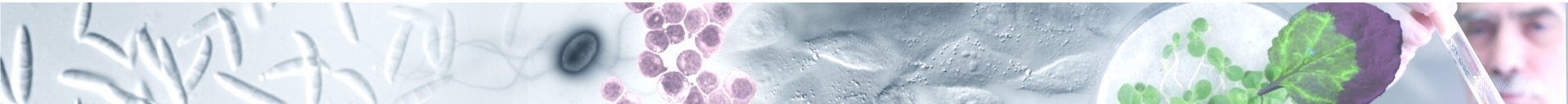
M. Göker

Clustering Optimization For Molecular Taxonomy



Molecular taxonomy

Definition:
Establishing a (informal or even formal) taxonomy of organisms based only on molecular sequences

Uses:

- Revision of taxon boundaries
- Detection of cryptic and pseudocryptic species
- Detection of misidentifications and mislabelled sequences in public databases
- Identification of juvenile specimens
- Analysis of environmental samples (e.g. metagenomics)

Threshold-based clustering

- Calculate distance *d(i,j)* between each pair of sequences *i* and *j*
- Define a threshold *T*
- Principle: if *d(i,j)* <= *T*, assign *i* and *j* to the same molecular operational taxonomic unit (MOTU)

From Meier et al. (2006)

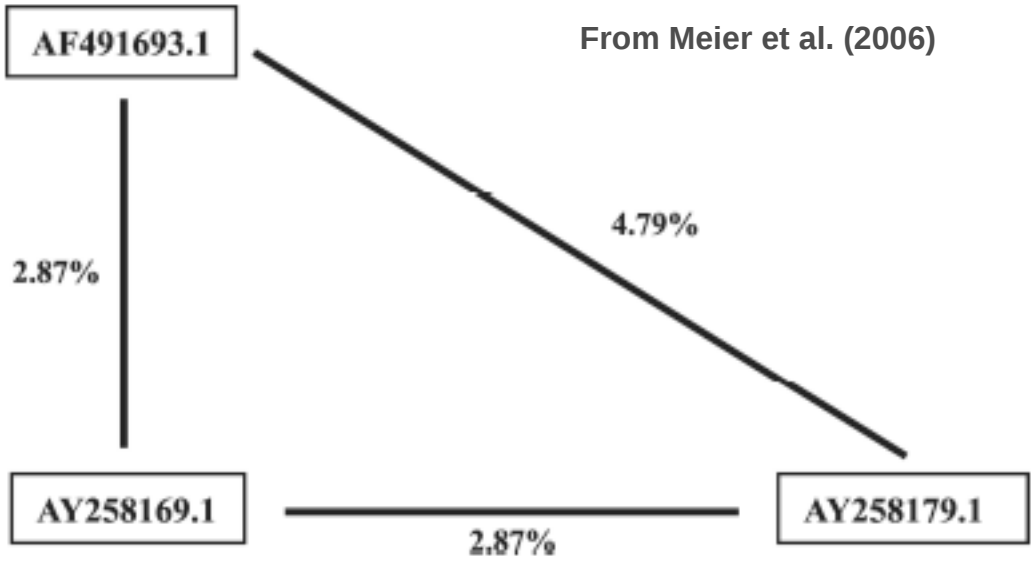


FIGURE 1. Pairwise distances for three *Anopheles* sequences (AF491693.1, AY258179.1 = *A. maculipennis*; AY258169.1 = *A. messae*). All belong to the same 3% DNA profile, although one pairwise distance exceeds the threshold.

=> Can lead to inconsistencies if formulated in that way

Impact of the clustering algorithm

- A distance *d(i,j)* <= *T* is called link
- An additional parameter, the “linkage fraction” *F*, determines how many links between an object and a cluster are necessary to include the object in the cluster

From Meier et al. (2006)

