BOTANY &
MYCOLOGY
University of Tübingen

Clustering optimisation techniques to define biologically meaningful molecular operational taxonomic units

M. Göker

BOTANY & MYCOLOGY
University of Tübingen

Contents

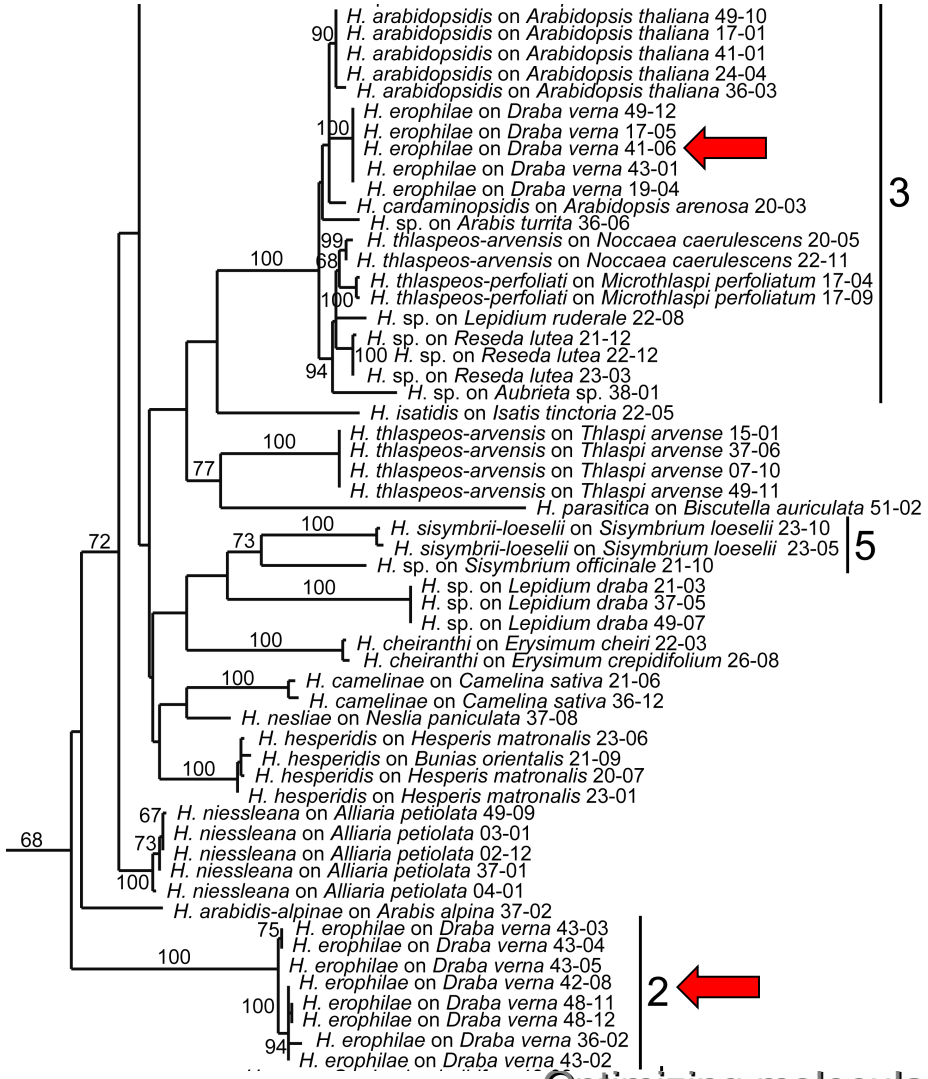
- The need for molecular taxonomy
- Methodological problems in molecular taxonomy
- Clustering optimization as a potential solution
- Outlook

The need for molecular taxonomy

- 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)

BOTANY & MYCOLOGY
University of Tübingen

(Pseudo-)cryptic species



- Example from the genus *Hyaloperonospora* (Peronosporales, Oomycetes)
- Combined ITS/LSU rDNA analysis (Göker et al., in press)
- Two genetically distinct but microscopically identical species on *Draba verna* host plants

Mislabelled sequences

- Example *Tuber* (743 ITS rDNA sequences): about 64 misidentifications
- Foraminifera (306 SSU rDNA sequences): 4 misidentifications, >= 5 unknown species, >= 1 synonym, >= 3 cryptic species
- *Hyaloperonospora*: majority of *H. parasitica* GenBank sequences is *H. arabidopsidis*

BOTANY & MYCOLOGY
University of Tübingen

Methodological problems in molecular taxonomy

- Distance threshold
- Clustering algorithm
- Distance calculation
- Sequence alignment

Threshold-based clustering

- Calculate pair-wise distances
- Principle: if distance between two sequences is <= predefined threshold, both belong to the same molecular operational taxonomic unit (MOTU)

