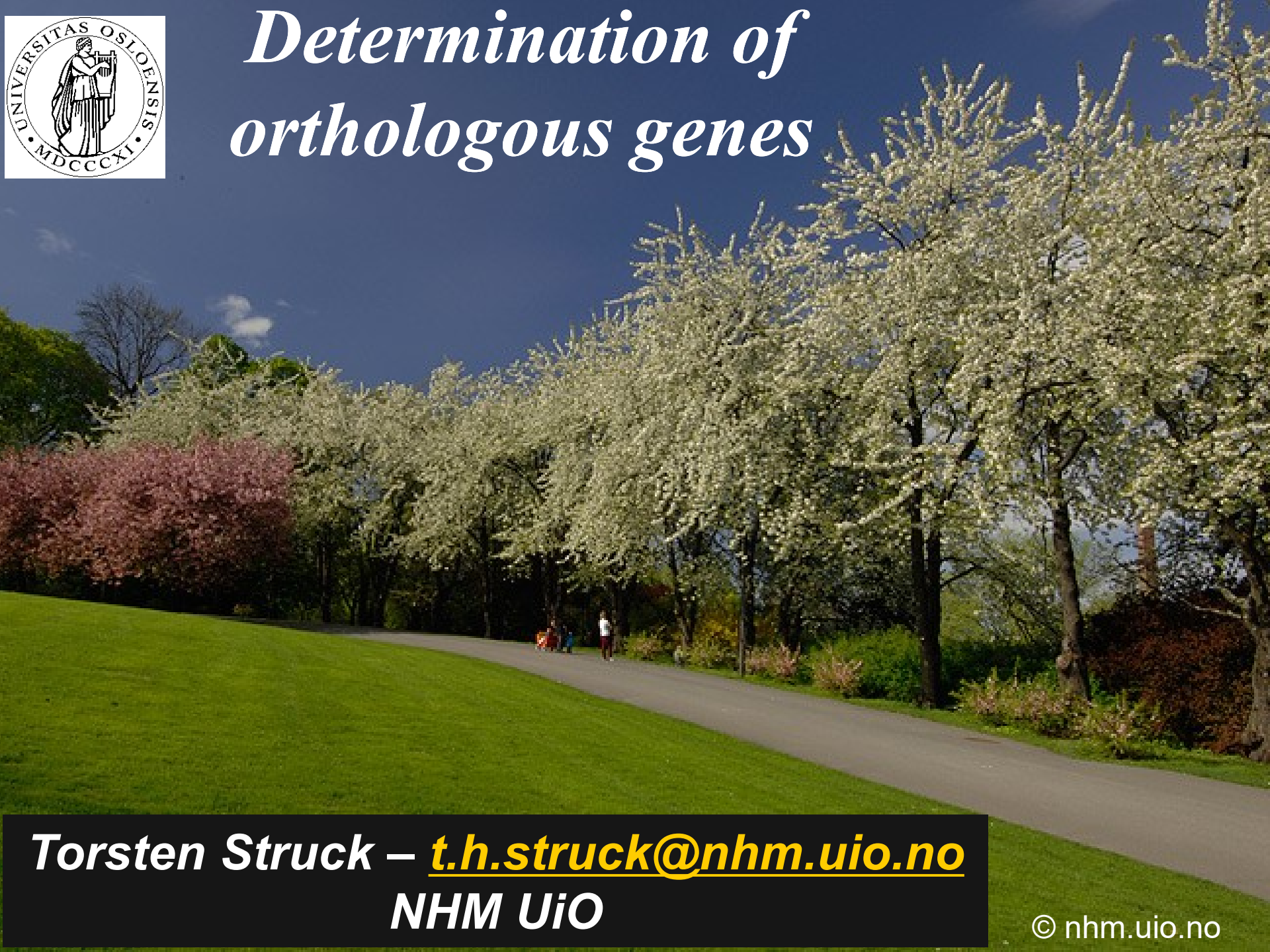




Determination of orthologous genes



Torsten Struck – t.h.struck@nhm.uio.no