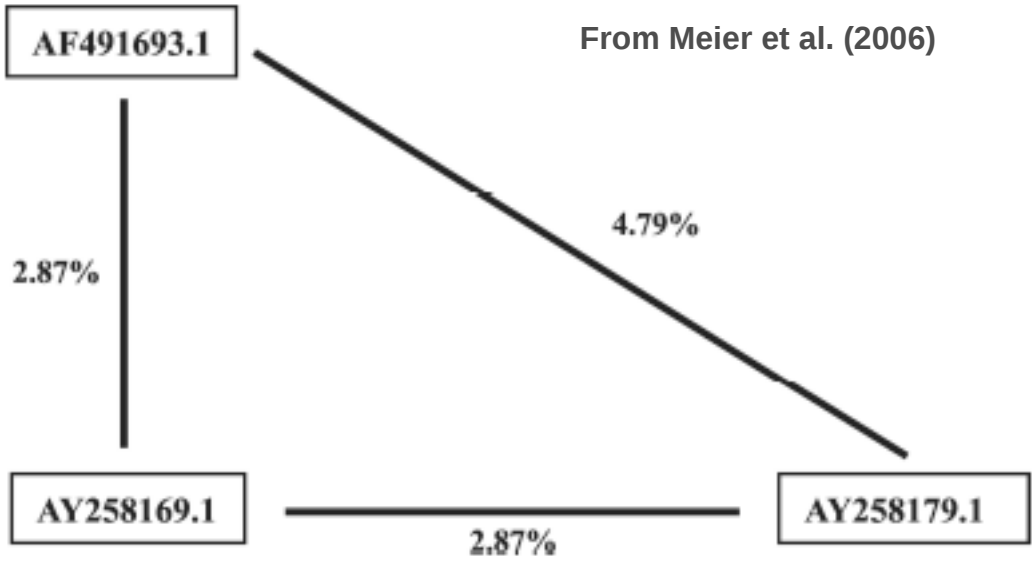


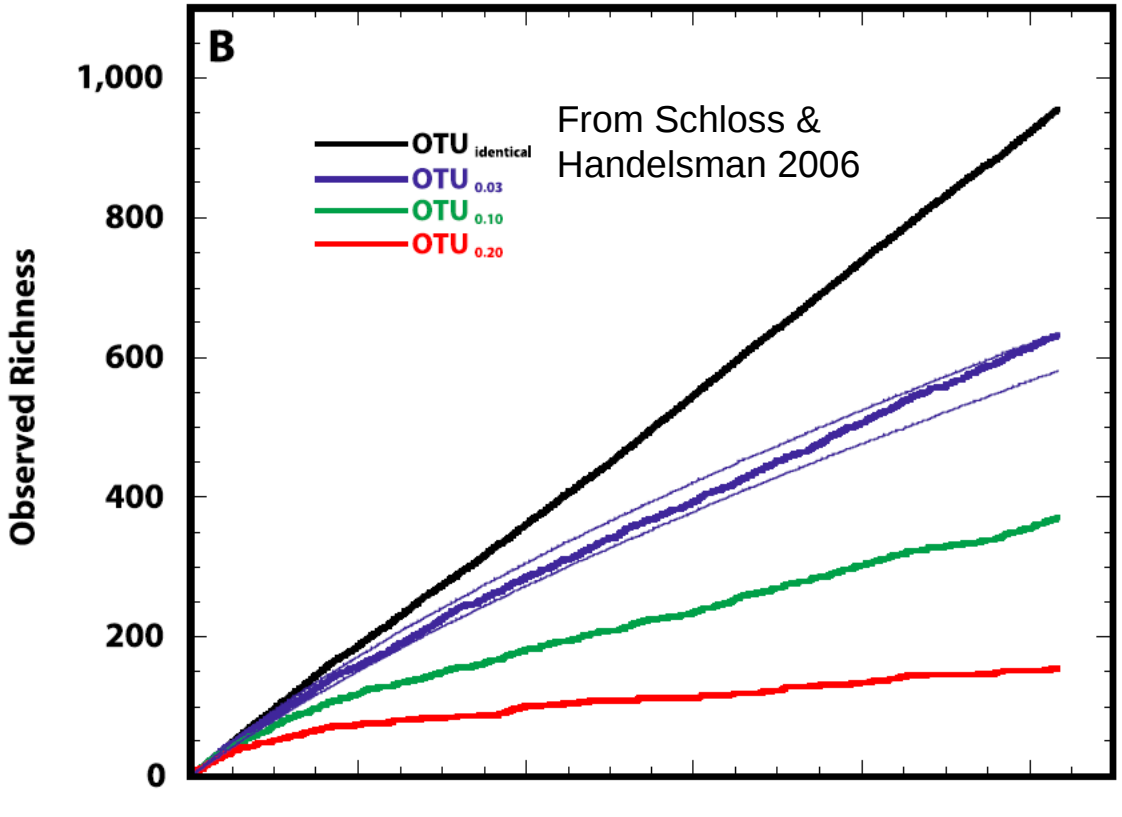
FIGURE 1. Pairwise distances for three *Anopheles* sequences (AF491693.1, AY258179.1 = *A. maculipennis*; AY258169.1 = *A. messae*). All belong to the same 3% DNA profile, although one pairwise distance exceeds the threshold.

=> Here, 1 cluster for *F* <= 0.5, but 2 clusters otherwise!
=> Lower *F* values allow higher within-cluster divergence

How to choose the clustering parameters?

Example:

- Species richness of soil bacteria estimated from 16S rDNA sequences
- Question: Has saturation been obtained?
- Obvious dependency on *T* (hidden one on *F*)



=> Choice of parameters has serious consequences for total biodiversity estimates

Criticisms of molecular taxonomy

Ongoing intense (and sometimes hostile) debate between molecular taxonomists and traditional morphologists, particularly in the context of DNA barcoding:

- Values of *T* used for clustering differ in the literature, even if applied to the same groups of organisms and molecular markers
- Values of *T* are often based on subjective criteria or on a tradition that emerged in recent years for the sake of comparability between studies
- Genetic divergence may differ between morphologically defined lineages
- A smaller distance (or a higher similarity) does not necessarily indicate a closer phylogenetic relationship

=> How can we maximize the agreement between traditional and molecular taxonomy?

Clustering optimization

- Partition := non-hierarchical, non-overlapping classification
- Many biological data are represented as partitions (e.g. assignment of sequences to species):
- Non-hierarchical clustering also results in a partition, e.g.:

Accession number	Organism
EF050035	Pseudoperonospora cubensis
EF174888	Peronospora aestivalis
EF174890	Peronospora sepium
EF174891	Peronospora fulva
EF174894	Peronospora lathyri-verni
EF174944	Peronospora orobi
...	...

Accession number	Cluster number
EF050035	29
EF174888	26
EF174890	25
EF174891	24
EF174894	27
EF174944	27
...	...

Approach:

- Use set of specimens identified using traditional techniques as reference points
- Determine the clustering parameters that maximize the agreement with a reference partition
- Do not require that full agreement can be obtained

Comparing partitions

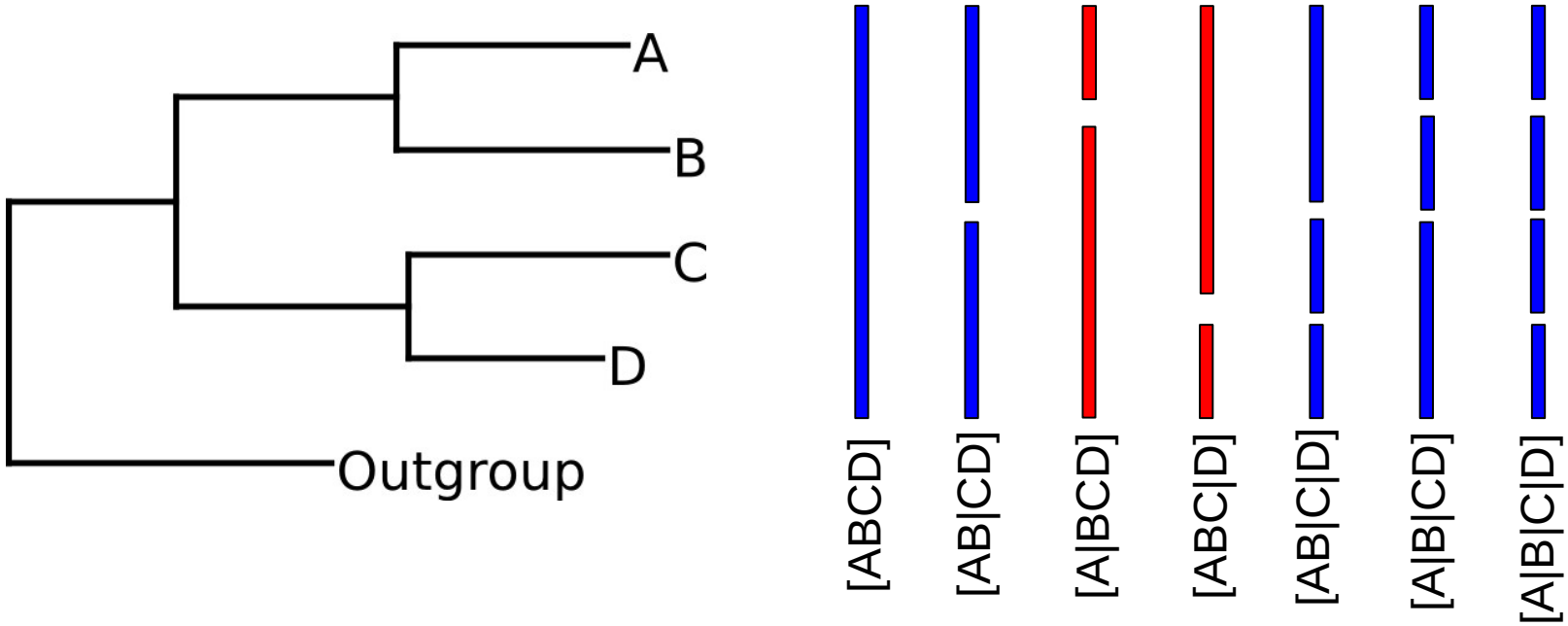
3 example partitions of 7 objects

Object	Fruit type	Colour	Condition
A	Apple	Green	Fresh
B	Lemon	Yellow	Fresh
C	Cherry	Red	Rotten
D	Apple	Green	Fresh
E	Cherry	Red	Fresh
F	Lemon	Yellow	Rotten
G	Apple	Green	Rotten

