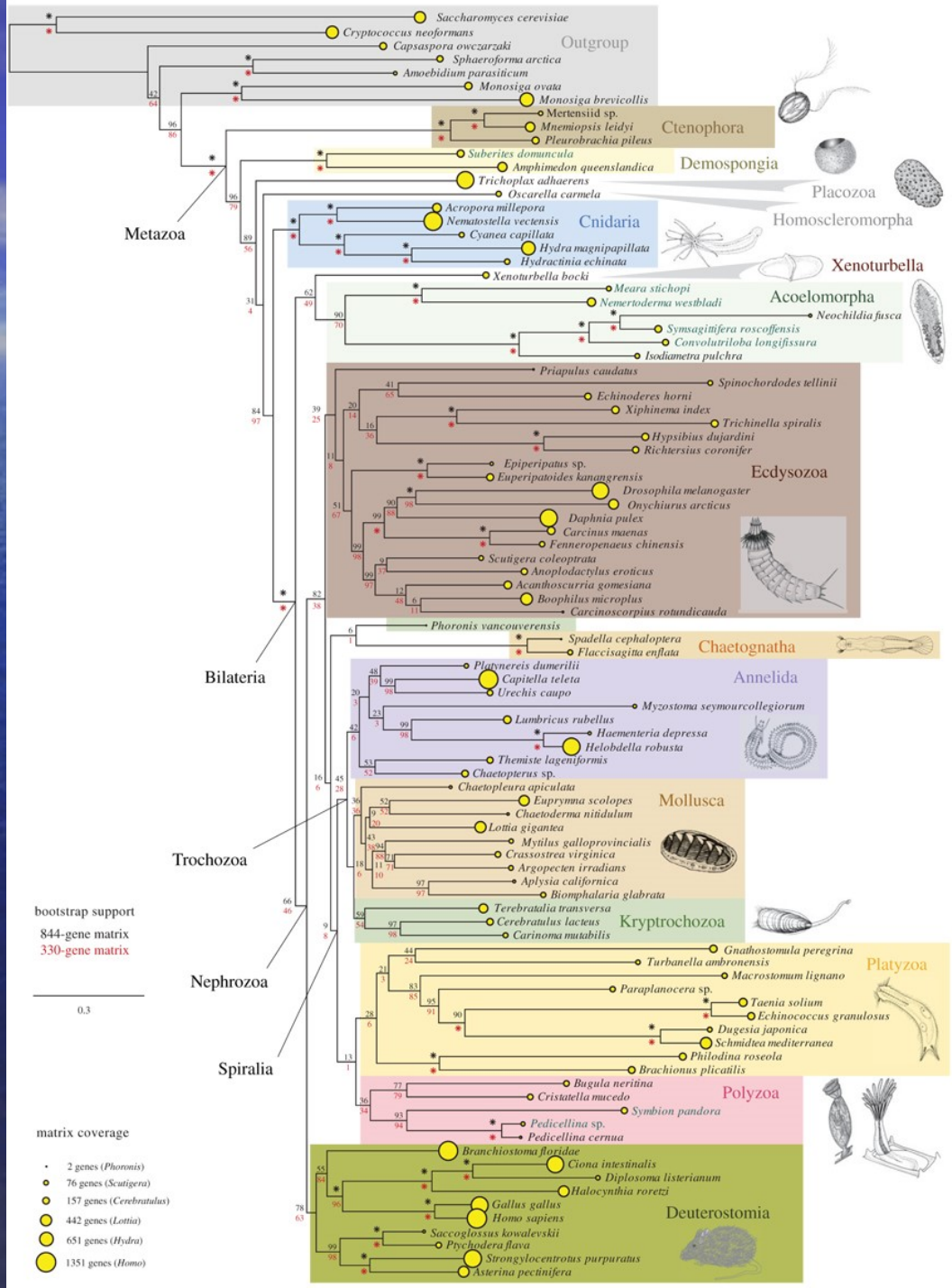
Missingness & Phylogenetic signal

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

Missing data

Coverage of taxa per genes can be different and similarly of genes per taxon.
Introduction of missing data in the tree.
Mislead analyses.

