapping to species with encoded proteins

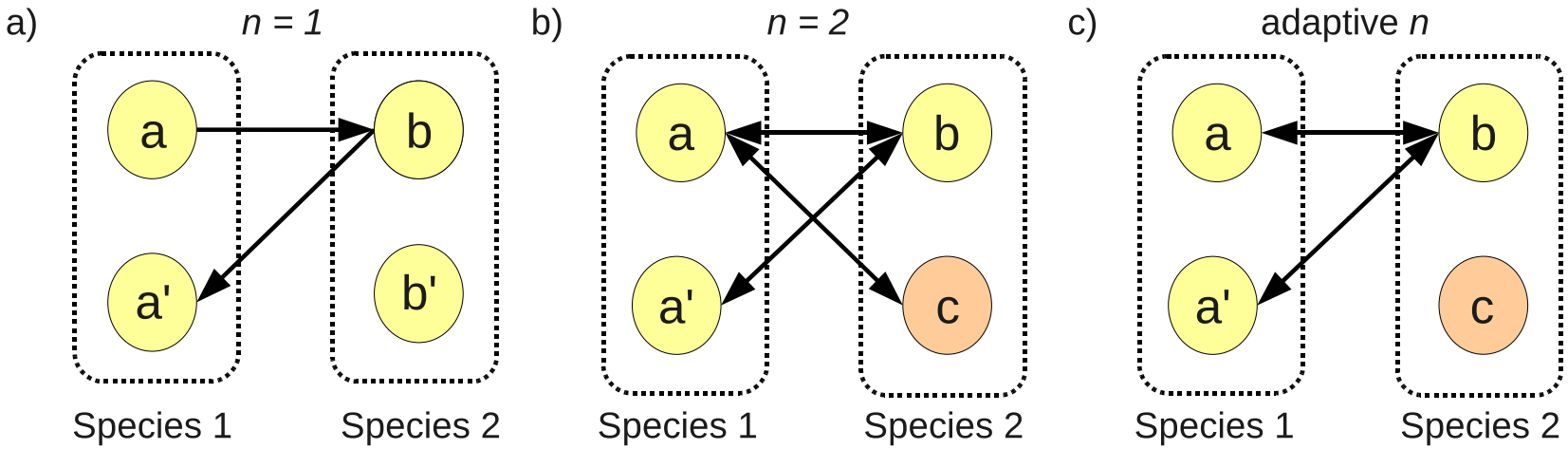
Features

Adaptive best blast hit



Features

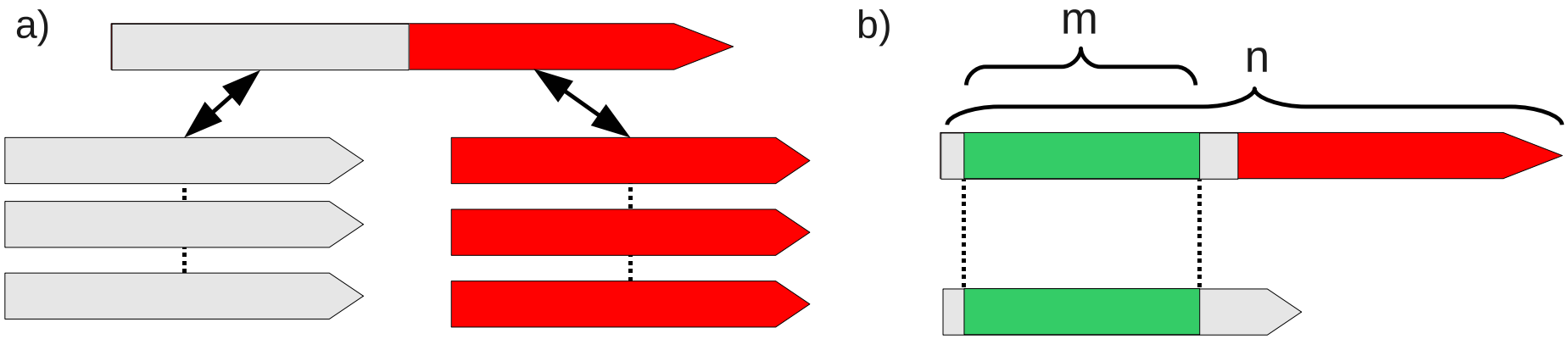
Adaptive best blast hit



$$s(candidate) = \frac{best + candidate}{best} - 1 < 0.95$$

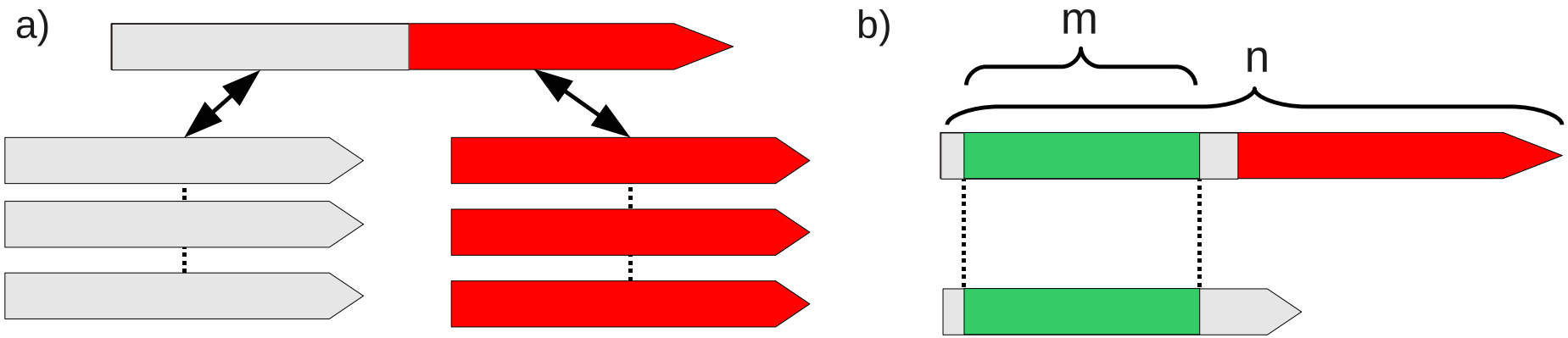
Features