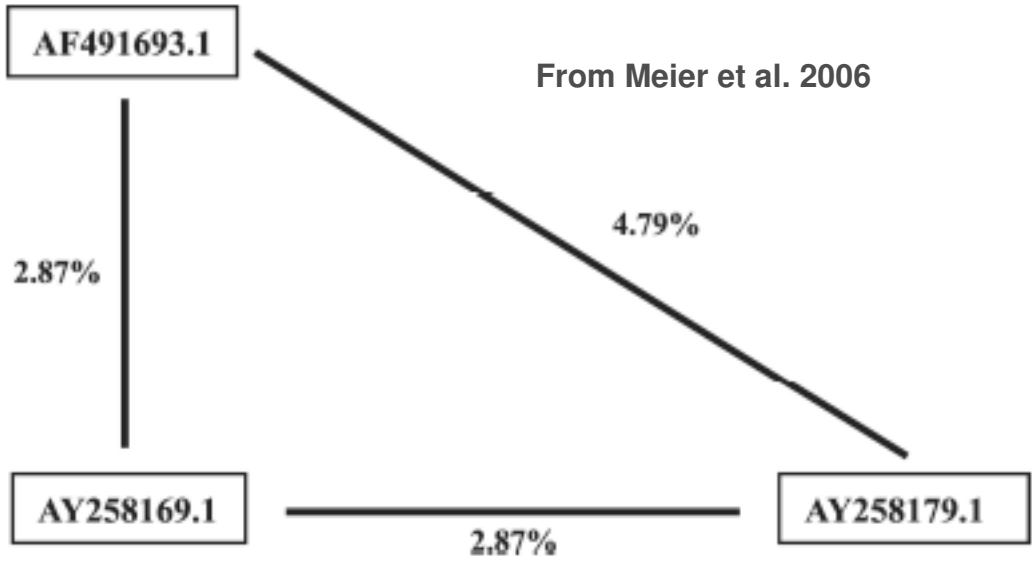


FIGURE 1. Pairwise distances for three *Anopheles* sequences (AF491693.1, AY258179.1 = *A. maculipennis*; AY258169.1 = *A. messae*). All belong to the same 3% DNA profile, although one pairwise distance exceeds the threshold.

- Can lead to inconsistencies if formulated in that way

Clustering algorithm

- A distance <= threshold is called link
- Additional parameter "linkage fraction" determines % of distances of object to a sequences in cluster that must be links for object to be included in cluster

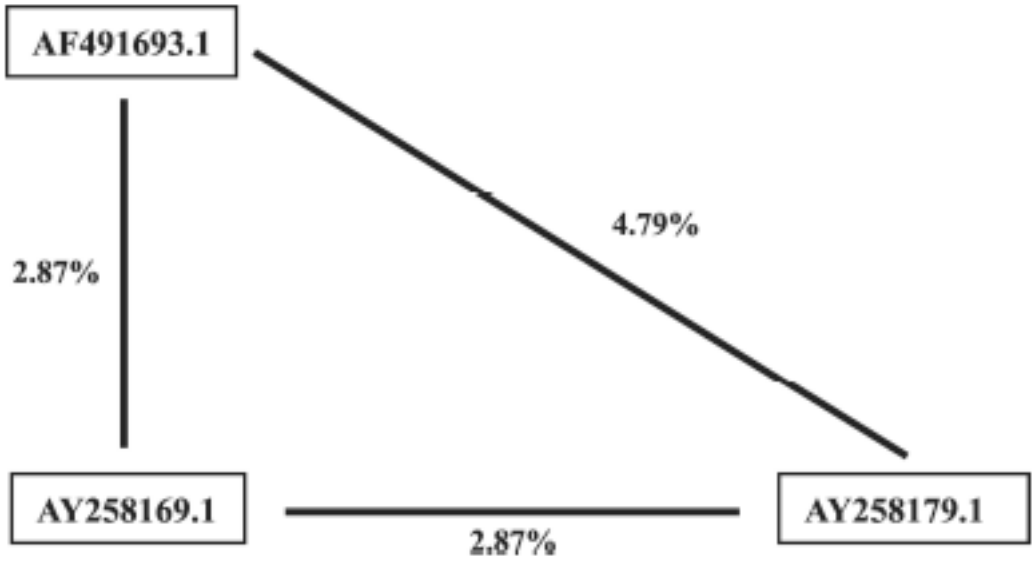


FIGURE 1. Pairwise distances for three *Anopheles* sequences (AF491693.1, AY258179.1 = *A. maculipennis*; AY258169.1 = *A. messae*). All belong to the same 3% DNA profile, although one pairwise distance exceeds the threshold.

- Here, single cluster for linkage fractions <= 50%, two clusters otherwise

Distance calculation

- Many distance formula available
- Uncorrected (“p”) distances used many times in lierature
- May be too simple for many datasets
- More complex formulae => which too choose, and how to optimize their parameters?

Alternatives to clustering

- Bayesian methods (Nielsen & Matz 2006)
- Decision theory (Abdo & Golding 2007)
- Neural networks (Zhang et al. 2008)

Features:

- Complexity: specific assumptions (e.g., from population genetics) and/or long training times
- Require 100% correct reference dataset for training
- But in many cases we do not know the true taxon boundaries
- Require sequence alignment