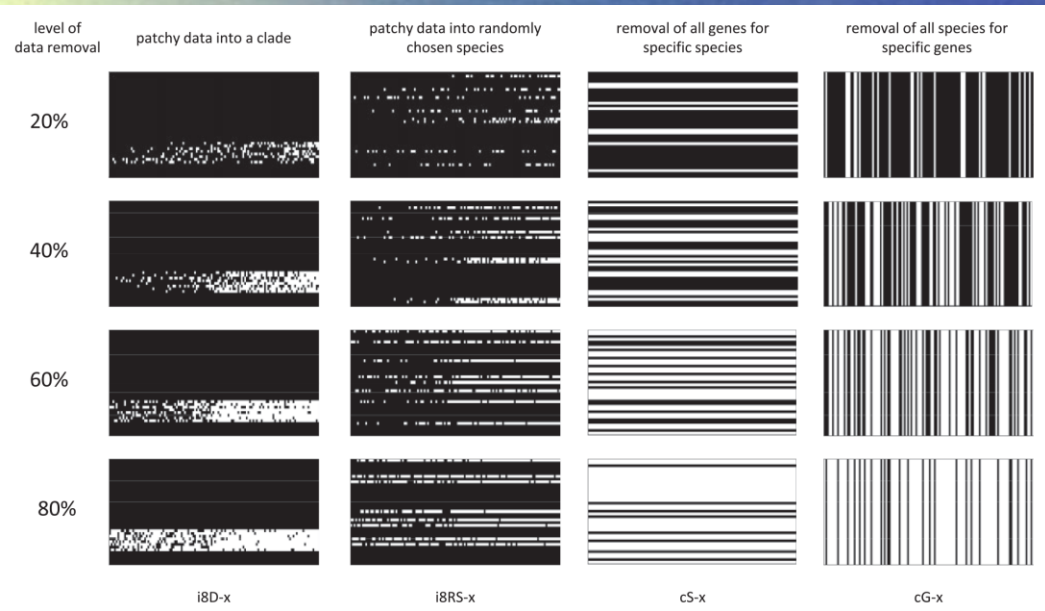


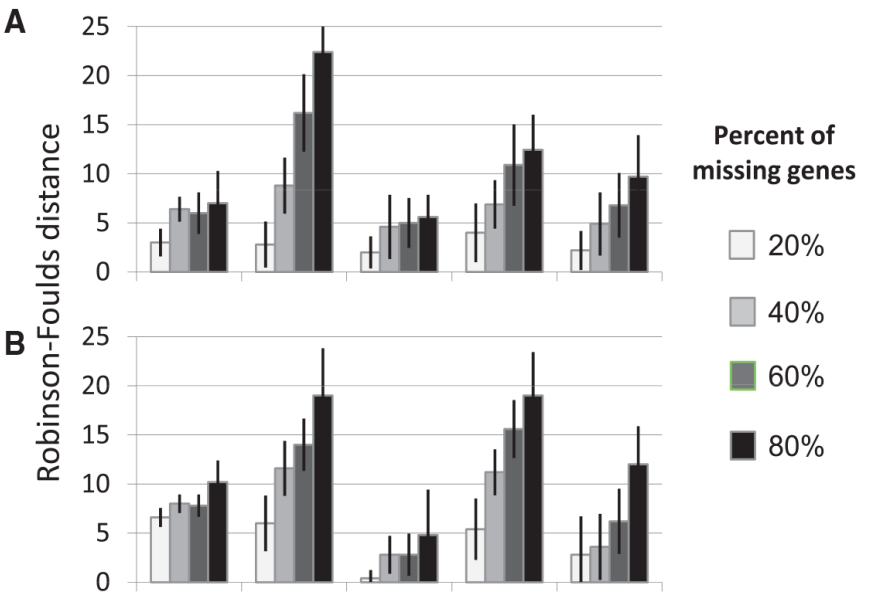
Hejnol et al. (2009). Proc. R. Soc. B: Biol. Sci.

Simulation studies

Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets

Béatrice Roure,[1] Denis Baurain,[‡,2] and Hervé Philippe*,[1]





Internodal support|phylogenetic signal

Inferring ancient divergences requires genes with strong phylogenetic signals

Leonidas Salichos[1] & Antonis Rokas[1]

To tackle incongruence, the topological conflict between different gene trees, phylogenomic studies couple concatenation with practices such as rogue taxon removal or the use of slowly evolving genes. Phylogenomic analysis of 1,070 orthologues from 23 yeast genomes identified 1,070 distinct gene trees, which were all incongruent with the phylogeny inferred from concatenation. Incongruence severity increased for shorter internodes located deeper in the

results in analyses of vertebrate and metazoan phylogenomic data sets. These results question the exclusive reliance on concatenation and associated practices, and argue that selecting genes with strong phylogenetic signals and demonstrating the absence of significant incongruence are essential for accurately reconstructing ancient divergences.

Internode support could be the single gene trees or the bootstrap support values.

Hughes et al. (2018)