Fusion and fission of genes



Features

Fusion and fission of genes



$$n < 2m$$

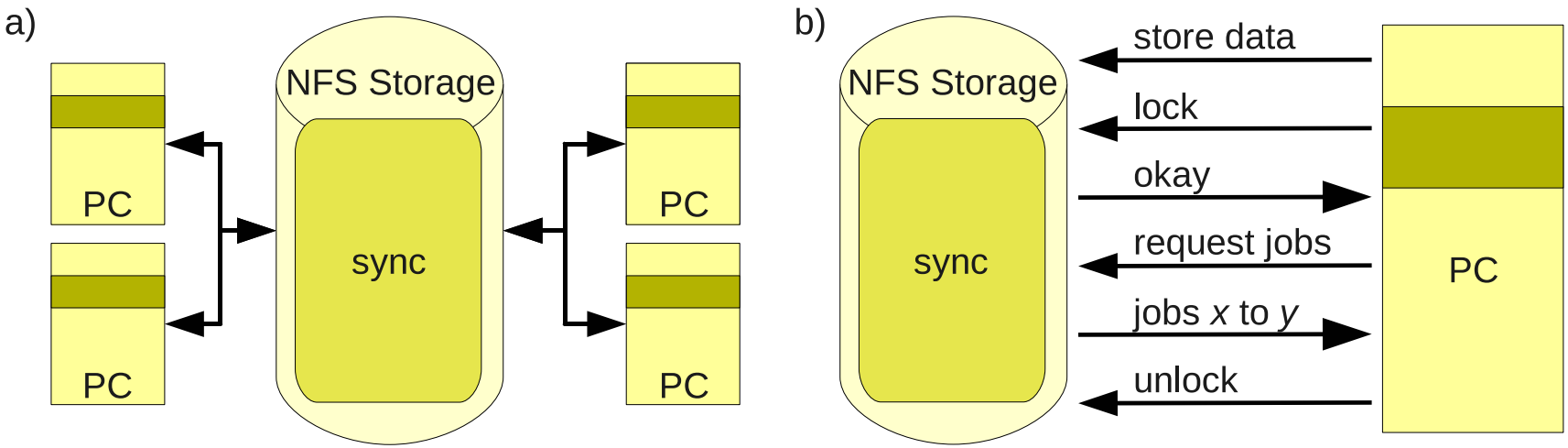
Features

More general

- detection of (co-)orthologous protein coding genes
- designed for high-throughput
- behaves nicely in memory consumption
- capable of distributed computing