BOTANY & MYCOLOGY
University of Tübingen

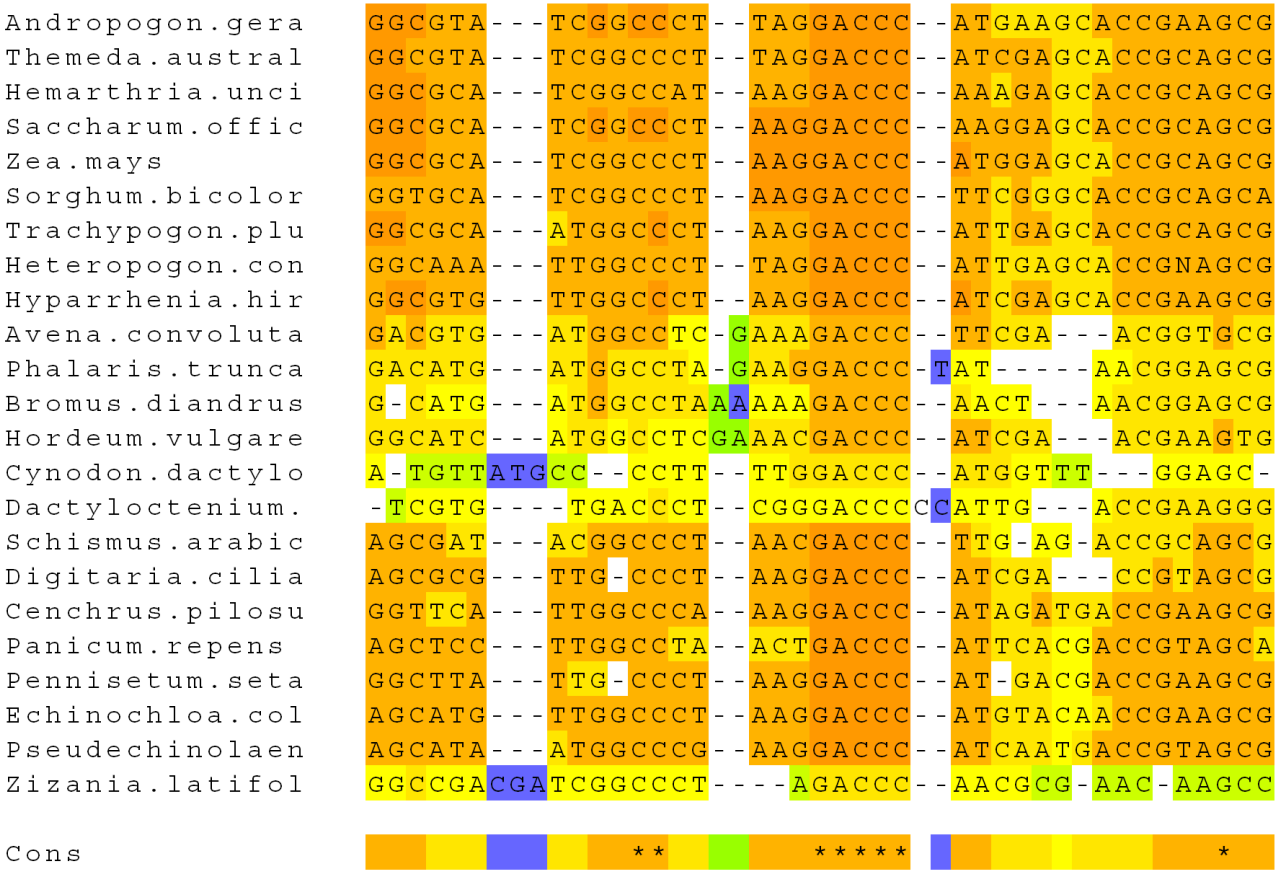
Alternatives to clustering

Phylogenetic trees

- What do we do if query sequences is placed as sister group of a subtree comprising only species X, not within? Is it X or not?
- What do we do if subtree is taxonomically heterogenous?
- In many cases we do not know the true taxon boundaries
- Mostly requires sequence alignment

BOTANY & MYCOLOGY
University of Tübingen

Sequence alignment



Alignment ambiguity present in many fast-evolving, non-coding markers such as ITS rDNA

Sequence alignment

- Phylogenetic inference may be more heavily influenced by sequence alignments then by inference methods
- Alignment artefacts known from phylogenetic literature
- Parallel problems likely in molecular taxonomy

BOTANY & MYCOLOGY
University of Tübingen

Sequence alignment

Alignment software and -options as far as deviating from the defaults	M749.NX: Bootstrap support for the split *Halophytophthora batemanensis* & *Pythium vexans* vs. other taxa	M751.NX: Bootstrap support for the split *Peronospora sparsa*, *Phytophthora arecae*, *Phytophthora palmivora* and *Phytophthora megakarya* vs. other taxa
Original Treebase alignment	100%	96%
Clustalw v1.81	95%	95%
Dialign v2.2.1 (-fa -n)	<50%	<50%
MAFFT v5.732 (maxiterate 1000)	!70%	57%
MUSCLE v3.52 (-stable)	!69%	75%
POA v2 (-do_progressive)	<50%	70%
POA v2	!70%	<50%
POA v2 (-do_global)	!99%	<50%

! = conflicting support

Re-examination of the Oomycetes ITS rDNA data presented in Cooke et al. (2000): Alignment instability of main results

BOTANY & MYCOLOGY
University of Tübingen

Alignment-free methods

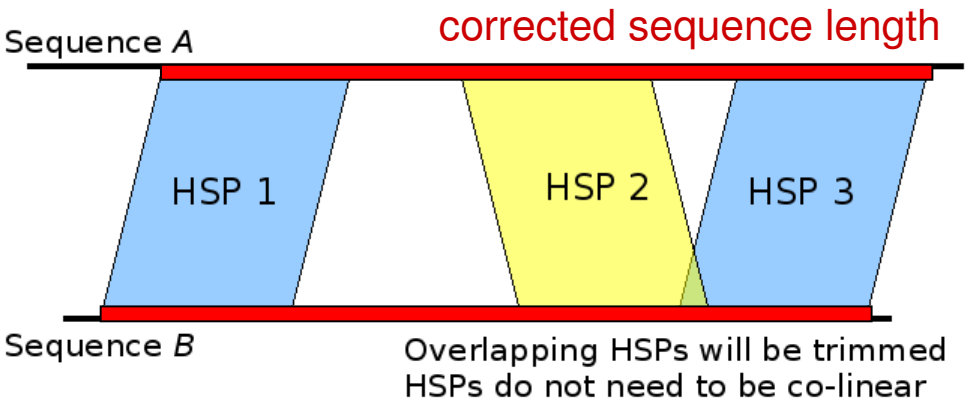
Example: GBDP, Gen(om)e BLAST distance phylogeny

Distance calculation: $d_1 := 1 - I / g$

Logarithmic: $d_2 := -\log (I / g)$

I **:= different coverage metrics using identity or BLAST score**

g **:= length of whole sequence(s) or *corrected* length**



Clustering optimization: the principle

- Partition := non-hierarchic, non-overlapping classification
- Many biological data represented as partitions (e.g., assignment of sequences to species)
- Non-hierarchical clustering also results in partition
- Approach: determine clustering parameters that maximize agreement with reference partition