NHM UiO

Functional Annotation – kind of

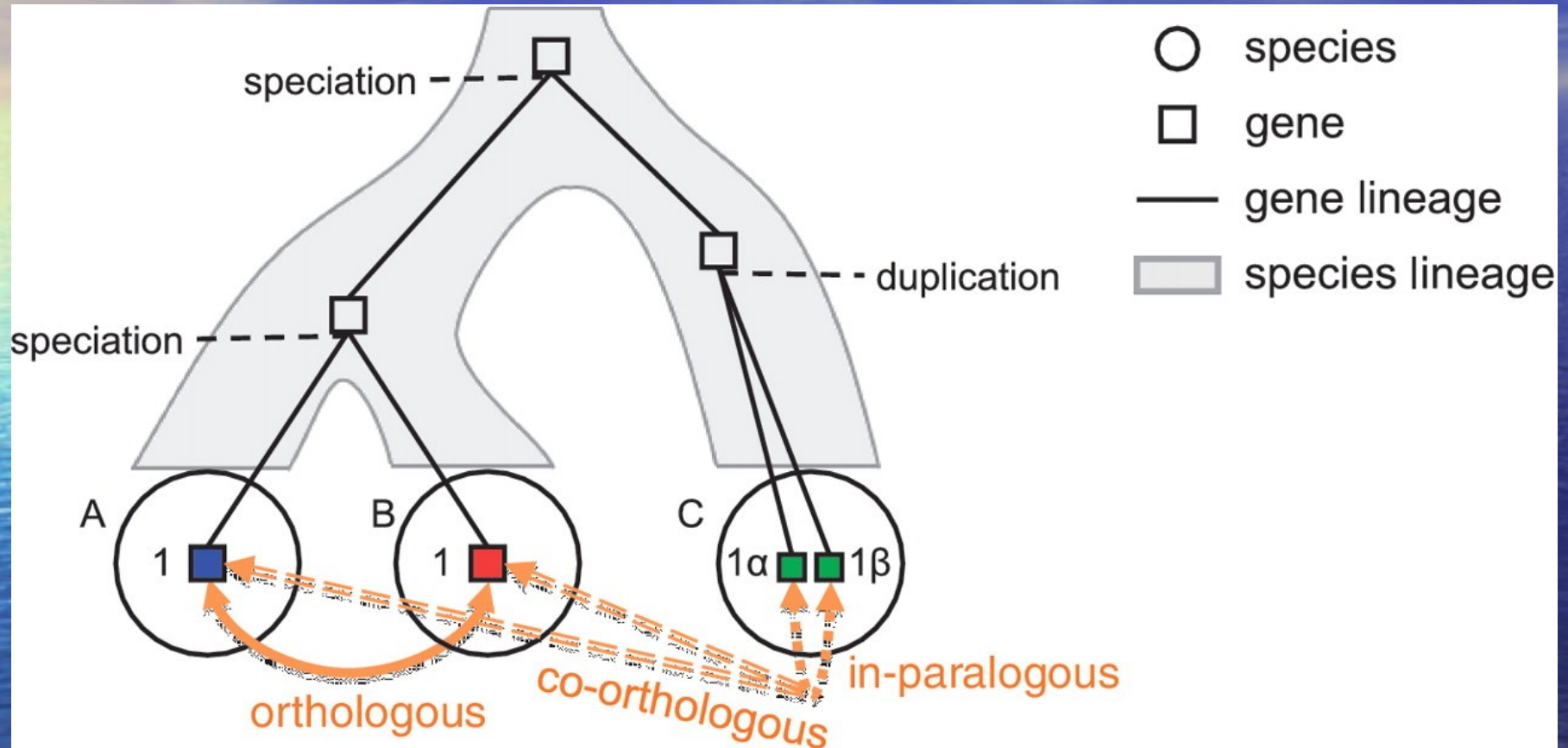
1.) Functional annotation with gene names