Observed values	Same	Different	Same	Different
Same	5	0	1	4
Different	0	16	8	8
Expected values	**Same**	**Different**	**Same**	**Different**
Same	1.19	3.81	2.14	2.86
Different	3.81	12.19	6.86	9.14
Rand Index	(5+16)/(5+0+0+16) = 1.0		(1+8)/(1+4+8+8) = 0.43	
Expected Index	(1.19+12.19)/ (5+0+0+16) = 0.64		(2.14+9.14)/(1+4+8+8) = 0.54	
Modified Rand Index (MRI)	(1.0-0.64)/(1.0-0.64) = 1.0		(0.43-0.54)/(1.0-0.54) = -0.24	

- Rand index (Rand 1971): traverse all pairs of objects and determine proportion of those being in the same cluster in *both* partitions or in a different cluster in *both* partitions
- Modified Rand index (Hubert & Arabie 1985): corrects for chance (by relating to the expected Rand index for two random partitions with the same cluster number and sizes)

Why trees don't help

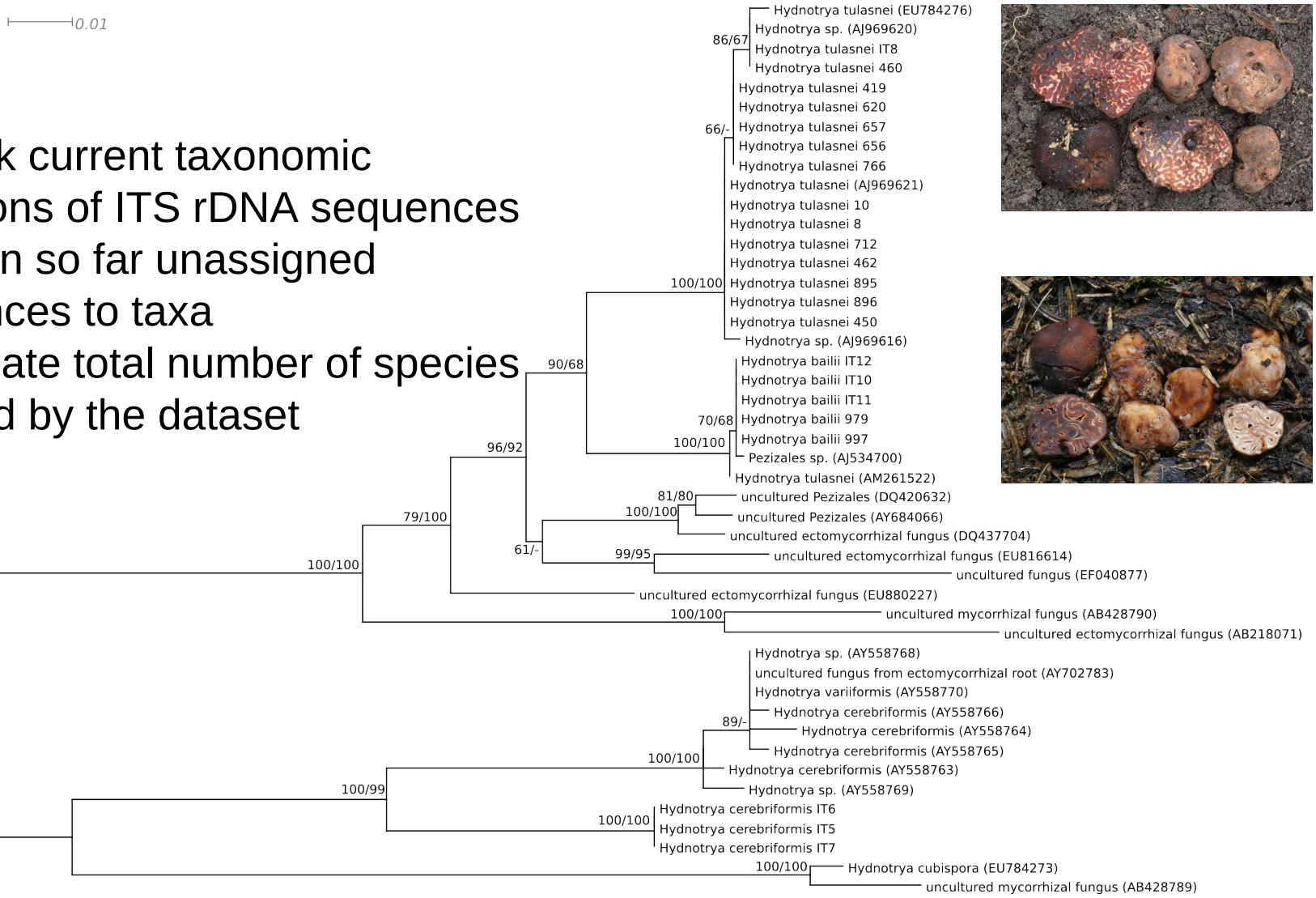


A phylogenetic tree rules out certain classifications (e.g. red ones), but is compatible with many others (blue ones)

Example 1: *Hydnotrya*

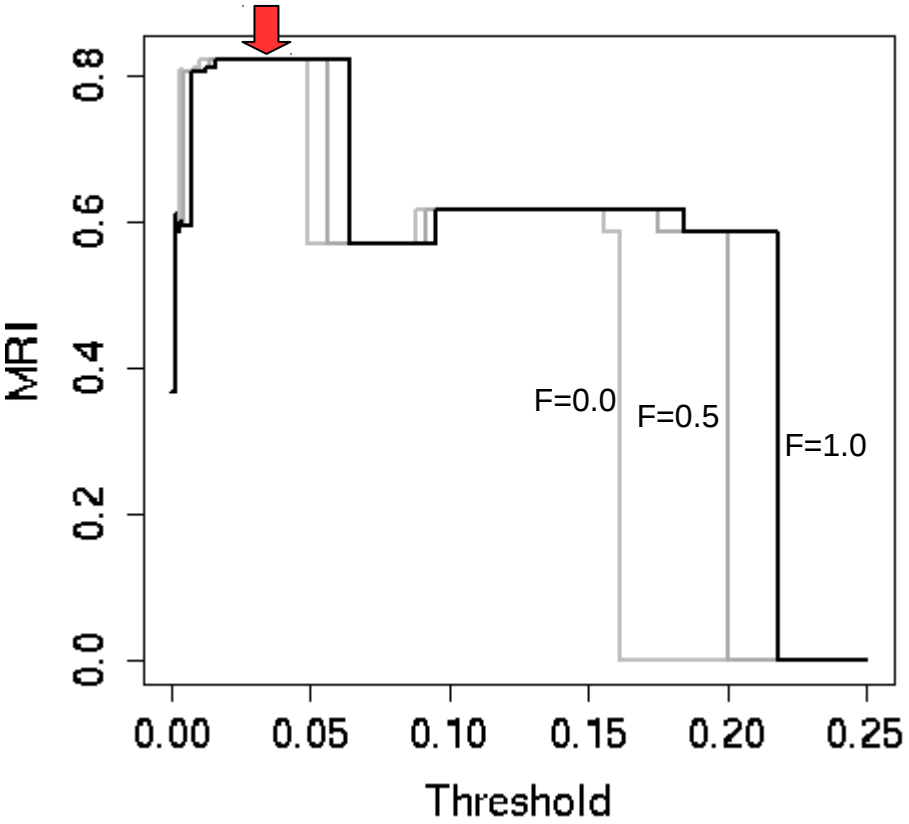
Tasks

- Check current taxonomic affiliations of ITS rDNA sequences
- Assign so far unassigned sequences to taxa
- Estimate total number of species covered by the dataset