Agreement between partitions

Modified Rand index (MRI; Hubert & Arabie 1985):

- **Agreement is lower if more pairs of objects are in the same cluster in one partition, but in different clusters in the other partition**
- **Corrects for random agreement and partition size**
- **Maximum at 1.0 for full agreement**
- **Approximately 0.0 for random partitions**

BOTANY & MYCOLOGY
University of Tübingen

Multidimensional optimization

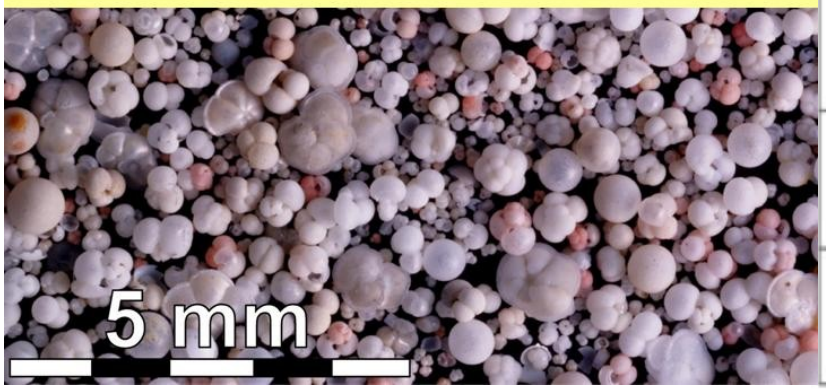
- *Ceteris paribus* principle: if other parameters are fixed, modified one responsible for changes in outcome
- Different sources of errors in reference partition and clustering partition => no bias
- Partition agreement used to optimize everything necessary to obtain clusters from unaligned sequences
- Clustering optimization implemented in OPTSIL program

BOTANY & MYCOLOGY
University of Tübingen

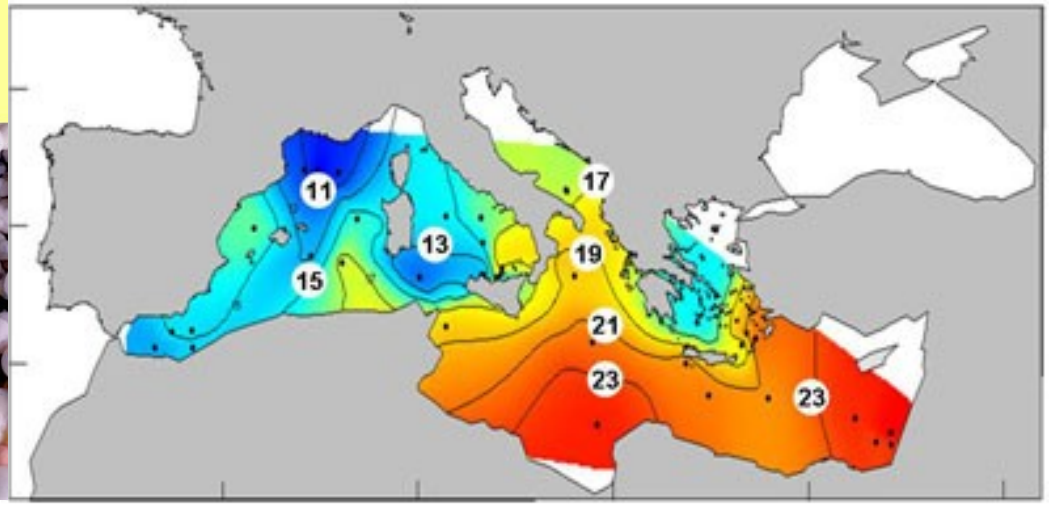
Example dataset

- Planktonic Foraminifera: one of the most important tools in biostratigraphy and paleoclimatology
- 306 partial or complete SSU rDNA sequences (50% from GenBank)
- Extremely difficult to align (up to 50% of the length unique expansion segments)
- Morphotaxonomy as reference partition: frequent misidentification

Planktonic foraminifera: shells of 25 – 50 % of all individuals are preserved in deep-sea sediments. Probably 1-10 % of all individuals that lived throughout the last 70 million years are preserved as fossils.

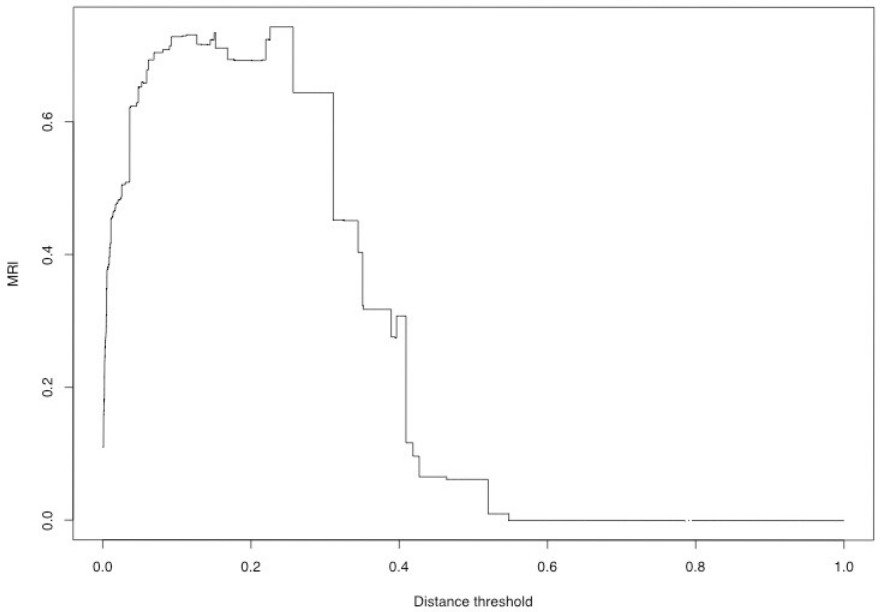


Summer temperature of ice-age Mediterranean reconstructed from fossil planktonic foraminifera