e.g. Trinotate → same gene name = orthology

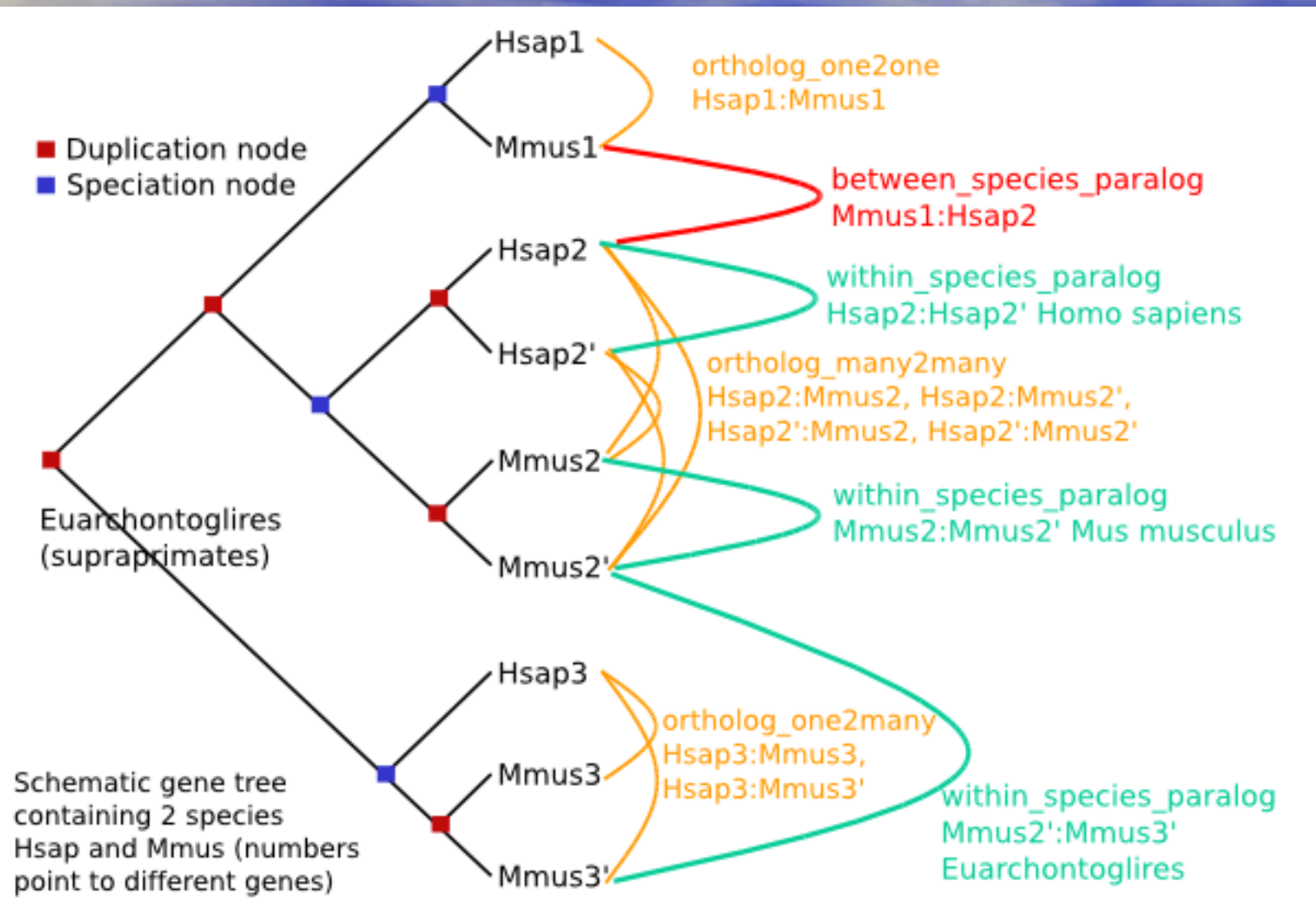
Problem:

gene name does not always imply orthology → paralogs

Orthology vs. Paralogy



Orthology vs. Paralogy



Functional Annotation – kind of

1.) Functional annotation with gene names

e.g. Trinotate → same gene name = orthology

Problem:

gene name does not always imply orthology → paralogs

→ sequence-based approach

Functional Annotation – kind of

1.) Functional annotation with gene names

e.g. Trinotate → same gene name = orthology

Problem:

gene name does not always imply orthology → paralogs

2.) All – vs. – All Blast

e.g. OrthoMCL → no prior knowledge necessary (from scratch)
→ slow & not expandable with new taxa

3.) Reciprocal All – vs. – All Blast using a core ortholog set

e.g. Orthograph → prior knowledge necessary (core set)
→ fast & easily expandable with new taxa