Procedure

1) Restrict dataset to sequences with taxonomic affiliations



Accession number	Organism	Species name present?
EU784276	Hydnotrya tulasnei	Yes
AJ969620	Hydnotrya sp. G-Ht	No
AJ969621	Hydnotrya tulasnei	Yes
AJ969616	Hydnotrya sp. LB-Ht	No
AJ534700	Pezizales sp. B48	No
AM261522	Hydnotrya tulasnei	Yes
DQ420632	uncultured Pezizales	No
...	...	...

2) Conduct clustering optimization with reduced dataset

3) Place sequences without taxonomic affiliations back in the dataset

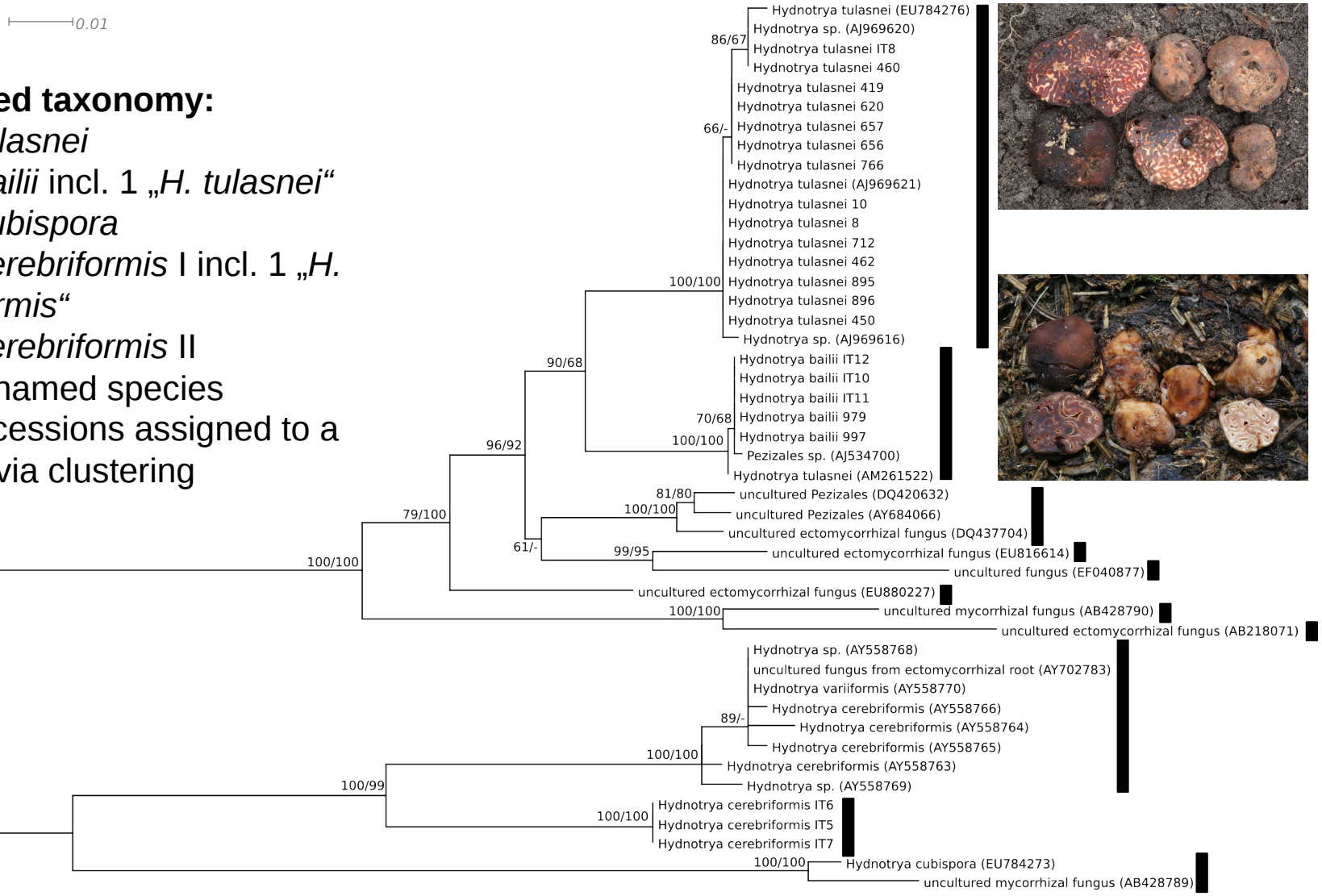
4) Conduct clustering with all sequences and optimized parameters

Stielow et al., The neglected hypogeous fungus *Hydnotrya bailii* Soehner (1959) is a widespread sister taxon of *Hydnotrya tulasnei* (Berk.) Berk. & Broome (1846). Mycological Progress 9: 195-203, 2010

Up to 50% of the MOTUs are novel species

Revised taxonomy:

- *H. tulasnei*
- *H. bailii* incl. 1 „*H. tulasnei*“
- *H. cubispora*
- *H. cerebriformis* I incl. 1 „*H. variiformis*“
- *H. cerebriformis* II
- 6 unnamed species
- 7 accessions assigned to a taxon via clustering



Self-cleaning of Genbank data

1) Restrict dataset to (a) sequences with taxonomic affiliations and (b) sequences with interpretable host names

Accession number	Organism	Specific host	Species name present?	Host present?
EF614964	Peronospora variabilis	Chenopodium album	Yes	Yes
EF614958	Peronospora sp. SMK20063	Chenopodium ambrosioides	No	Yes
EF614957	Peronospora sp. DAR45530	Chenopodium ambrosioides	No	Yes
EF614955	Peronospora farinosa f. sp. chenopodii	Chenopodium hybridum	Yes	Yes
EF174939	Peronospora sp. GG133		No	No
EF174924	Peronospora sp. HV956		No	No
EF174970	Peronospora trifoliorum		Yes	No
EF174963	Peronospora trifoliorum		Yes	No

2) Conduct clustering optimization with reduced datasets (a) and (b)

3) Check for coincidence of results (i.e. of optimal clustering parameters)