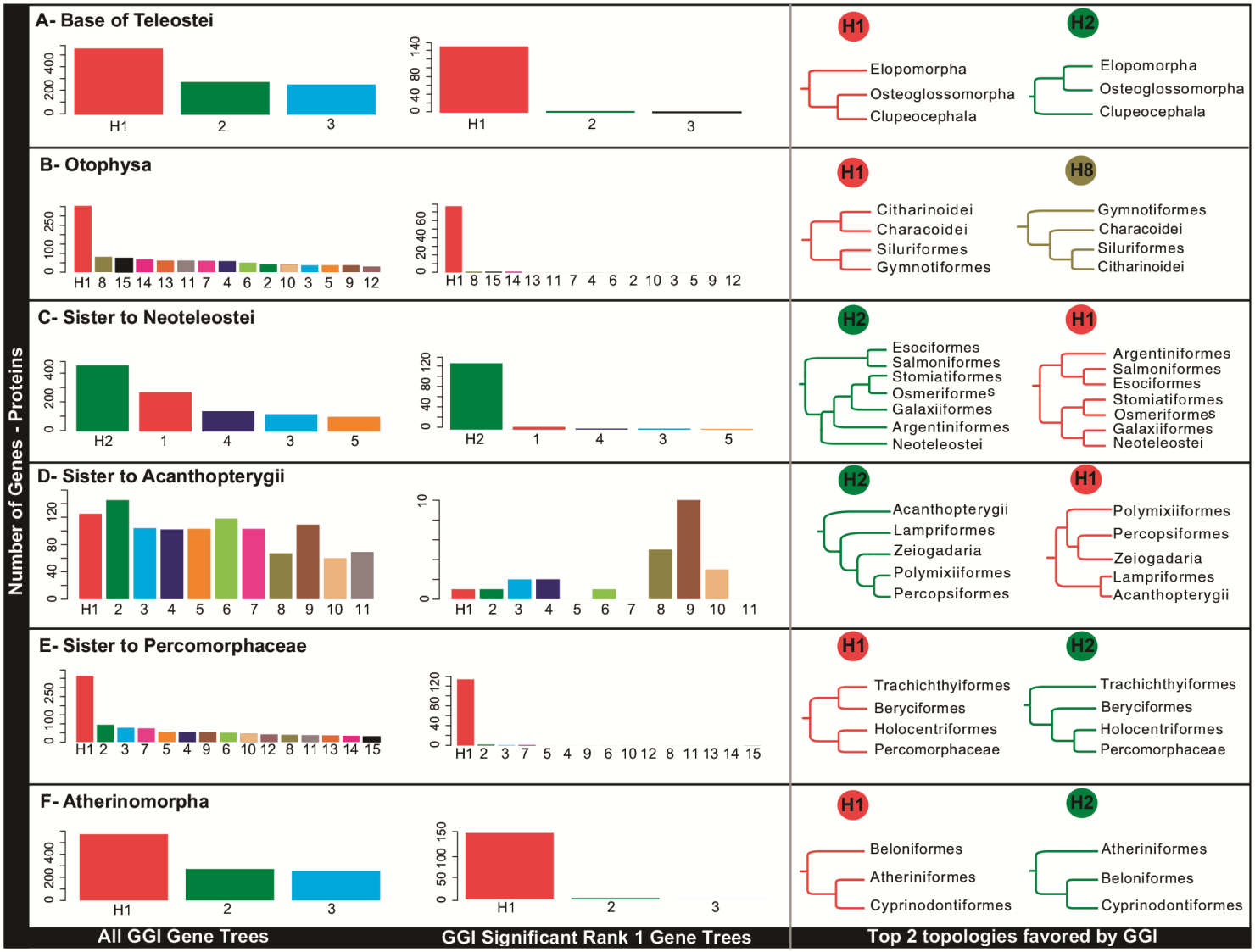
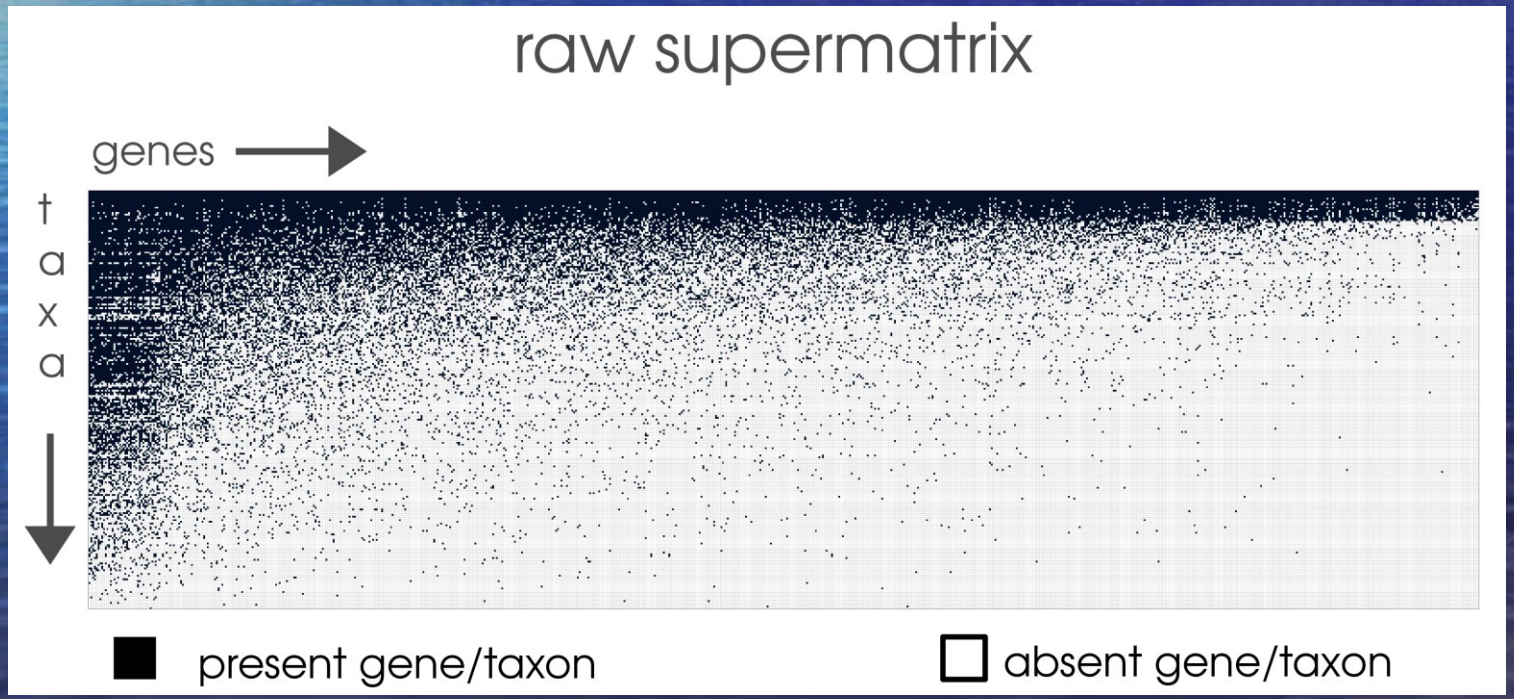


Fig. 3. GGI results based on protein alignments. For each specific hypothesis tested (*A*–*F*), the distribution of all gene trees supporting each alternative hypothesis (*Left*) or only the significantly supported hypotheses (*Middle*) are shown. The top two topologies favored by these tests are shown on the *Right* and in Fig. 2. GGI results based on nucleotide alignments are shown in *SI Appendix*, Fig. S7.

Matrix reduction

Matrix reduction based on an optimality function taksing into account signal (treelikeness) within genes/partitions, distribution of present data and signal heterogeneity and overall information content (IC) of the supermatrix.



Misof et al. (2013). BMC Bioinf.

New reduction heuristic

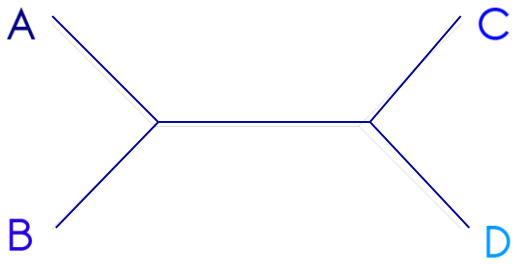
Selection of an optimal data subset (SOS):

- Information content of genes (extended Geometry Quartet Mapping eGM)
- overlap between genes and taxa (quasibiclique like)
- optimisation towards a high information content of a (sub)matrix with as many taxa as possible

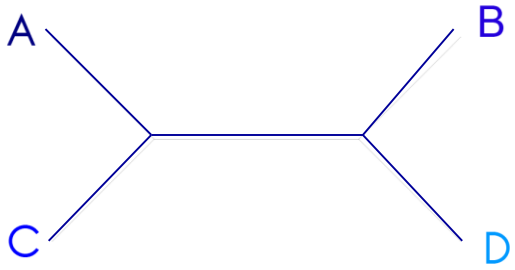
Misof et al. (2013). BMC Bioinf.

extended Geometry Quartet Mapping (eGM)