OrthoMCL

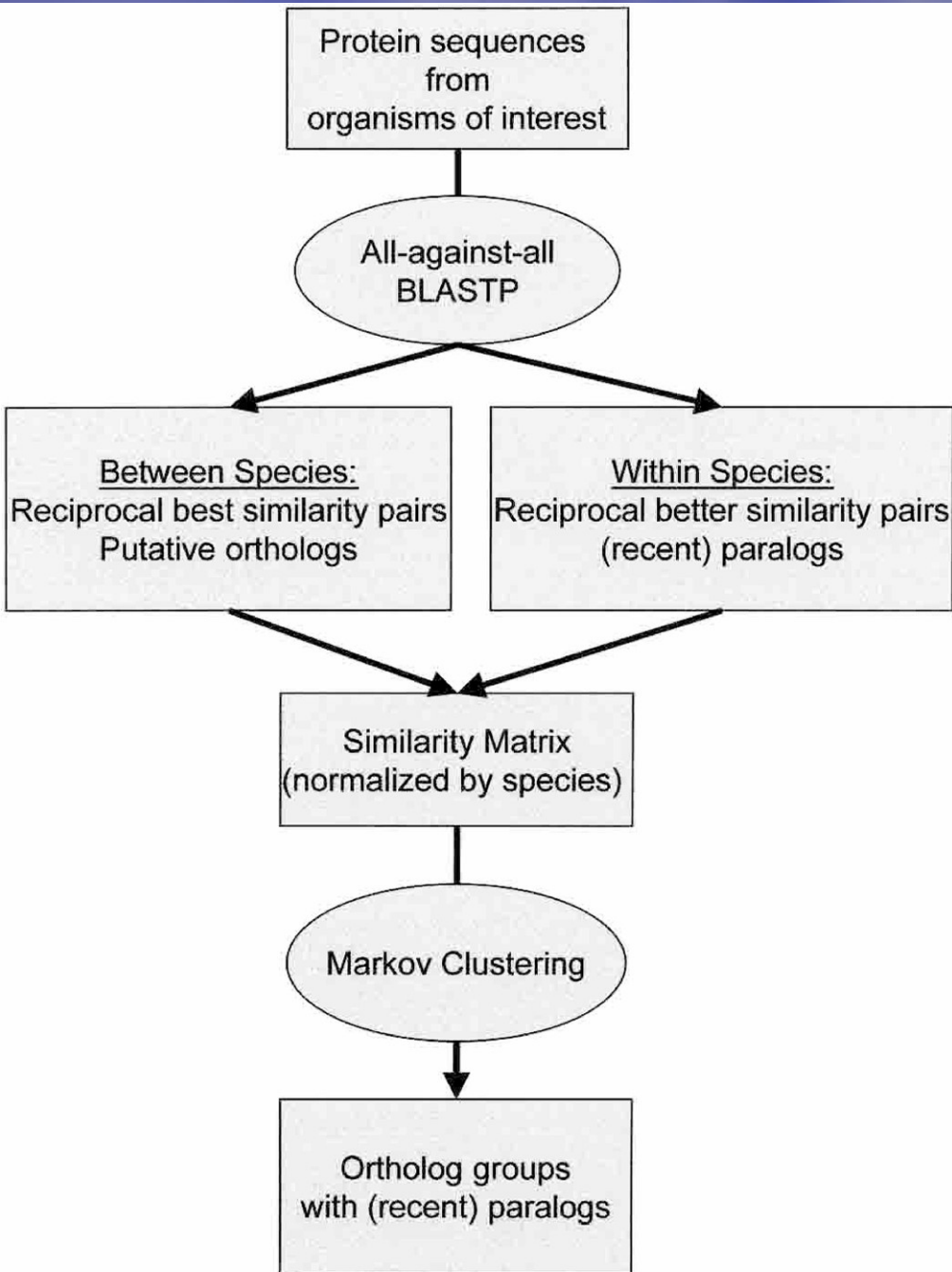


Figure 1 Flow chart of the OrthoMCL algorithm for clustering orthologous proteins.