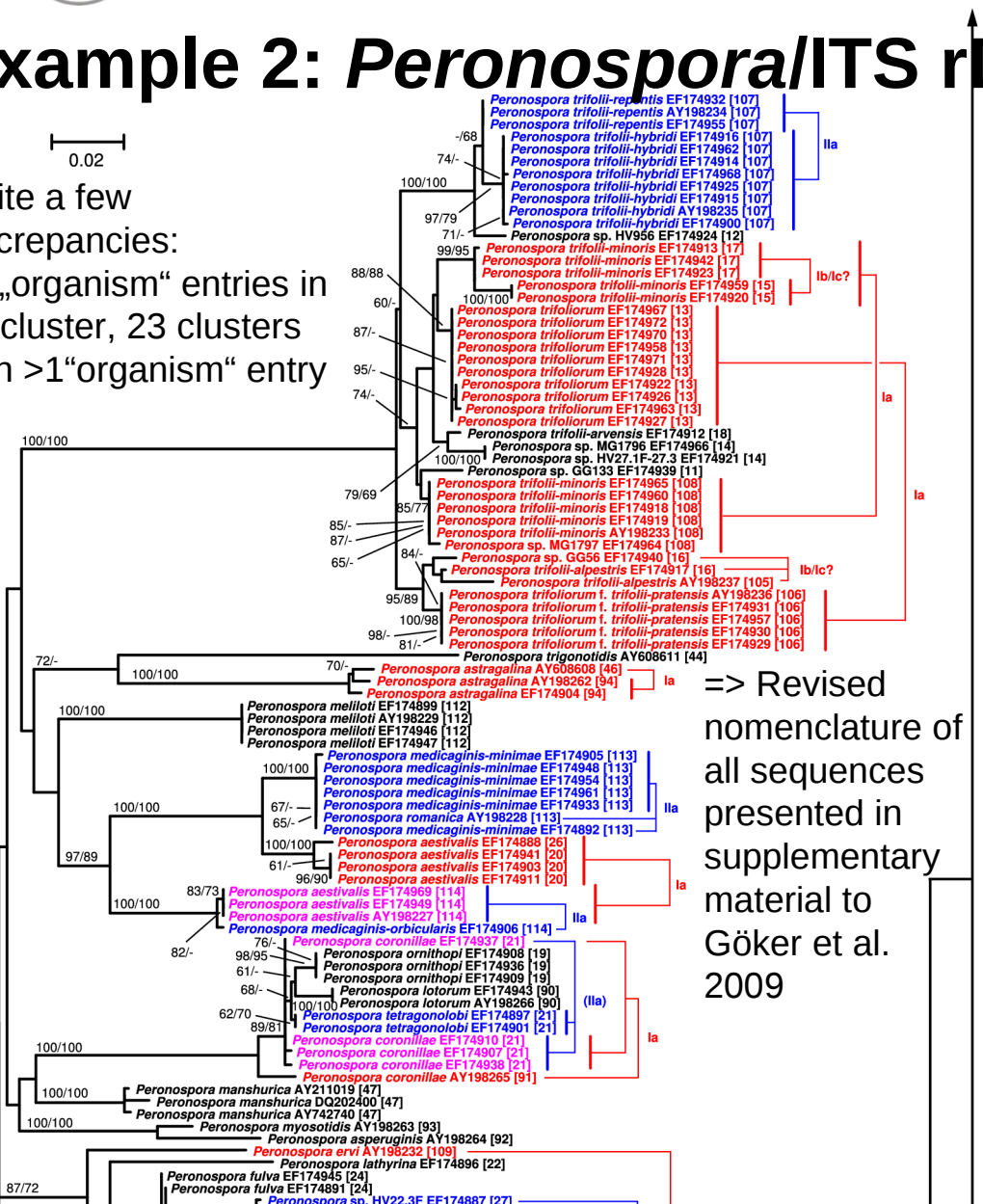
4) Place sequences without taxonomic affiliations or host information back in the dataset

5) Conduct clustering with all sequences and optimized parameters

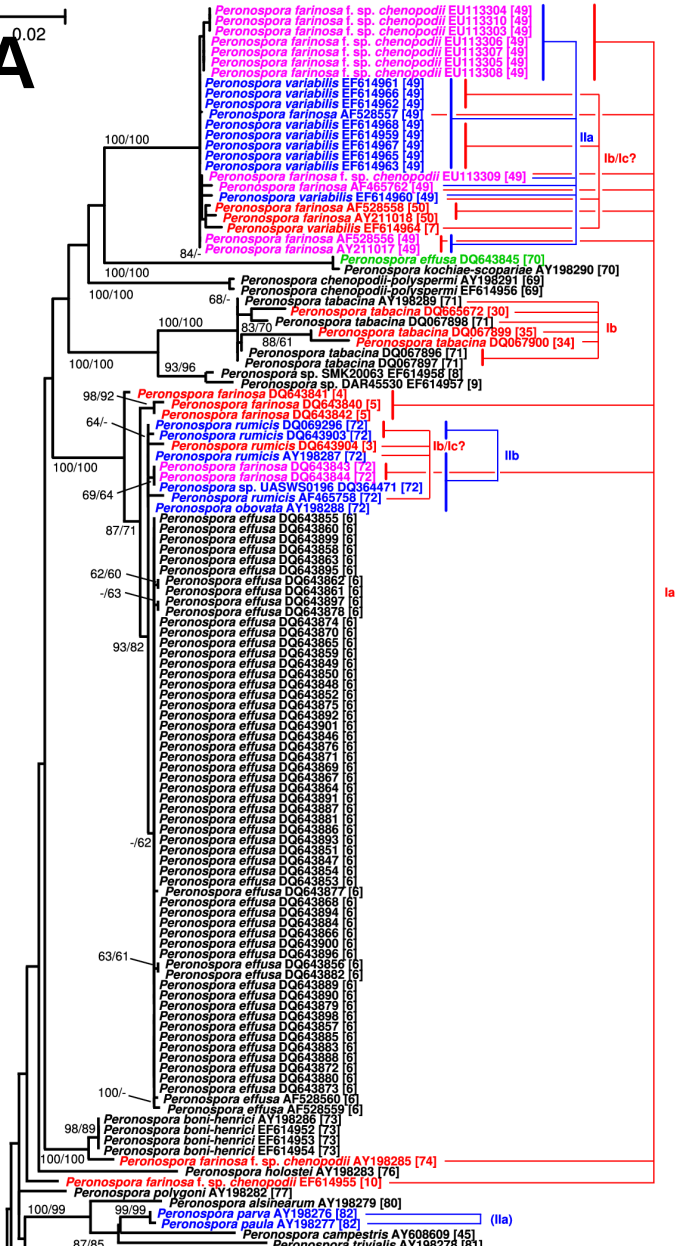
Göker et al., Molecular taxonomy of phytopathogenic fungi: a case study in *Peronospora*. PLoS ONE 4: e6319, 2009