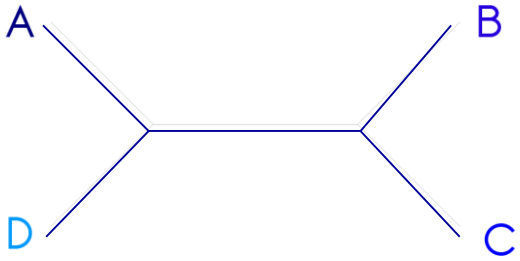
For each quartet, 3 topologies are possible:



$\delta_1 = \sum_N (AB \mid CD)$



$\delta_2 = \sum_N (AC \mid BD)$



$\delta_3 = \sum_N (AD \mid BC)$

δ: support for each topology = ∑ of all branch lengths of a quartet, calculated from distances

Nieselt Struwe & von Haeseler (2001) Mol. Biol. Evol.

extended Geometry Quartet Mapping (eGM)

Support values δ_i for the topology of each quartet are transformed in relative support values s_i:

$s_1 = \delta_1 / (\delta_1 + \delta_2 + \delta_3)$ e.g. $s_1 = 0.5$

$s_2 = \delta_2 / (\delta_1 + \delta_2 + \delta_3)$ e.g. $s_2 = 0.5$

$s_3 = \delta_3 / (\delta_1 + \delta_2 + \delta_3)$ e.g. $s_1 = 0.0$

$$\sum s_i = 1$$

Draw relative support values into a 2-D simplex graph S as baricentric coordinates (vectors s_1,s_2,s_3)

Nieselt Struwe & von Haeseler (2001) Mol. Biol. Evol.

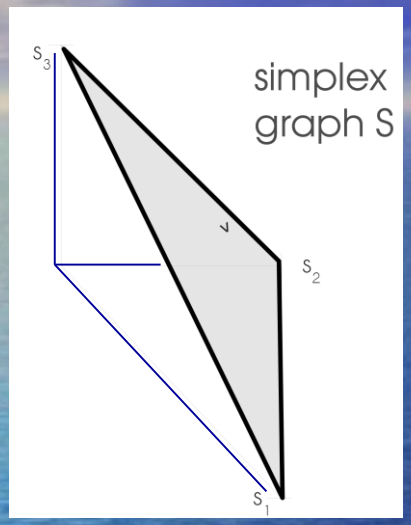
extended Geometry Quartet Mapping (eGM)

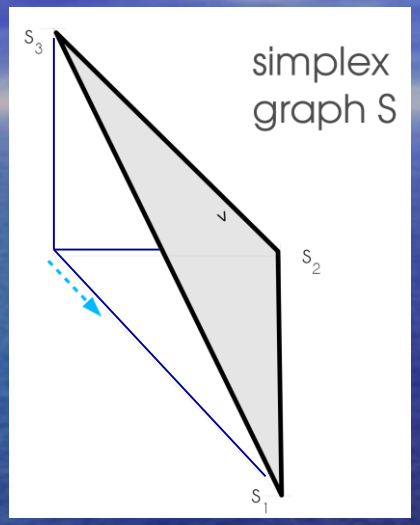
- starlike topology (unresolved):
  $s_1 = 0.333,\ s_2 = 0.333\ ,\ s_3 = 0.333$
- only topology 1 supported:
  $s_1 = 1.0,\ s_2 = 0.0\ ,\ s_3 = 0.0$
- support for topology 1 and 2 similar; topology 3 not supported:
  $s_1 = 0.5,\ s_2 = 0.5\ ,\ s_3 = 0.0$

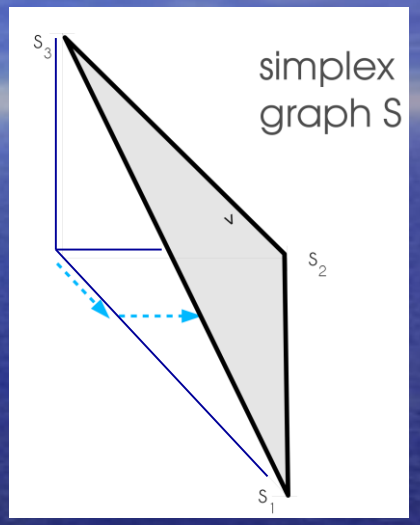
Nieselt Struwe & von Haeseler (2001) Mol. Biol. Evol.

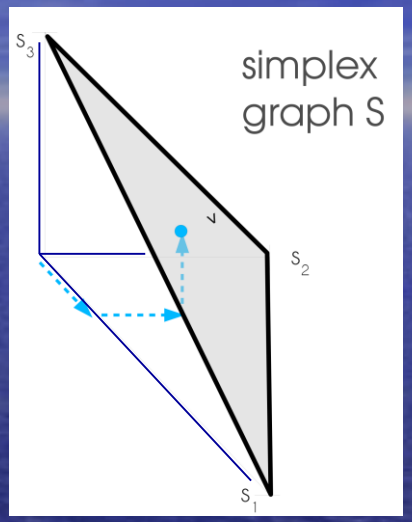
extended Geometry Quartet Mapping (eGM)