OrthoMCL – Similarity matrix

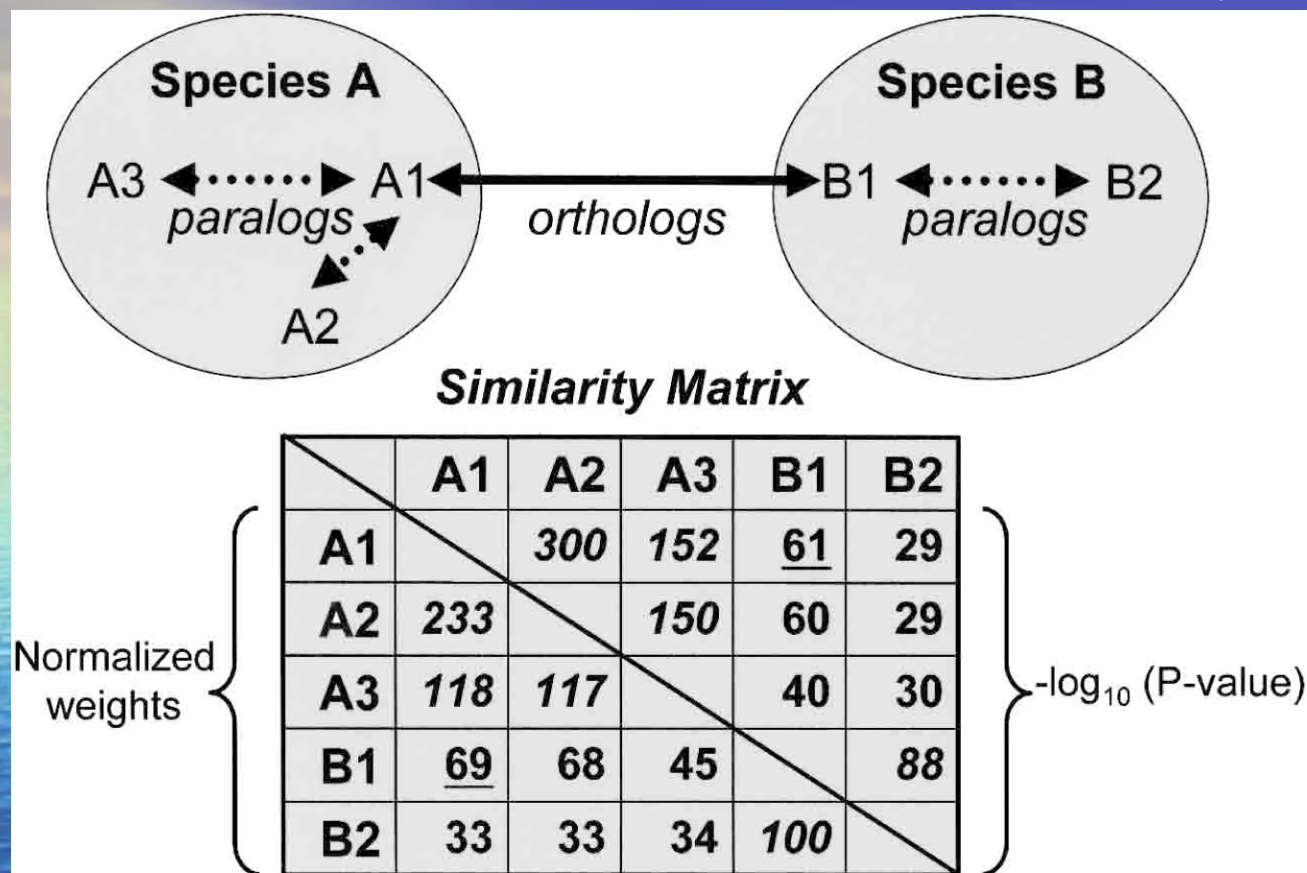