Example 2: *Peronospora*ITS rDNA

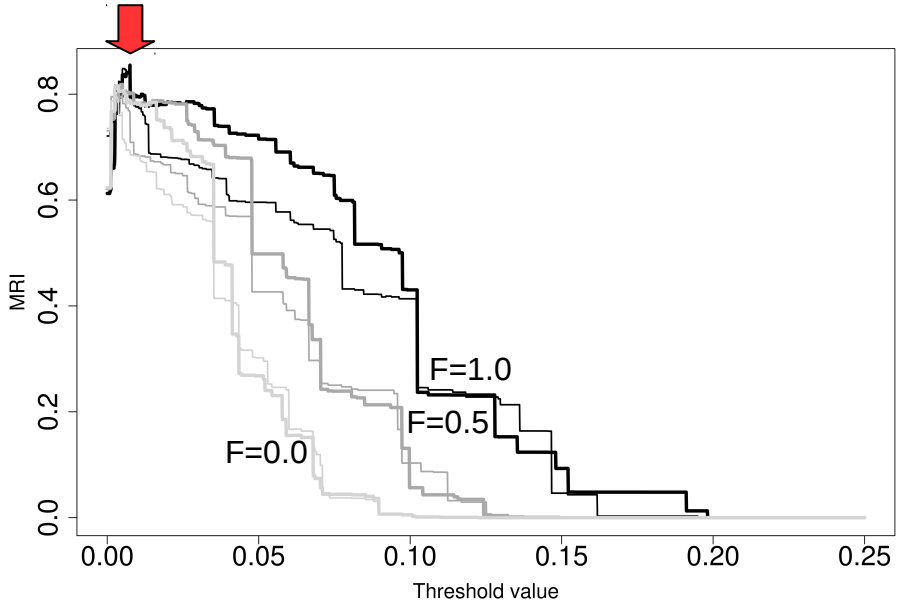
Quite a few discrepancies:
20 „organism“ entries in >1 cluster, 23 clusters with >1“organism“ entry

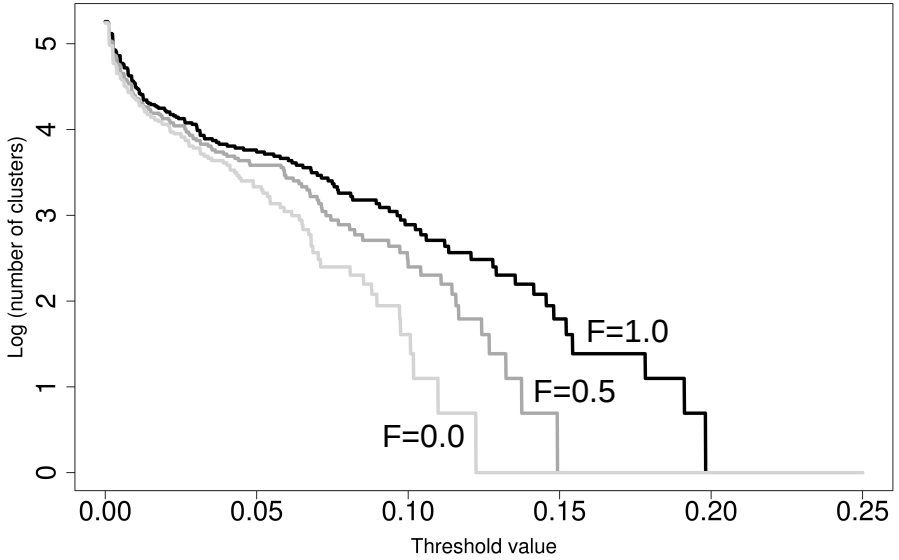


=> Revised nomenclature of all sequences presented in supplementary material to Göker et al. 2009



Host- and sequence-based species concept





- Taxonomy-based optimization: best result (MRI=0.85485) with T=0.0075 and F=1.0 (left picture: thick lines)
- Host-based optimization: best result (MRI=0.85204) with T=0.0075 and F=1.0 (left picture: thin lines) => *exactly the same optimum*
- Resulting in 117 clusters

Göker et al., Molecular taxonomy of phytopathogenic fungi: a case study in *Peronospora*. PLoS ONE 4: e6319, 2009

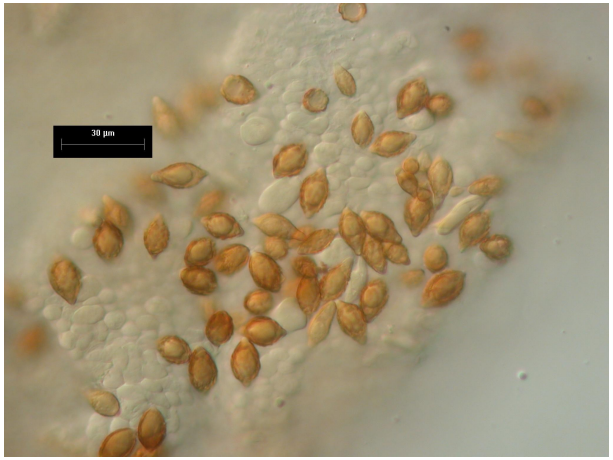
Example 3: *Hymenogaster* taxonomy

Tasks

- Determine the best morphological approach to species delimitation
- Once the best approach is identified, clarify remaining discrepancies with ITS rDNA data

Challenges in *Hymenogaster* morphology and taxonomy:

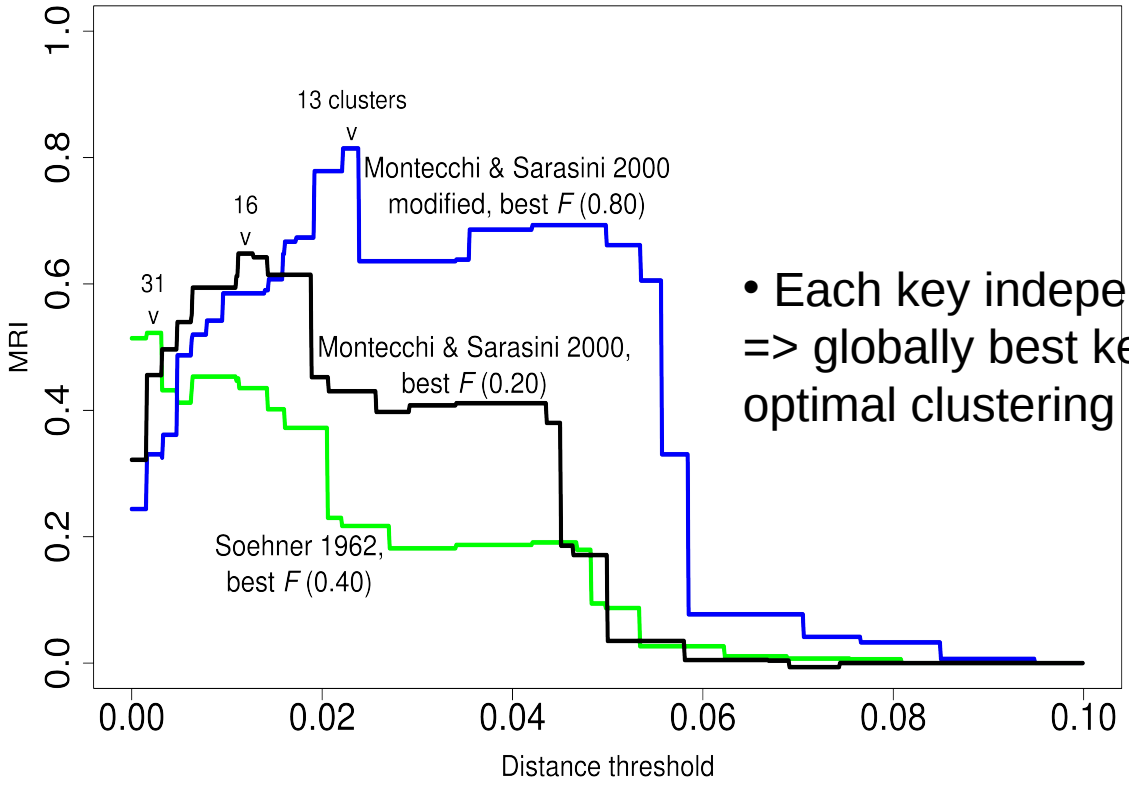
- Variability of basidiomata
- Variability of basidiospores
- Great variety in number of accepted species, e.g.
- Soehner (1960): 94 species
- Montecchi & Sarasini (2000): 17 species