Features

Distributed computing



Comparison

Challengers

`OrthoMCL`

- Similarity based Markov Clustering Algorithm

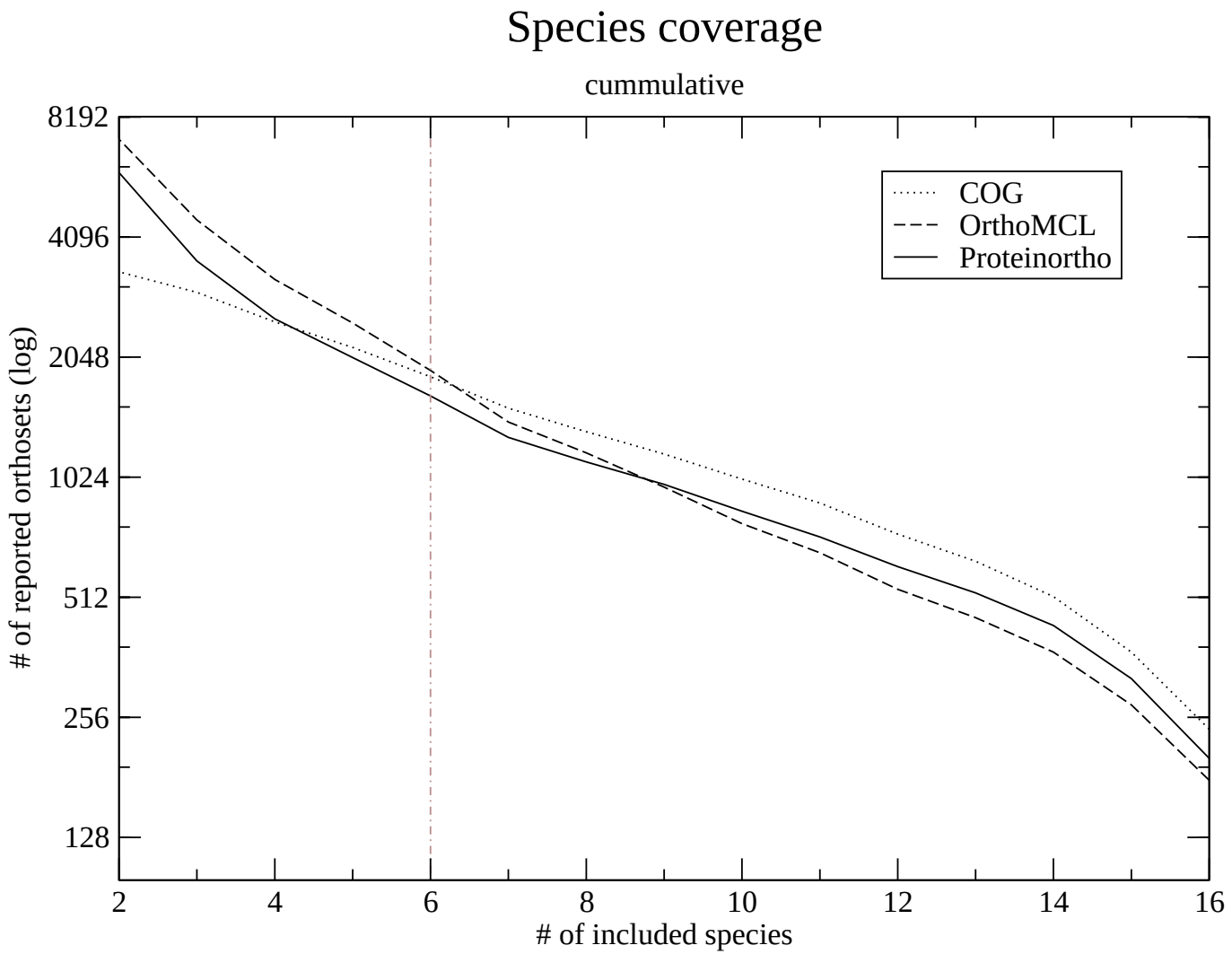
`COG - Clusters of Orthologous Groups`

- Manually curated database

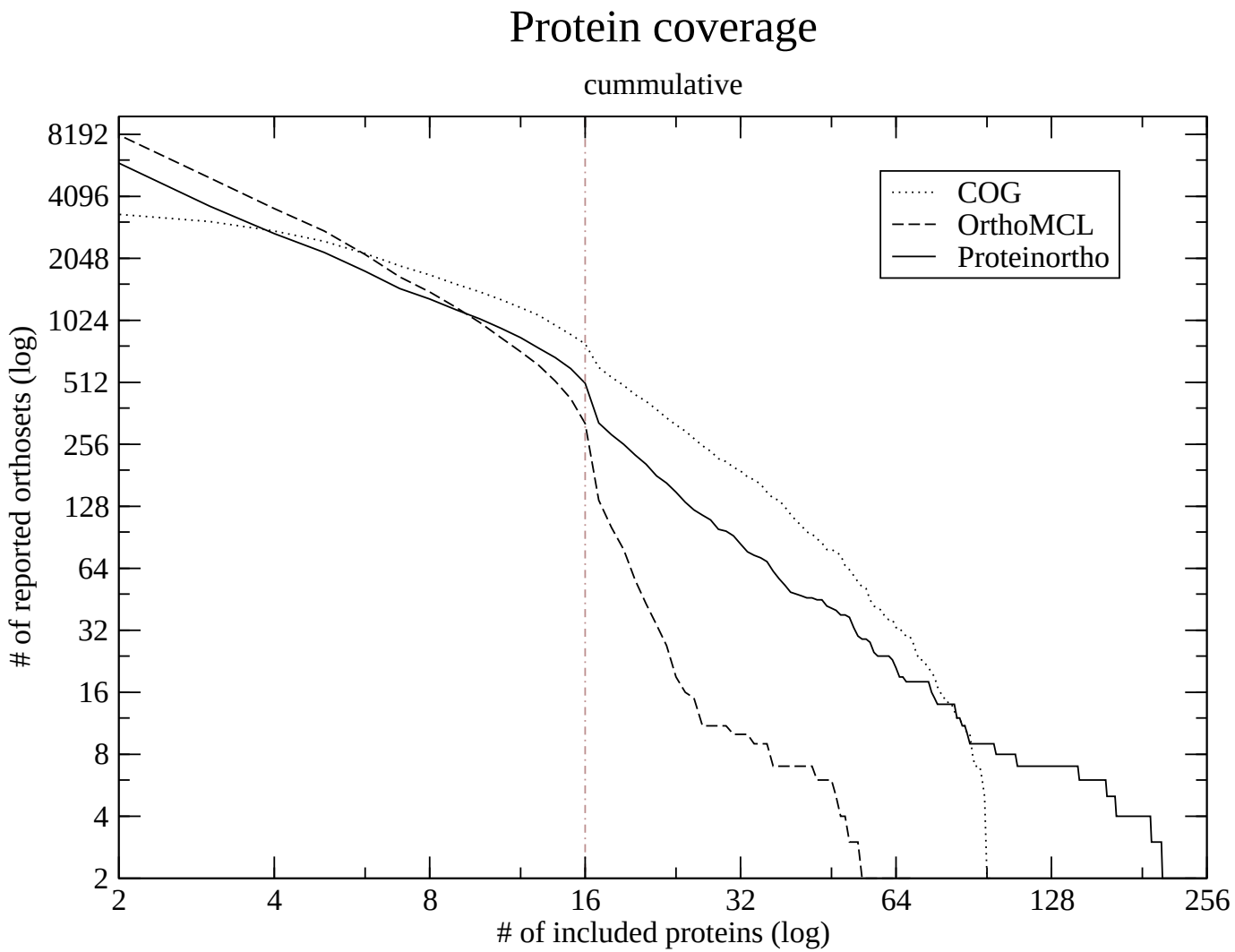
Species used for an example run

Species	Class
Bacillus halodurans	Bacilli (Gram-positive)
Bacillus subtilis	Bacilli (Gram-positive)
Lactococcus lactis	Bacilli (Gram-positive)
Listeria innocua	Bacilli (Gram-positive)
Streptococcus pneumoniae TIGR4	Bacilli (Gram-positive)
Streptococcus pyogenes M1 GAS	Bacilli (Gram-positive)
Buchnera sp. APS	Gamma proteobacteria
Escherichia coli K12	Gamma proteobacteria
Pasteurella multocida	Gamma proteobacteria
Salmonella typhimurium LT2	Gamma proteobacteria
Vibrio cholerae	Gamma proteobacteria
Yersinia pestis	Gamma proteobacteria
Brucella melitensis	Alpha proteobacteria
Caulobacter vibrioides	Alpha proteobacteria
Mesorhizobium loti	Alpha proteobacteria
Rickettsia prowazekii	Alpha proteobacteria