quartet 1
$s_1 = 0.333$
$s_2 = 0.333$
$s_3 = 0.333$

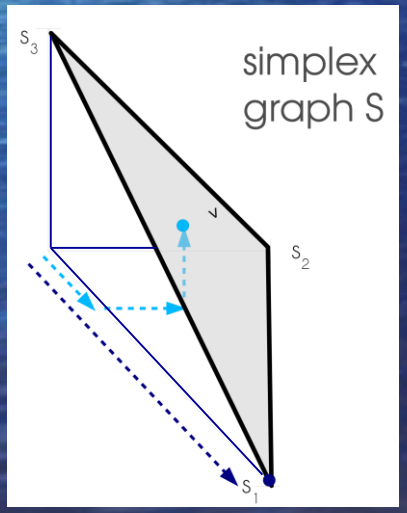




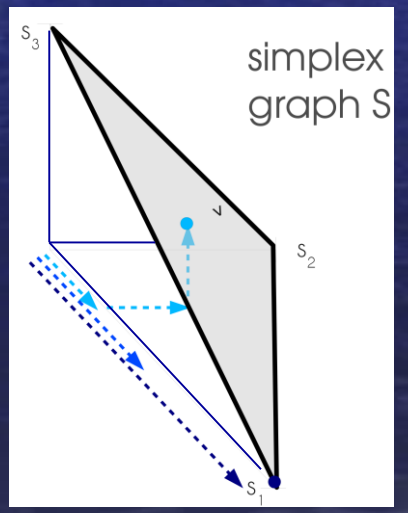


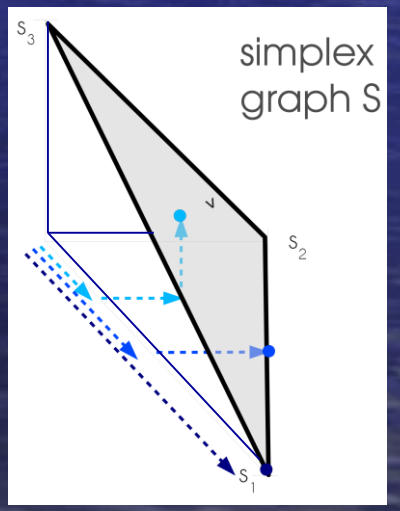


quartet 2
$s_1 = 1.0$
$s_2 = 0.0$
$s_3 = 0.0$



quartet 3
$s_1 = 0.5$
$s_2 = 0.5$
$s_3 = 0.0$

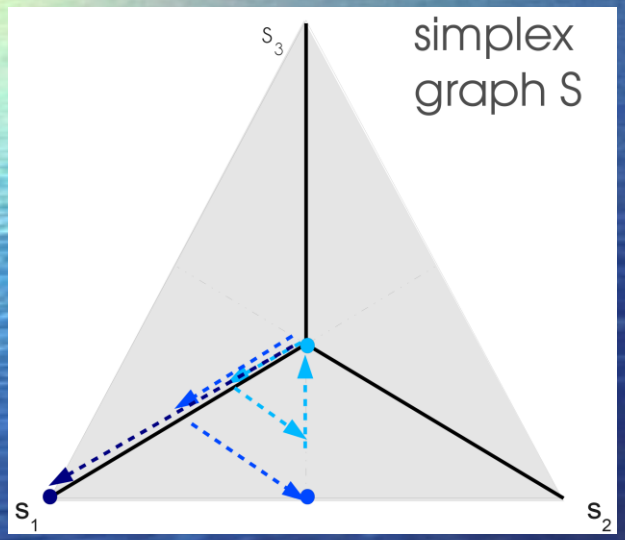


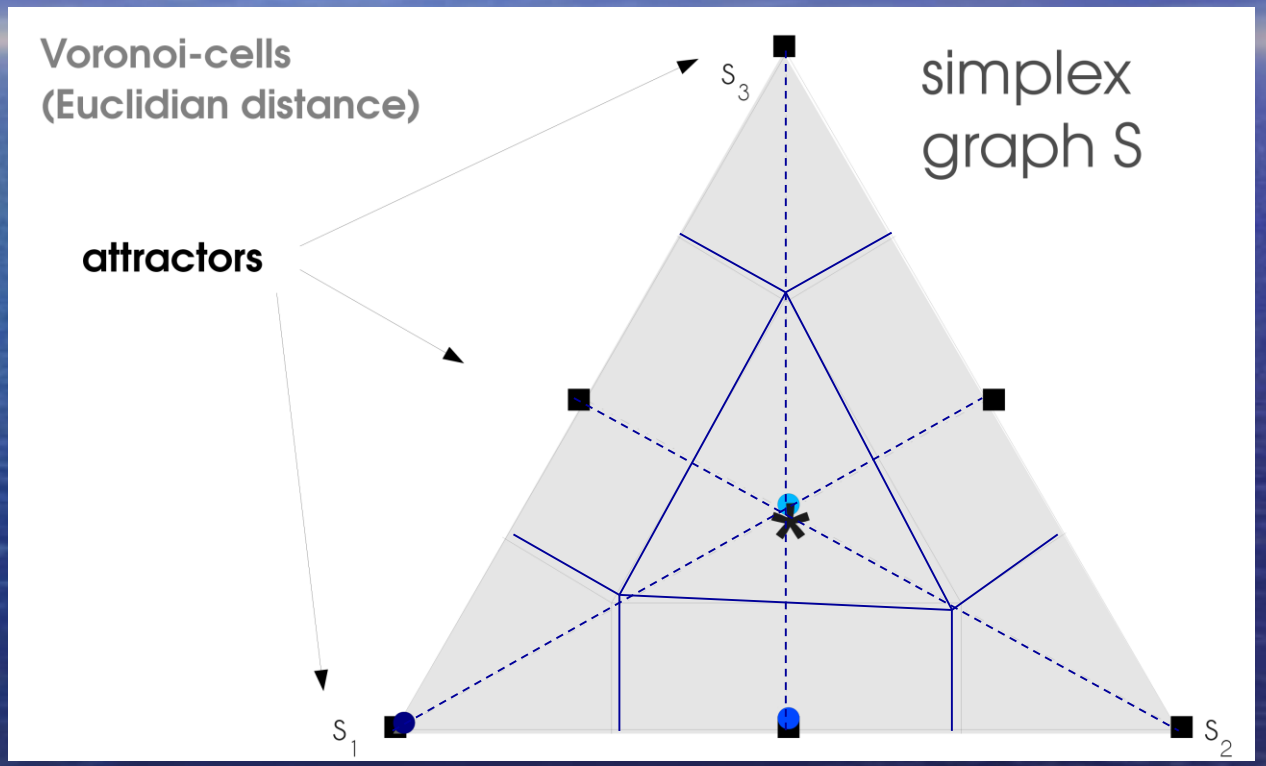


Nieselt Struwe & von Haeseler (2001) Mol. Biol. Evol.

extended Geometry Quartet Mapping (eGM)

quartet 1	quartet 2	quartet 3
$s_1 = 0.333$	$s_1 = 1.0$	$s_1 = 0.5$
$s_2 = 0.333$	$s_2 = 0.0$	$s_2 = 0.5$
$s_3 = 0.333$	$s_3 = 0.0$	$s_3 = 0.0$





Nieselt Struwe & von Haeseler (2001) Mol. Biol. Evol.