H. arenarius basidiospores and basidiomata

Objective comparison of identification keys

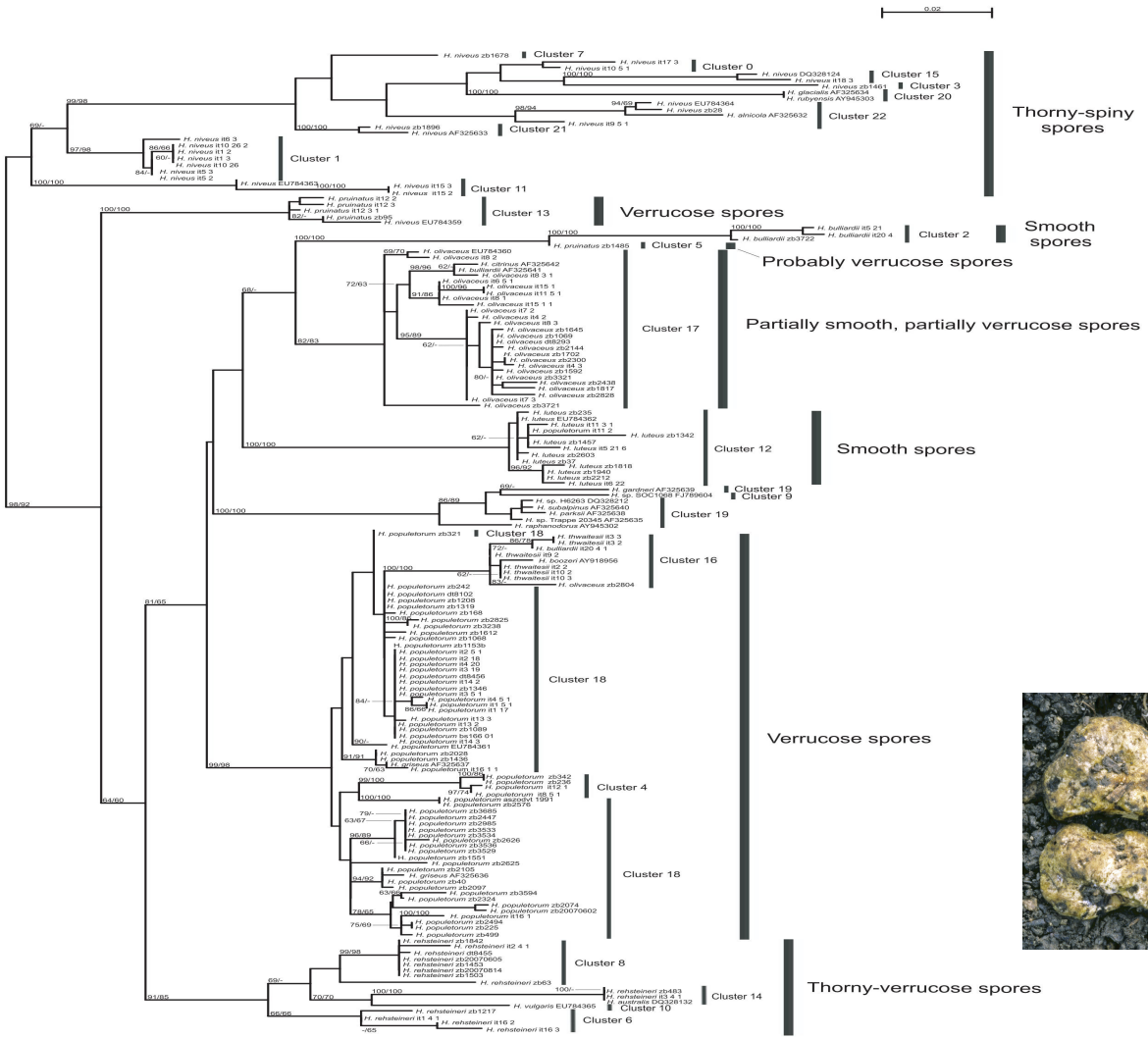
- ITS rDNA sequences obtained from 140 specimens from seven countries, mainly from Hungary and Germany
- Three keys (narrow vs. broad) used for identification => three reference partitions



- Each key independently optimized => globally best key ***and*** globally optimal clustering parameters

Stielow et al., Species delimitation in taxonomically difficult fungi: the case of *Hymenogaster*. PLoS ONE 6: e15614, 2011.

Revision of *Hymenogaster*



Clusters from optimal settings mapped on ITS rDNA ML tree

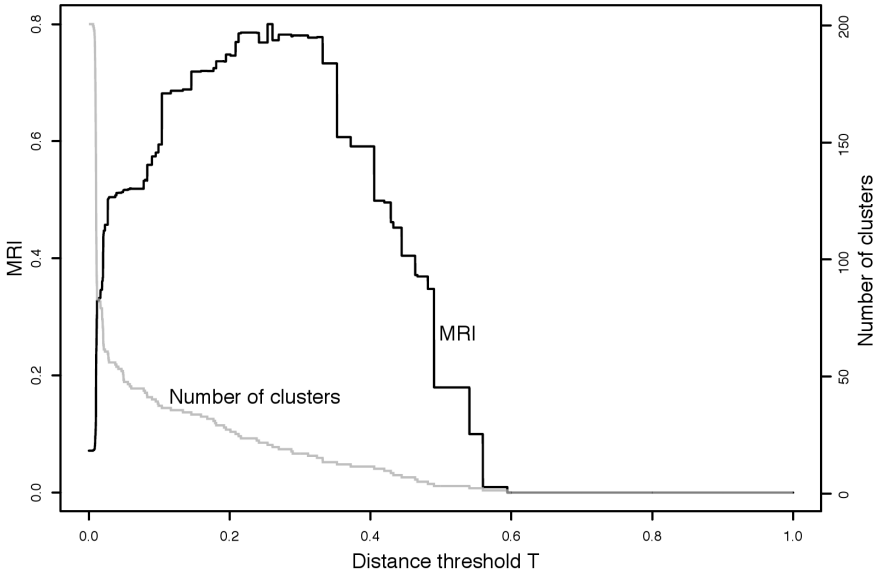
- Broadest species are optimal
- Just seasonal variability, no species boundary between *H. griseus* and *H. citrinus*
- Cryptic species in *H. niveus* remain
- Two novel species, *H. intermedius* and *H. huthii*
- Identification key for all European species according to new concept