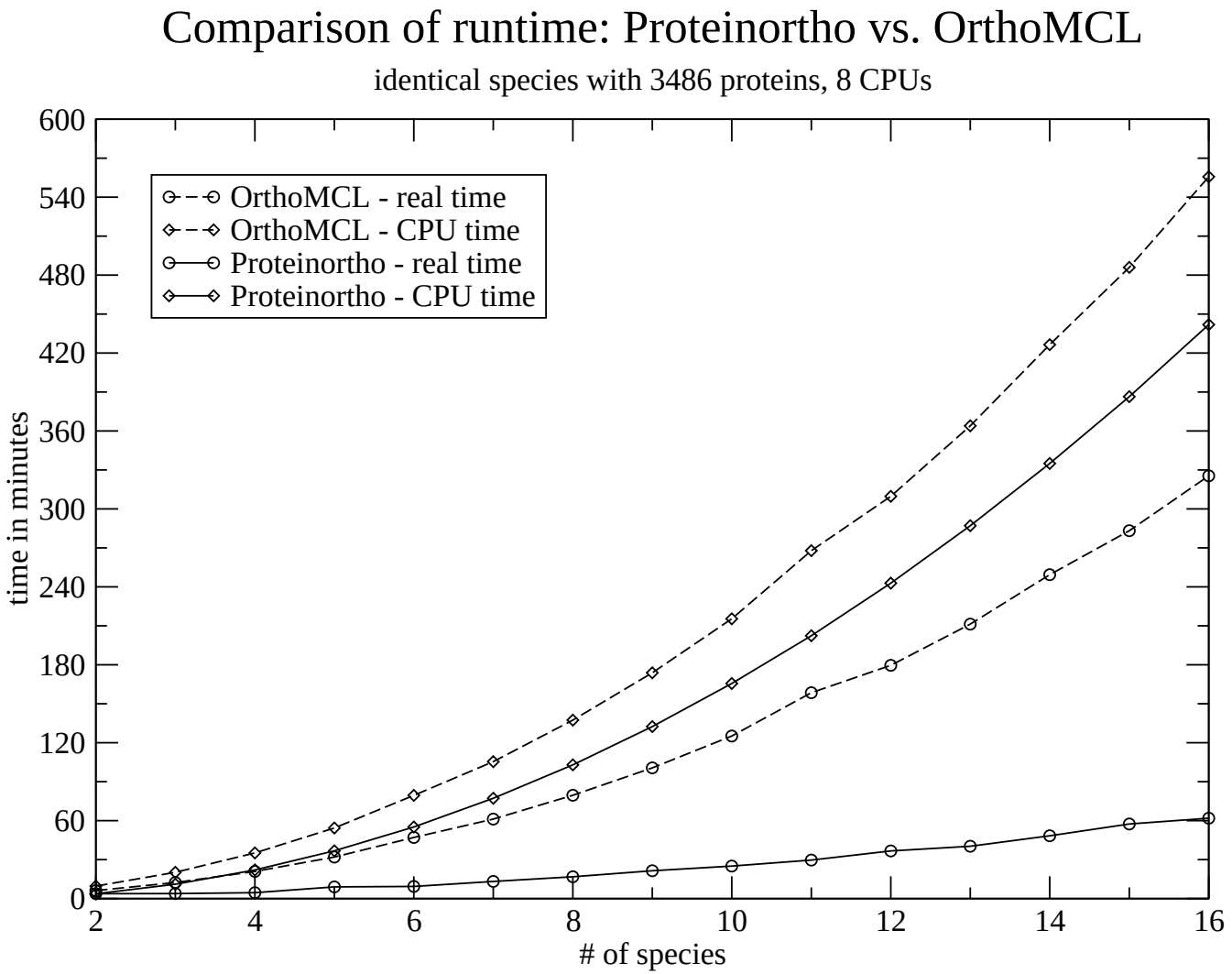
Species coverage



Protein coverage



Speed


Comparison of runtime: Proteinortho vs. OrthoMCL
identical species with 3486 proteins, 8 CPUs
OrthoMCL - real time
OrthoMCL - CPU time
Proteinortho - real time
Proteinortho - CPU time
time in minutes
# of species
0
60
120
180
240
300
360
420
480
540
600
2
4
6
8
10
12
14
16