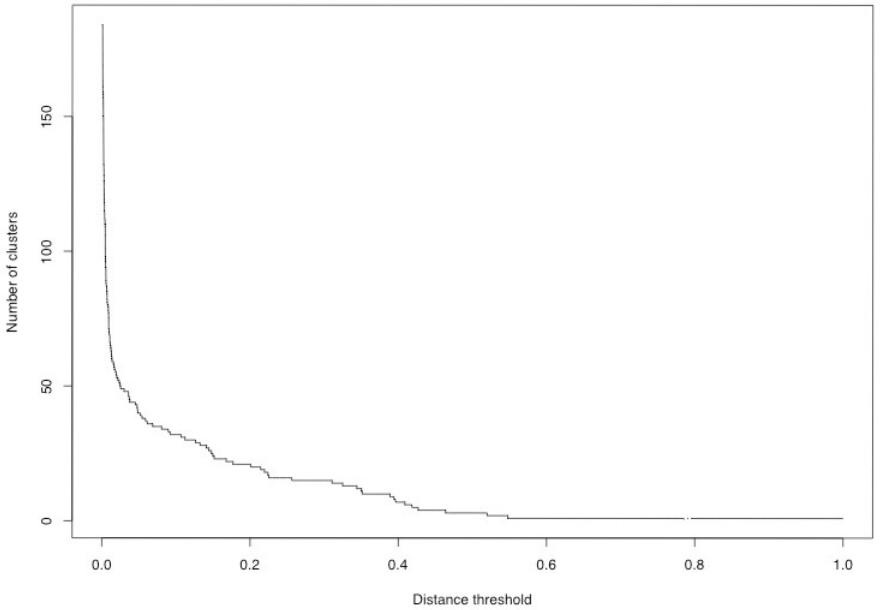


BOTANY & MYCOLOGY
University of Tübingen

Dimension 1: Distance threshold



Agreement with reference partition

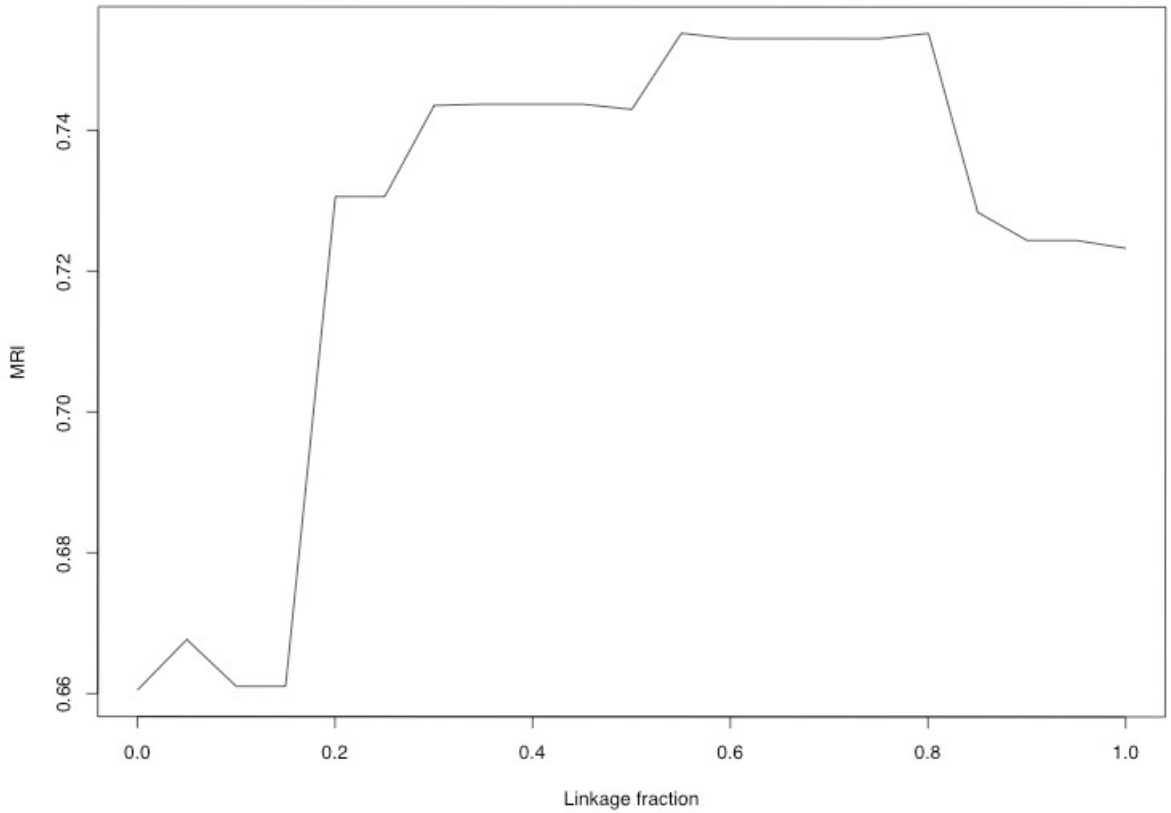


Number of clusters

Clustalw alignment, “p” distances, fixed linkage fractions of 0.5

BOTANY & MYCOLOGY
University of Tübingen

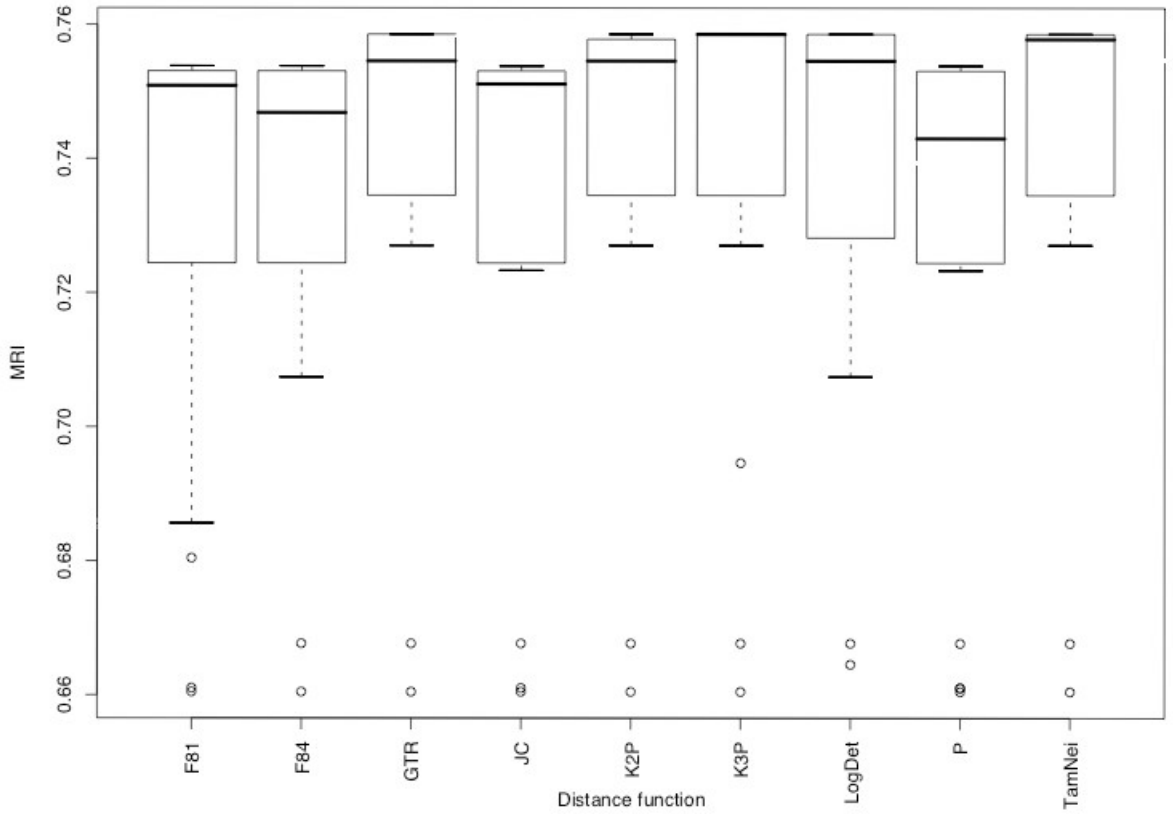
Dimension 2: Linkage fraction