Information content p of gene and taxon

outer region Φ_r

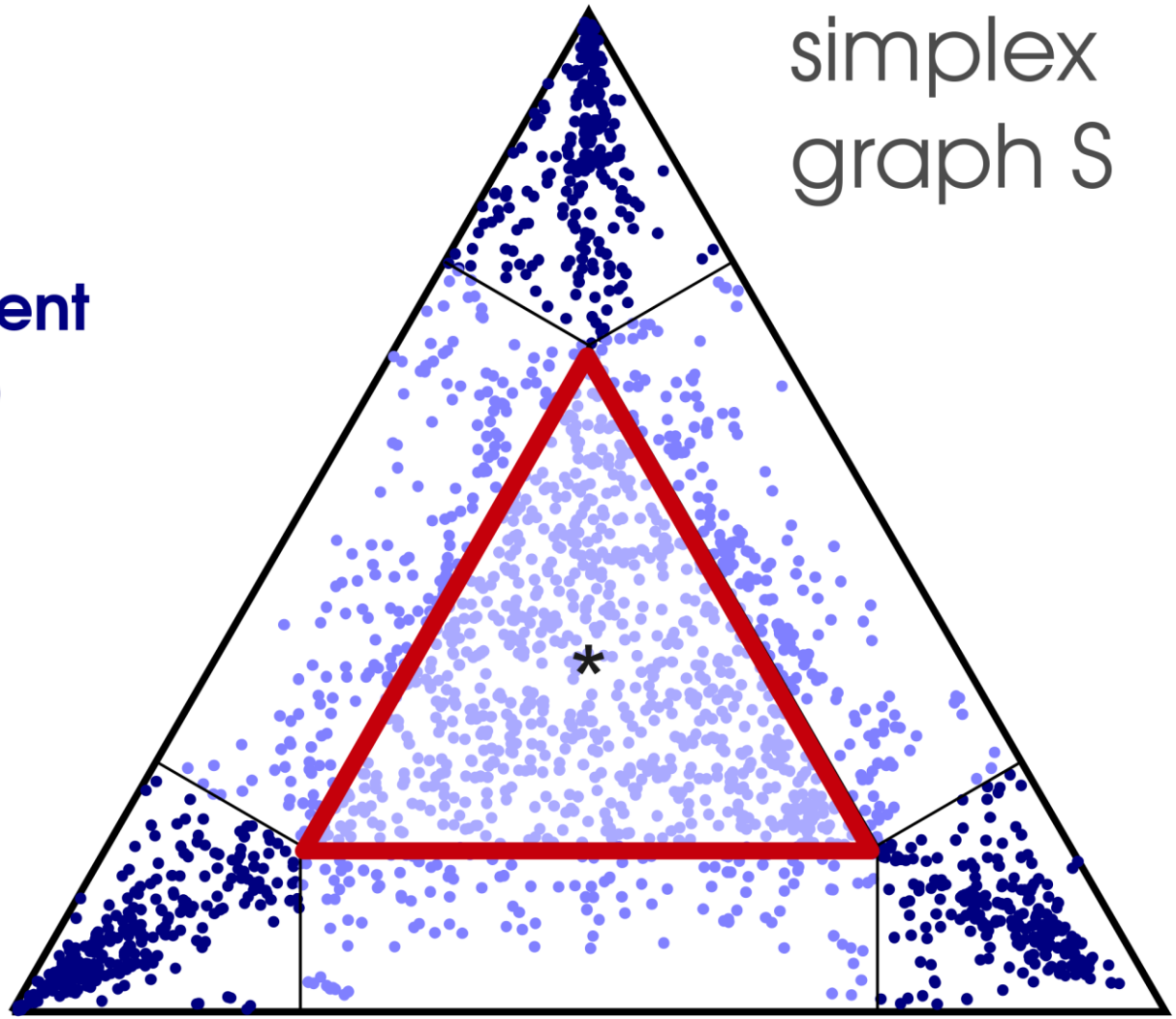
inner region Φ_*

Information content

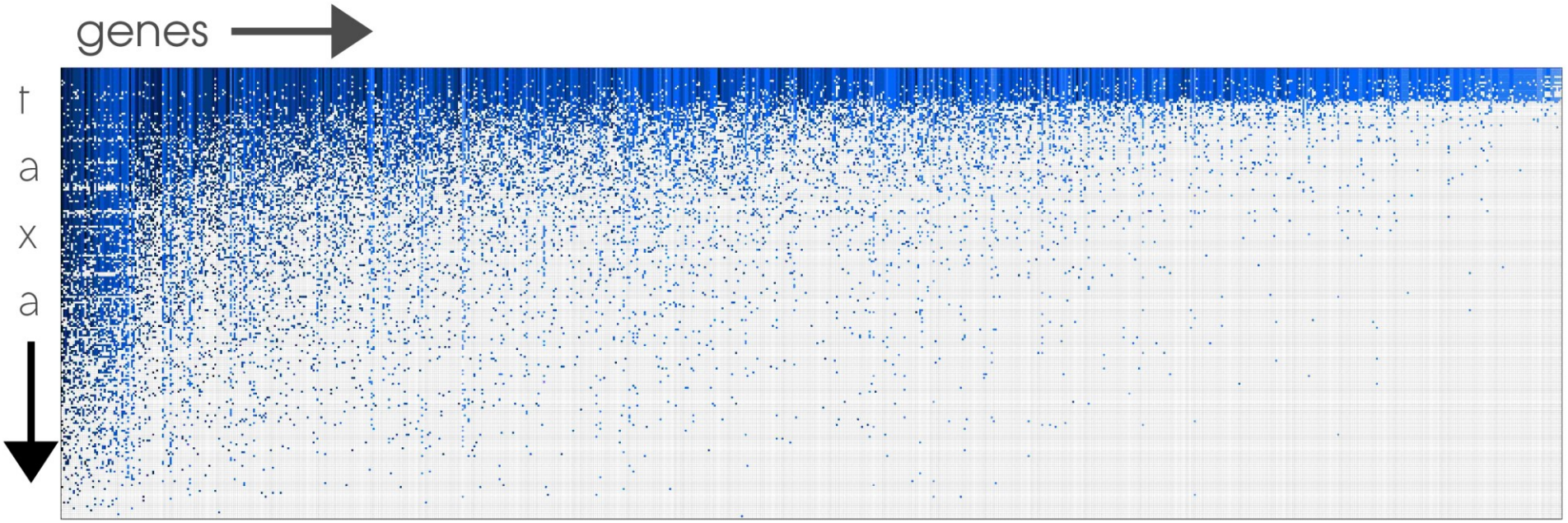
$p = \Phi_r / (\Phi_r + \Phi_*)$

if $0 \leq p \leq 1$

(0.61)