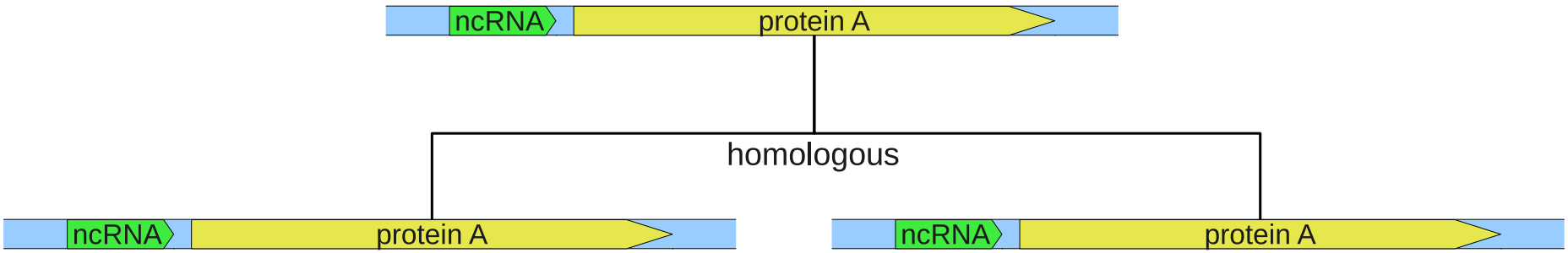
RNA application?

Problem

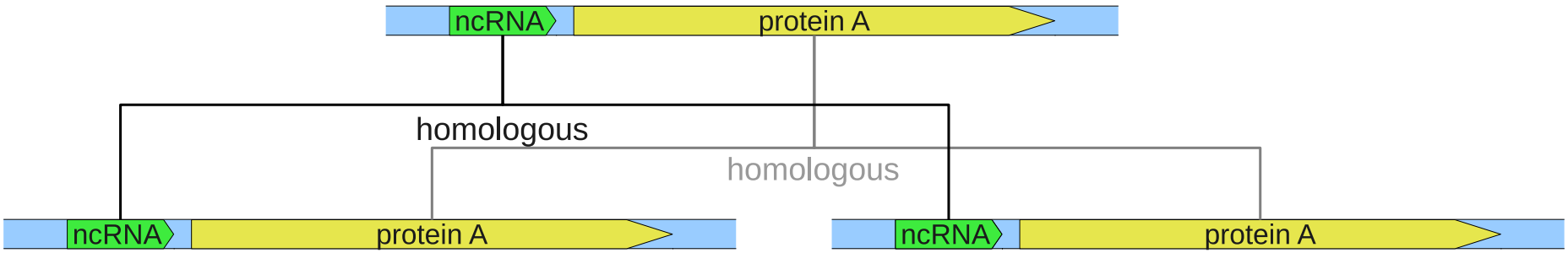
- very small sequences, compared to protein coding regions
- sequence AND structure necessary
- multiple RNA classes
- differing importance of both features

blast alone is no option

RNA application - From protein homology



RNA application - To ncRNA homology



The end

Thank you for listening!

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman.

Basic local alignment search tool.

J Mol Biol, 215(3):403–10, Oct 1990.

W M Fitch.

Homology a personal view on some of the problems.

Trends Genet, 16(5):227–31, May 2000.

E V Koonin.

Orthologs, paralogs, and evolutionary genomics.

Annu Rev Genet, 39:309–38, 2005.

L Li, C J Stoeckert, Jr, and D S Roos.

Orthomcl: identification of ortholog groups for eukaryotic genomes.

Genome Res, 13(9):2178–89, Sep 2003.