Clustalw alignment, "p" distances: Best MRI values independently determined for distinct linkage fractions, each time optimizing the threshold

BOTANY & MYCOLOGY
University of Tübingen

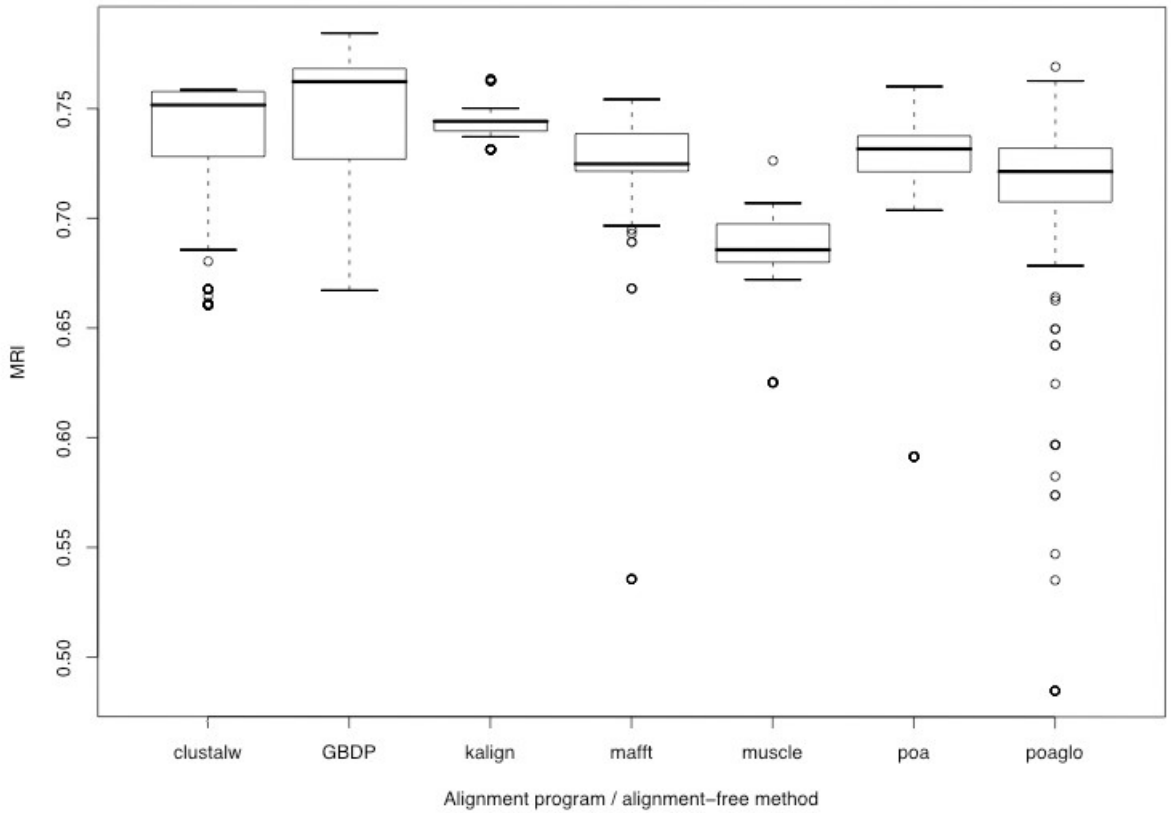
Dimension 3: Distance function



Clustalw alignment: Best MRI values independently determined for distinct distance functions, each time optimizing threshold and linkage fraction