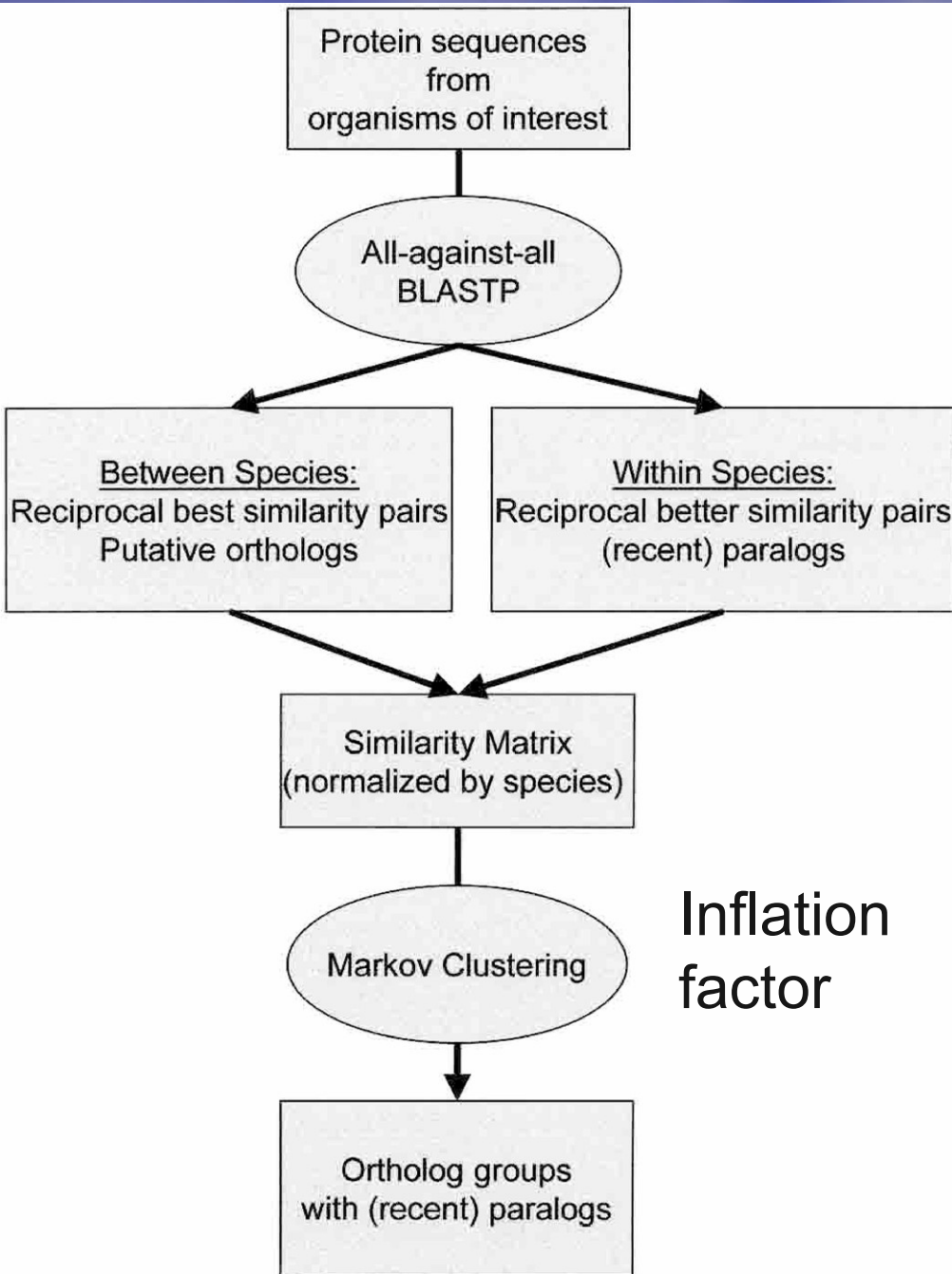