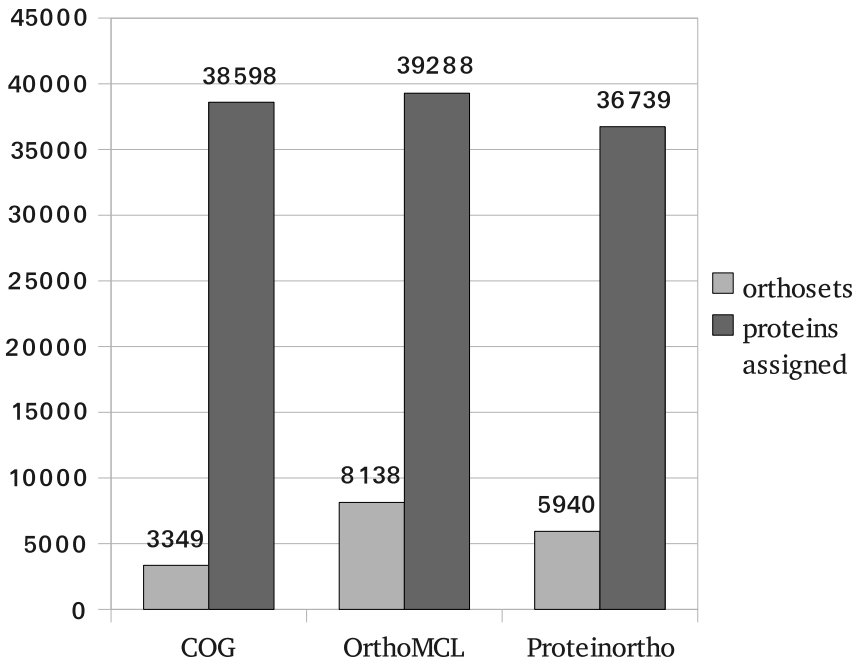
R L Tatusov, N D Fedorova, J D Jackson, A R Jacobs, B Kiryutin, E V Koonin, D M Krylov, R Mazumder, S L Mekhedov, A N Nikolskaya, B S Rao, S Smirnov, A V Sverdlov, S Vasudevan, Y I Wolf, J J Yin, and D A Natale.

The cog database: an updated version includes eukaryotes.

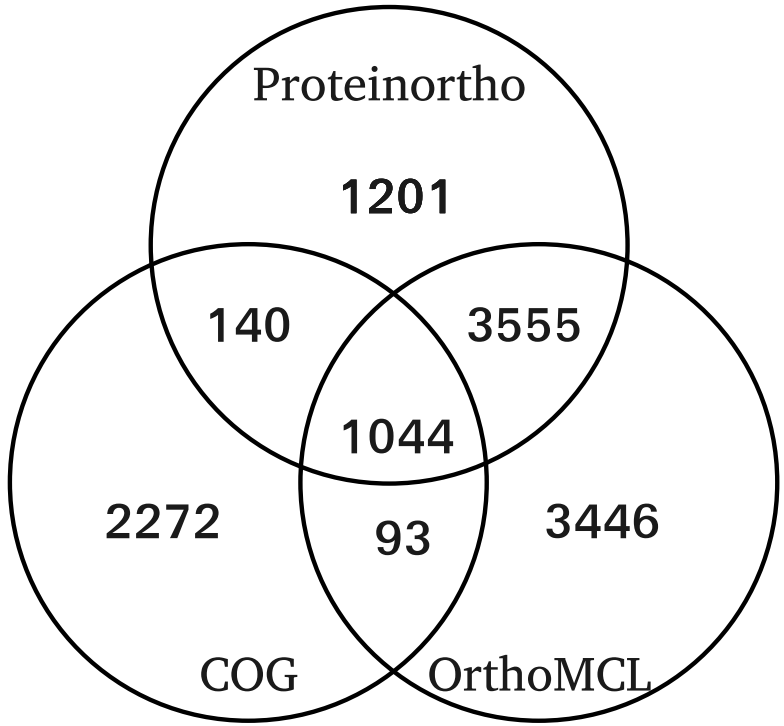
BMC Bioinformatics, 4:41, Sep 2003.

Appendix

a)


Output
53623 proteins
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
38598
39288
36739
3349
8138
5940
COG
OrthoMCL
Proteinortho
orthosets
proteins assigned

b)


Proteinortho
1201
140
3555
1044
2272
93
3446
COG
OrthoMCL