Information content & saturation



original matrix: Meusemann *et al.* 2010, MBE

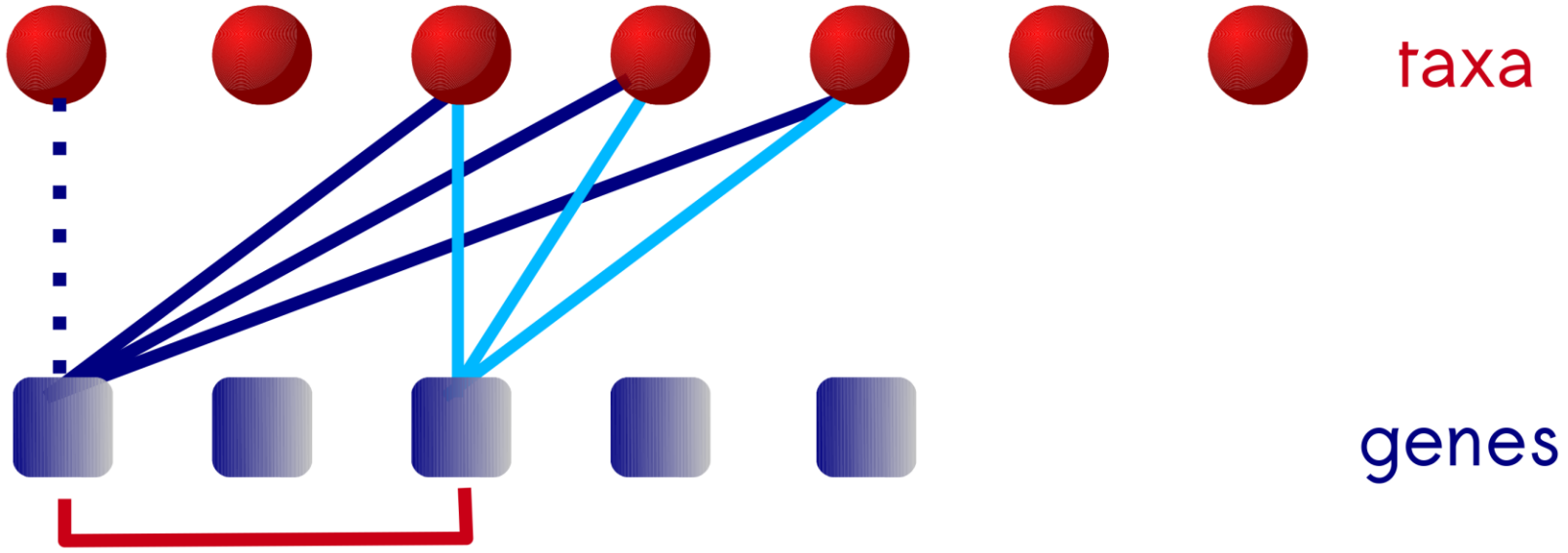
information content P of matrix B: 0.106
sum of tree-likeness values / sum of all entries

matrix saturation: 19.6%
sum of all present genes (entries = 1) / sum of all entries

Selection of optimal submatrix SOS

Gene and taxa overlap is monitored during reduction.

overlap: 2 genes/partitions must be connected by at least 3 taxa



Selection of optimal submatrix SOS

Stepwise reduction (gene or taxon with lowest IC is discarded; in case of ties, genes are excluded).

stepwise reduction

	G 1	G 2	G 3	G 4		∅
Tax 1	0.8	0.6	0.3	0.1		0.36
Tax 2	0.8	0.6	0.3	0.1		0.36
Tax 3	0.8	0.6	0.3	0.1		0.36
Tax 4	0.8	0.6	0	0.1		0.3
Tax 5	0.8	0	0.3	0		0.22
∅	0.8	0.48	0.24	0.08		

Selection of optimal submatrix SOS

With each reduction, a new matrix is generated;
IC P‘ of matrix B' and saturation are recalculated;
The matrix is resorted.

with every reduction step

→ **the size ratio λ decreases:**

$$\lambda = \frac{\#\text{ taxa x }\#\text{ partitions (matrix reduced)}}{\#\text{ taxa x }\#\text{ partitions (matrix original)}}$$

→ the IC P' of matrix B' increases

Hence, a penalty for the size is needed to avoid extreme reductions to only the very best IC p values.