Morphs of *H. citrinus*

Example 4: Planktonic Foraminifers

Tasks

- Determine the best alignment algorithm for PF SSU rDNA (highly length-variable and largely unalignable)
- Determine the best distance function
- Once the best approach is identified, clarify remaining discrepancies with morphology



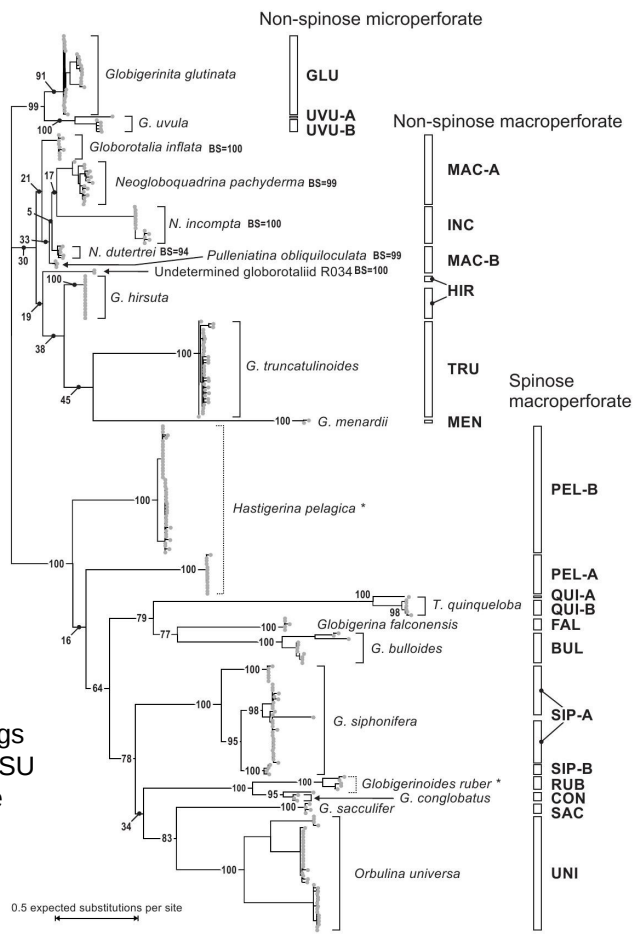
Globigerina bulloides
(http://www.foraminifera.eu/)

Solution

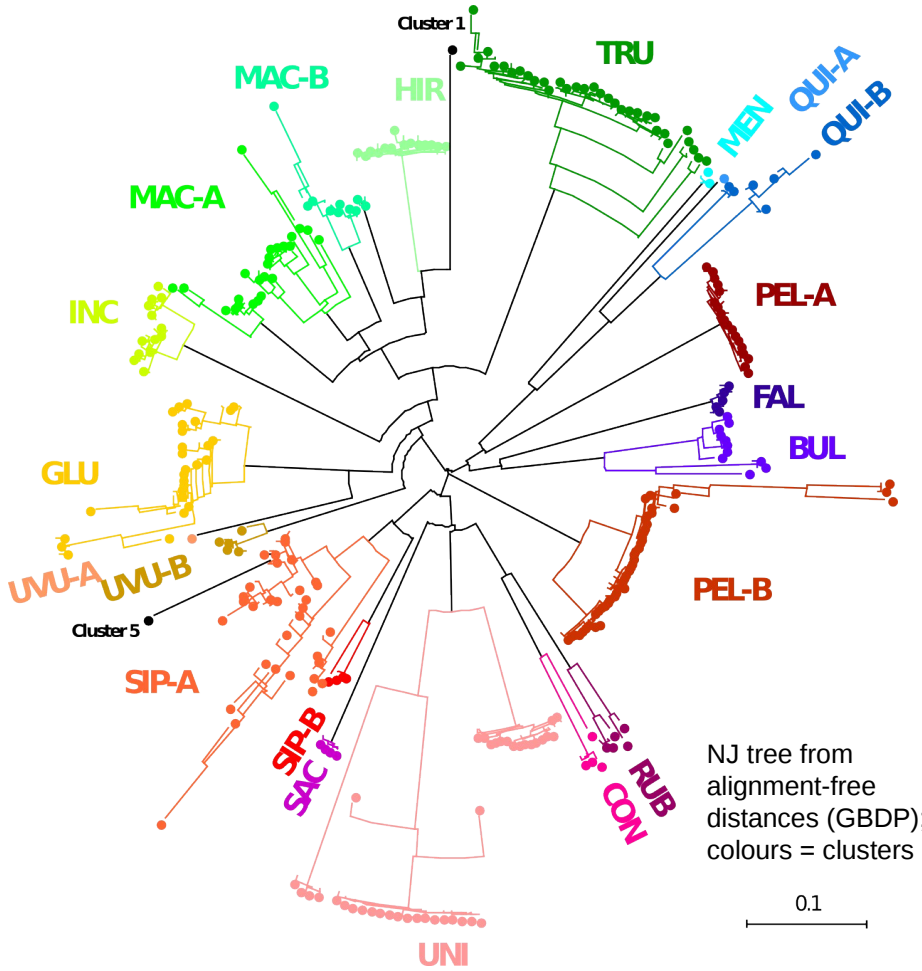
- Three-dimensional clustering optimization over (i) alignment; (ii) distance model; (iii) clustering parameters

Göker et al., A clustering optimization strategy for molecular taxonomy applied to planktonic foraminifera SSU rDNA. Evolutionary Bioinformatics 6: 97-112, 2010.

Alignment-free distances yield optimal clusters



Clusters from optimal settings mapped on SSU rDNA ML tree



NJ tree from alignment-free distances (GBDP); colours = clusters

Göker et al., A clustering optimization strategy for molecular taxonomy applied to planktonic foraminifera SSU rDNA. Evolutionary Bioinformatics 6: 97-112, 2010.

Summary: Clustering optimization...

- leads to MOTUs with highest agreement to traditional taxonomy, but it is **robust** against errors in such a reference partition
- connects traditional and modern taxonomic disciplines
- optimally accounts for both traditional species concepts and character divergence (maximizes both taxonomic **conservatism** and **consistency**)
- can be used to taxonomically cleanse data from INSDC (Genbank etc.)
- leads to biologically reasonable choices for alignment algorithms, distance functions and clustering parameters
- optimal parameters are also suitable for sequence **identification**
- is implemented in the **OPTSIL** software available at http://www.goeker.org/mg/clustering/ for all major operating systems