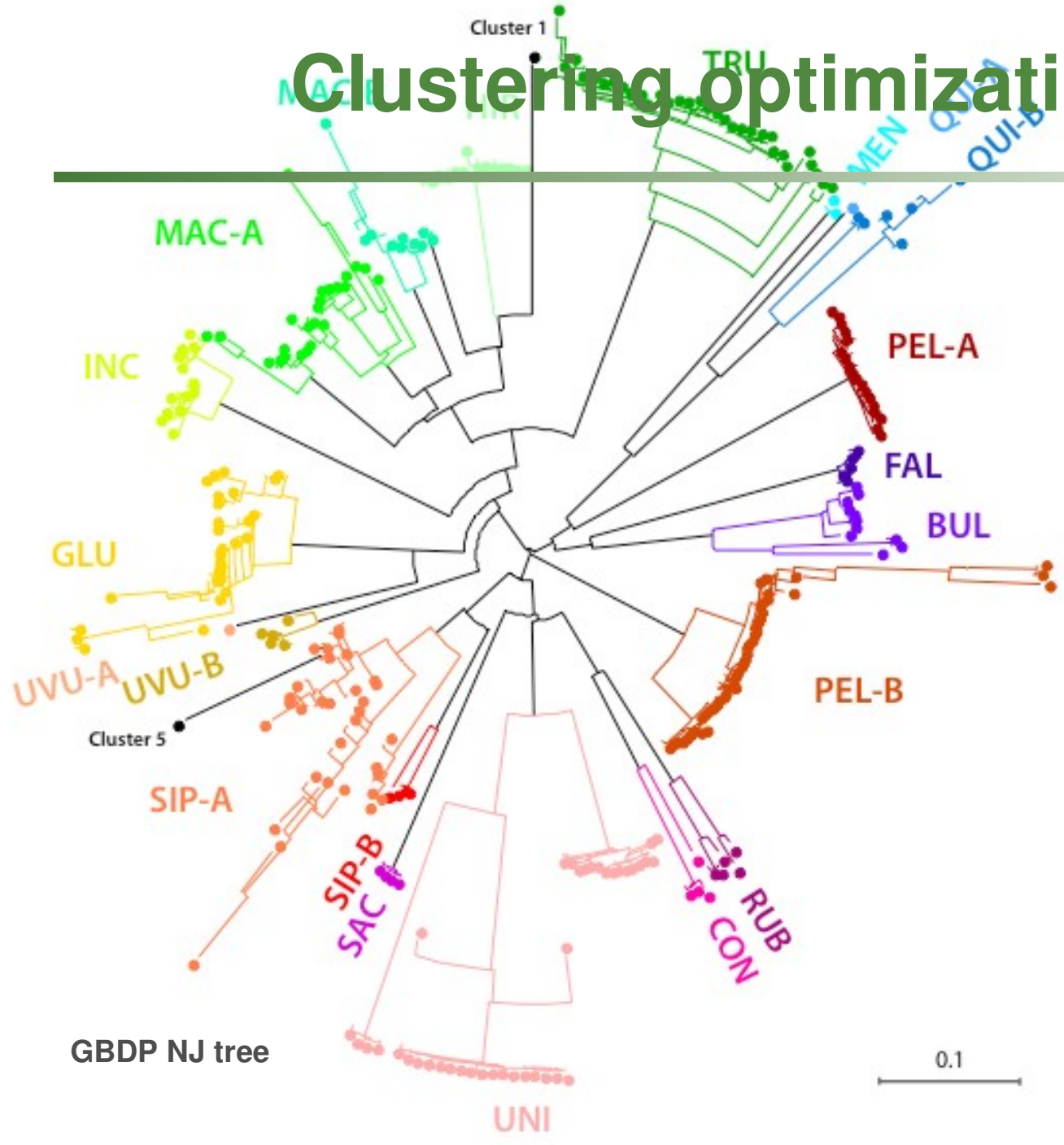
BOTANY & MYCOLOGY
University of Tübingen

Dimension 4: Sequence alignment



Best MRI values independently determined for distinct multiple sequence alignments programs and alignment-free methods, each time optimizing threshold, linkage fraction, and distance function

Clustering optimization: Result



GBDP NJ tree

- Highest MRI of 0.7845 obtained with corrected GBDP
- 23 clusters
- Best MRI for alignment program: 0.7693
- 3 misidentifications
- 1 misnomer from GenBank
- NCBI taxonomy synonyms resolved
- 5 sequences identified for the 1st time
- 2 cryptic species confirmed
- >= 1 cryptic species discovered