Figure 2 Illustration of sequence relationships and similarity matrix construction. Dotted arrows represent “recent” paralogy (duplication subsequent to speciation); solid arrows represent orthology. The *upper right* half of the matrix contains initial weights calculated as average $-\log_{10}$ (P-value) from pairwise WU-BLASTP similarities. The *lower left* half contains corrected weights supplied to the MCL algorithm; the edge weight connecting each pair of sequences w_{ij} is divided by W_{ij}/W , where W represents the average weight among all ortholog (underlined) and “recent” paralog (italicized) pairs, and W_{ij} represents the average edge weight among all ortholog pairs from species i and j . The net result of this normalization is to correct for systematic differences in comparisons between two species (e.g., differences attributable to nucleotide composition bias), and when $i = j$, to minimize the impact of “recent” paralogs (duplication within a given species) on the clustering of cross-species orthologs.

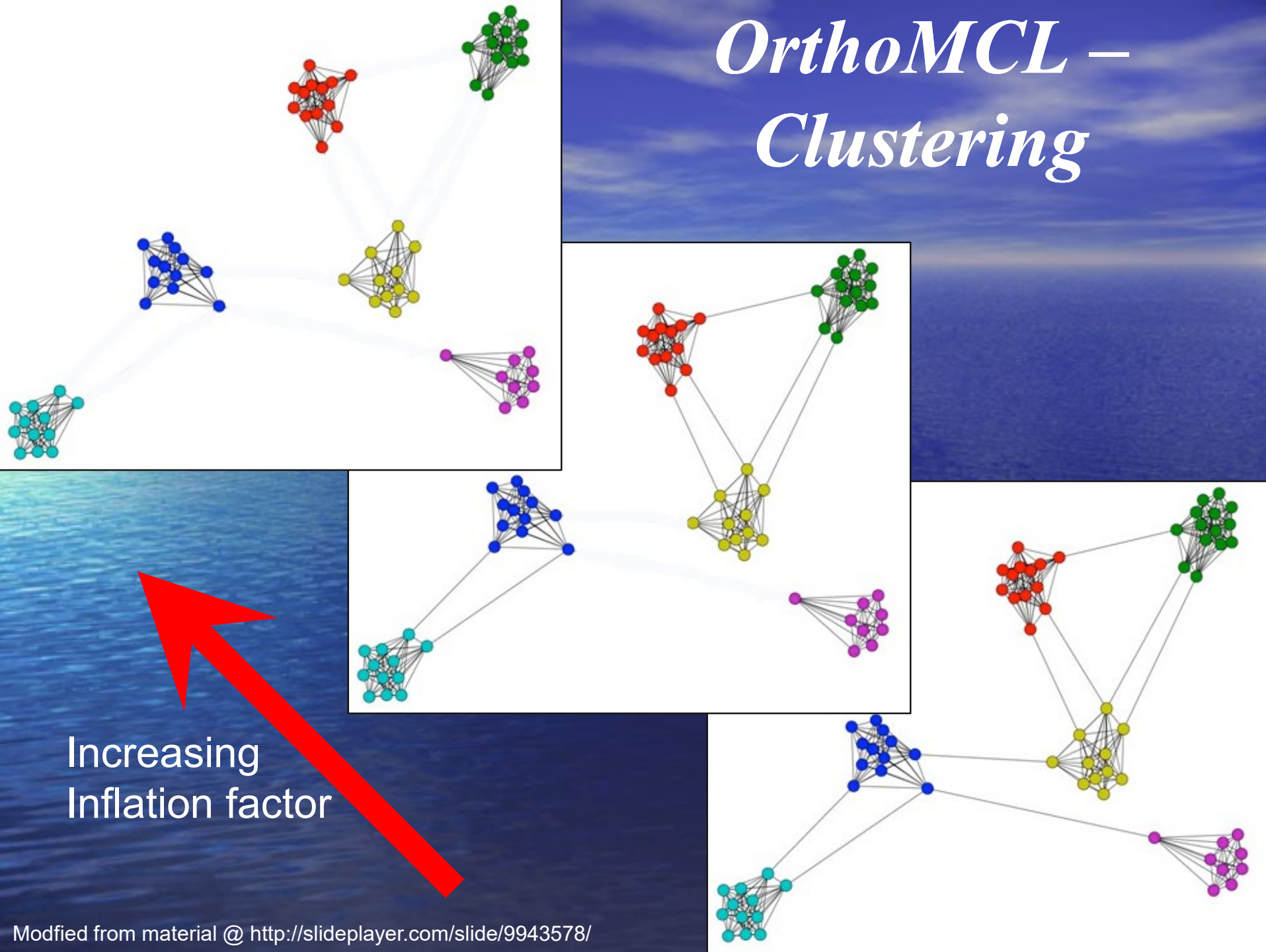
OrthoMCL



An important parameter in the MCL algorithm is the inflation value, regulating the cluster tightness (granularity); increasing the inflation value increases cluster tightness.

Figure 1 Flow chart of the OrthoMCL algorithm for clustering orthologous proteins.

OrthoMCL – Clustering



OrthoMCL – Inflation factor

Table 4. Consistency of OrthoMCL Groups with EC Assignments

Inflation (cluster tightness)	Total data set		Groups with ≥ 1 protein for which complete EC annotation is available			Groups with ≥ 2 proteins for which complete EC annotations are available			Consistent EC assignments ^c		
	Groups	Proteins (% of proteome) ^a	Groups	Proteins	EC-annotated (% of total) ^b	Groups	Proteins	EC-annotated (% of total)	Groups (% poss.)	Proteins	EC-annotated (% possible)
1.1	6,249	50,771 (50)	999	12,032	2,921 (82)	664	9,561	2,586 (73)	528 (80)	5,476	1,958 (76)
1.5	7,265	47,668 (47)	1,117	8,730	2,877 (81)	696	6,318	2,456 (69)	596 (86)	4,768	2,067 (84)
2.0	7,569	46,245 (46)	1,148	8,343	2,849 (80)	701	5,916	2,402 (67)	611 (87)	4,610	2,073 (86)
2.5	7,681	45,473 (45)	1,160	8,171	2,840 (80)	705	5,789	2,385 (67)	617 (88)	4,556	2,062 (86)
3.0	7,786	44,729 (44)	1,172	7,975	2,831 (79)	706	5,553	2,365 (66)	621 (88)	4,450	2,059 (87)
3.5	7,857	44,263 (44)	1,180	7,889	2,821 (79)	707	5,506	2,348 (66)	624 (88)	4,444	2,057 (88)
4.0	7,896	43,900 (43)	1,186	7,784	2,811 (79)	704	5,414	2,329 (65)	623 (88)	4,372	2,048 (88)