Optimality criterion

optimality criterion

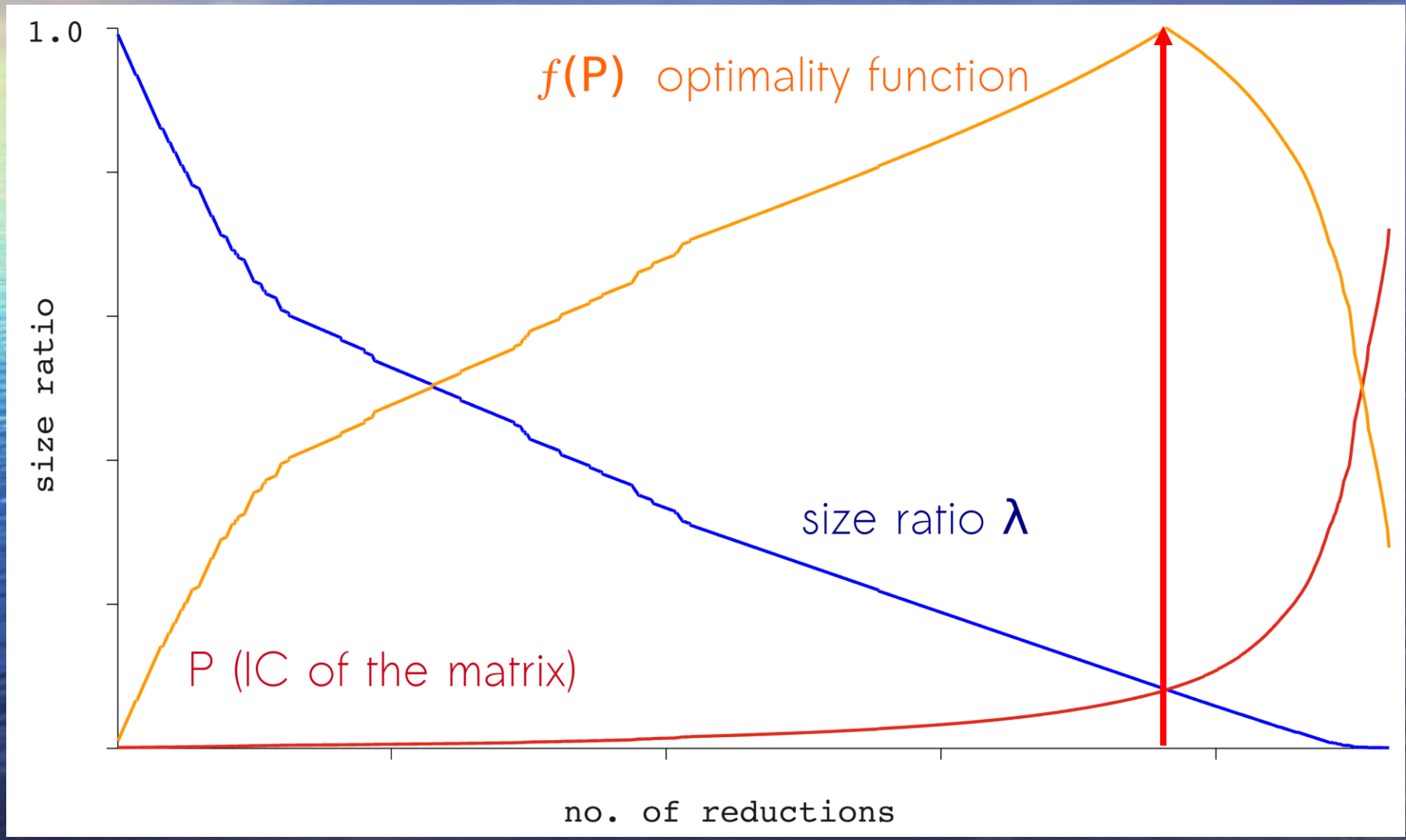
$$f(P) = 1 - |\lambda - P^{\alpha(1-P)}|$$

λ: size ratio

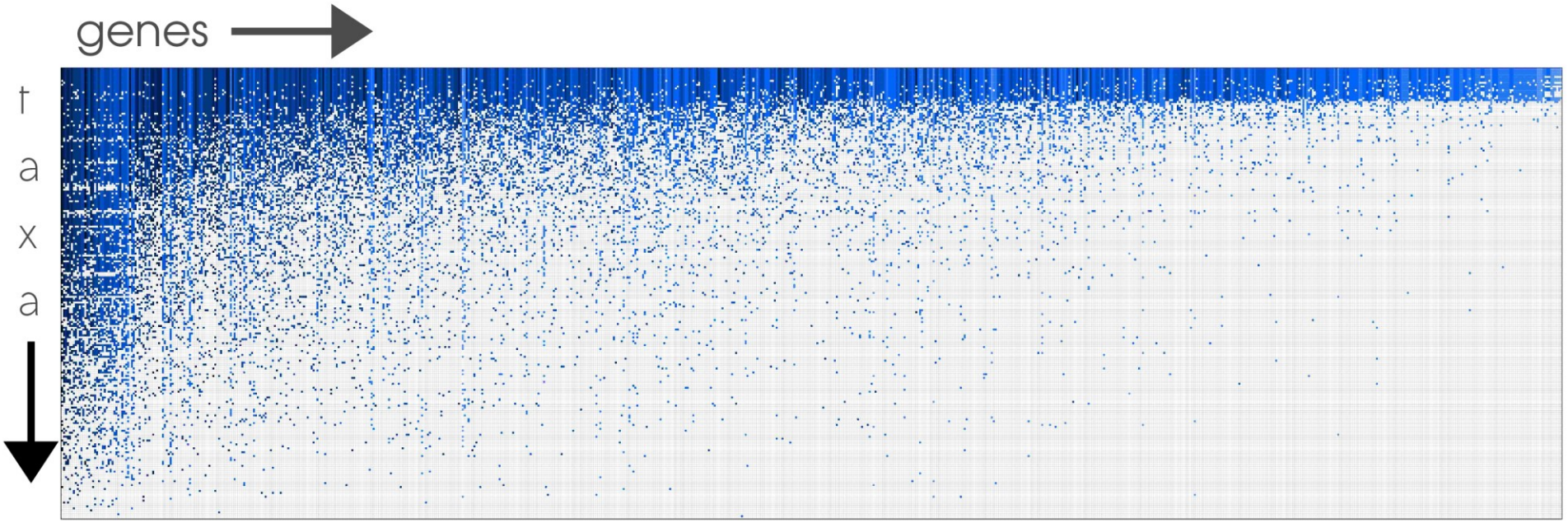
P: information content of the matrix

→ $f(P)$ is optimized

if $f(P)$ = 1, reduction stops



Original matrix



original matrix: Meusemann *et al.* 2010, MBE

information content P of matrix B: 0.106
sum of tree-likeness values / sum of all entries

matrix saturation: 19.6%
sum of all present genes (entries = 1) / sum of all entries

Optimized matrix

SOS Arthropoda

Taxa: 117
Genes: 129
Saturation: 62.3%
IC: ~ 0.43

Meusemann *et al.* 2010, MBE