BOTANY & MYCOLOGY
University of Tübingen

Robustness of clustering optimization

Figure 7

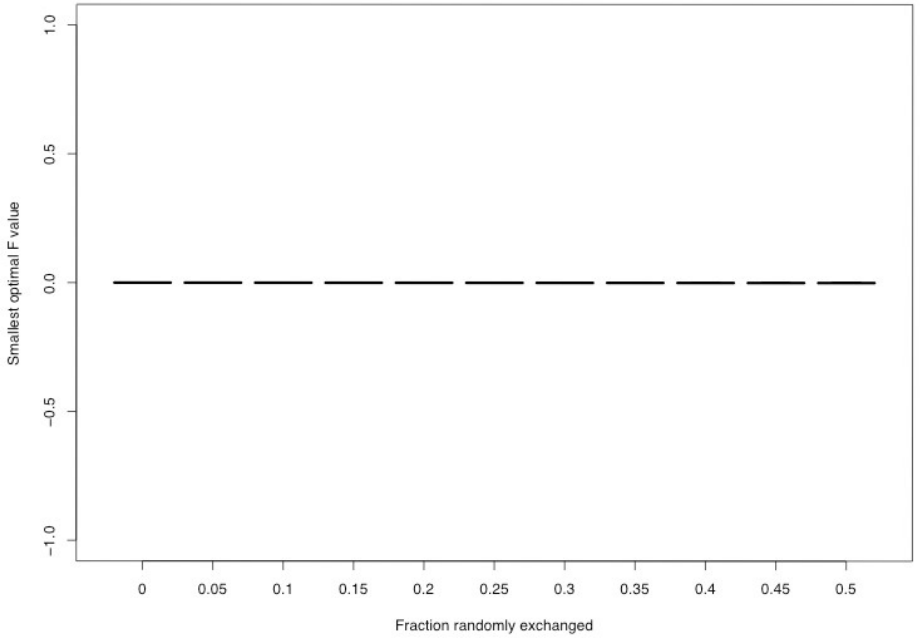
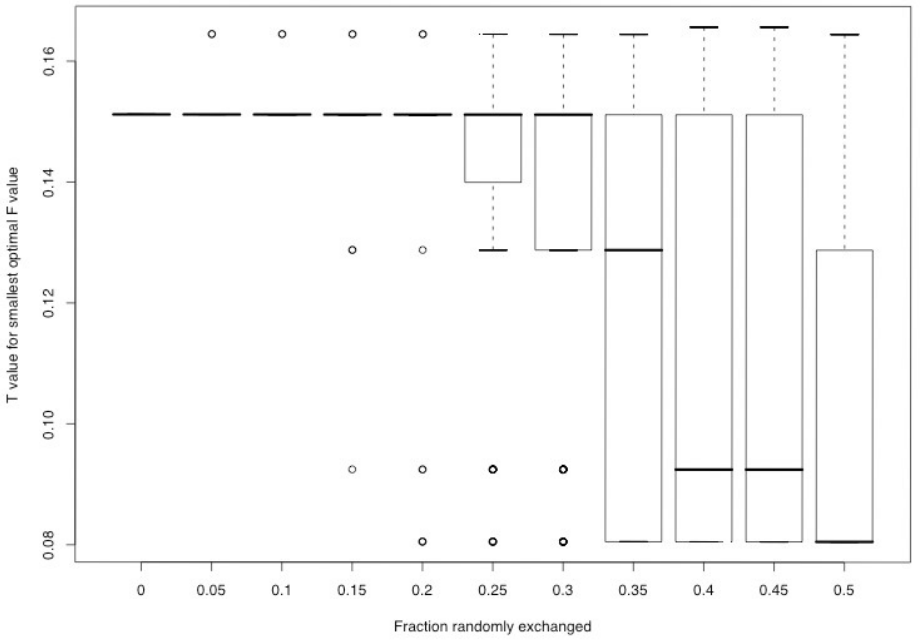


Figure 9



Random perturbation experiments with empirical data => optimal T and F values practically constant for up to 10-20% random exchanges (i.e., 15-25% introduced errors in reference partition).

Alternative reference partitions

- Taxonomy
- (Discrete) morphological characters
- Host species of parasites/mutualists
- Geographic origin
- Affiliation to environmental probes

BOTANY & MYCOLOGY
University of Tübingen

Alternative reference partitions

Organisms	Locus	Reference	Source	Number of sequences	Highest MRI
Planktonic Foraminifera	SSU rDNA	Morphospecies / NCBI taxonomy	Göker et al., in prep.	306	0.7845
Litoria spp.	cox2	Morphospecies / NCBI taxonomy	Schneider et al., 1998	178	1.0000
Astraptes spp.	cox2	Morphospecies / NCBI taxonomy	Hebert et al., 2004	466	0.9540
Tuber spp.	ITS rDNA	Morphospecies / NCBI taxonomy	Marjanovic et al., in prep.	743	0.9394
Peronospora spp.	ITS rDNA	Morphospecies / NCBI taxonomy	Garcia-Blazquez et al., in prep.	335	0.9162
Peronospora spp.	ITS rDNA	Plant host species	Garcia-Blazquez et al., in prep.	364	0.8646
Sebacinales	rDNA	Probe affiliation	Setaro et al., under review	81	0.6408

Outlook

- Inclusion of improved clustering methods
- Threshold-free distance methods combined with partially erroneous reference partitions possible?
- Automated correction of sequence sets from public databases
- Determination of the best-suited loci for molecular taxonomy and barcoding

BOTANY & MYCOLOGY
University of Tübingen

Summary

Optimization of the agreement between partitions...

- **leads to MOTUs with highest agreement with traditional taxonomy, even if the latter contains errors**
- **can be used together with different types of reference partition (e.g., if a traditional taxonomy is unavailable)**
- **is robust against errors in the reference partitions**
- **leads to biologically reasonable choices for alignment, distance, and clustering methods and their parameters**
- **leads to method and method parameter choices that are also suitable for sequence identification**

BOTANY & MYCOLOGY
University of Tübingen

Acknowledgements

Cordial thanks to I. Kottke (Invitation!), S. Setaro (Sebacinales), Z. Marjanovic, T. Grebenc (*Tuber*), M. Kucera, G. Grimm, R. Aurahs (Foraminifera), and A. Auch (GBDP)