^aTotal proteome size = 101,047 (*Arabidopsis thaliana*, 25,009 sequences; *Caenorhabditis elegans*, 19,774; *Drosophila melanogaster*, 13,288; *Homo sapiens*, 27,049; *Plasmodium falciparum*, 5279; *Saccharomyces cerevisiae*, 6358; *Escherichia coli*, 4290).

^bA total of 3562 EC-annotated proteins were obtained from the ENZYME database (*A. thaliana*, 370; *C. elegans*, 269; *D. melanogaster*, 210; *H. sapiens*, 1160; *S. cerevisiae*, 778; *E. coli*, 775).

^cAll EC-annotated sequences in the group were assigned the same EC number. Percentages indicate fraction of ortholog groups containing at least two complete EC assignments (the only data set for which consistency can be assessed), or percentage of EC-annotated sequences properly identified.

HaMStR – ancestor of Orthograph

Build your own custom set of core orthologous genes

Different resources are available, which provide well curated set of orthologous genes.

e.g. EMBL UniProt, InParanoid, OrthoDB9

The screenshot shows the OrthoDB v9 website. At the top, there are logos for the University of Geneva, the Zdobnov's Computational Evolutionary Genomics group, and the SIB (Swiss Institute of Bioinformatics). The main heading is "OrthoDB v9 The Hierarchical Catalog of Orthologs". Below this, a paragraph explains the concept of orthology. To the right, there is a "Build your query" section with various input fields and a "Submit" button. Green arrows and text annotations highlight specific features: (1) "ENTER GENE NAME, ANNOTATION KEYWORD, ETC." points to the "Text search" field; (2) "SUBMIT YOUR QUERY" points to the "Submit" button; "OPTION EVOLUTION FILTER" points to the "Phyloprofile" dropdowns; "SELECT SPECIES and REFINE" points to the "Species to display" and "Select species" fields. The bottom of the page contains links for "Data downloads", "OrthoDB software", "BUSCO", and "OrthoDB-News".

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics group

SIB

OrthoDB v9
The Hierarchical Catalog of Orthologs

Orthology is the cornerstone of comparative genomics and gene function prediction. OrthoDB aims to classify protein-coding genes from the increasing number of available sequenced genomes into groups of orthologs descended from a single gene of the last common ancestor (LCA) of each clade of species. Applying this concept to the major species radiations of the hierarchy of LCAs along the species phylogeny results in multiple levels of orthology: the more closely-related the species, the more finely-resolved the orthologous relations.

Examples of how you can query OrthoDB
[Cytochrome P450](#), [protease](#) | [peptidase](#), [kinase -serine](#), [FBgn0036816](#), [GO:0006950](#), [immune response](#), [stress response](#), [breast cancer](#), [diabetes](#).

Please cite
OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software.
EV Kriventseva, F Tegenfeldt, TJ Petty, RM Waterhouse, FA Simao, IA Pozdnyakov, P Ioannidis, and EM Zdobnov NAR, Jan 2015, [PMID:25428351](#)

Email: [support\[at\]orthodb.org](mailto:support[at]orthodb.org)

Data downloads Protein sequences and orthologous group annotations for major clades.
OrthoDB software Can be used to compute orthologs on custom data.
BUSCO Assessing completeness of genome assembly and annotation with single-copy genes.
OrthoDB-News Join the mailing list to keep abreast of the latest developments.

Build your query Search by sequence

Text search:

Phyloprofile:

Search at:

Species to display: Clear all

Submit

Select species: Search species by name:

(1) ENTER GENE NAME, ANNOTATION KEYWORD, ETC.

(2) SUBMIT YOUR QUERY

OPTION EVOLUTION FILTER

SELECT SPECIES and REFINE

Reciprocal BLAST searches

