Biologia (2022) 77:569–582
https://doi.org/10.1007/s11756-021-00971-y

ORIGINAL ARTICLE

18S and ITS2 rDNA sequence-structure phylogeny of *Prototheca* (Chlorophyta, Trebouxiophyceae)

Tanja Plieger[1] · **Matthias Wolf**[1]

Received: 29 September 2021 / Accepted: 18 November 2021 / Published online: 2 December 2021

Abstract
Protothecosis is an infectious disease caused by organisms currently classified within the green algal genus *Prototheca*. The disease can manifest as cutaneous lesions, olecranon bursitis or disseminated or systemic infections in both immunocompetent and immunosuppressed patients. Concerning diagnostics, taxonomic validity is important. *Prototheca*, closely related to the *Chlorella* species complex, is known to be polyphyletic, branching with *Auxenochlorella* and *Helicosporidium*. The phylogeny of *Prototheca* was discussed and revisited several times in the last decade; new species have been described. Phylogenetic analyses were performed using ribosomal DNA (rDNA) and partial mitochondrial cytochrome b (*cytb*) sequence data. In this work we use Internal Transcribed Spacer 2 (ITS2) as well as 18S rDNA data. However, for the first time, we reconstruct phylogenetic relationships of *Prototheca* using primary sequence and RNA secondary structure information simultaneously, a concept shown to increase robustness and accuracy of phylogenetic tree estimation. Using encoded sequence-structure data, Neighbor-Joining, Maximum-Parsimony and Maximum-Likelihood methods yielded well-supported trees in agreement with other trees calculated on rDNA; but differ in several aspects from trees using *cytb* as a phylogenetic marker. ITS2 secondary structures of *Prototheca* sequences are in agreement with the well-known common core structure of eukaryotes but show unusual differences in their helix lengths. An elongation of the fourth helix of some species seems to have occurred independently in the course of evolution.

Keywords 18S · ITS2 · Phylogeny · *Prototheca* · Secondary structure

Abbreviations

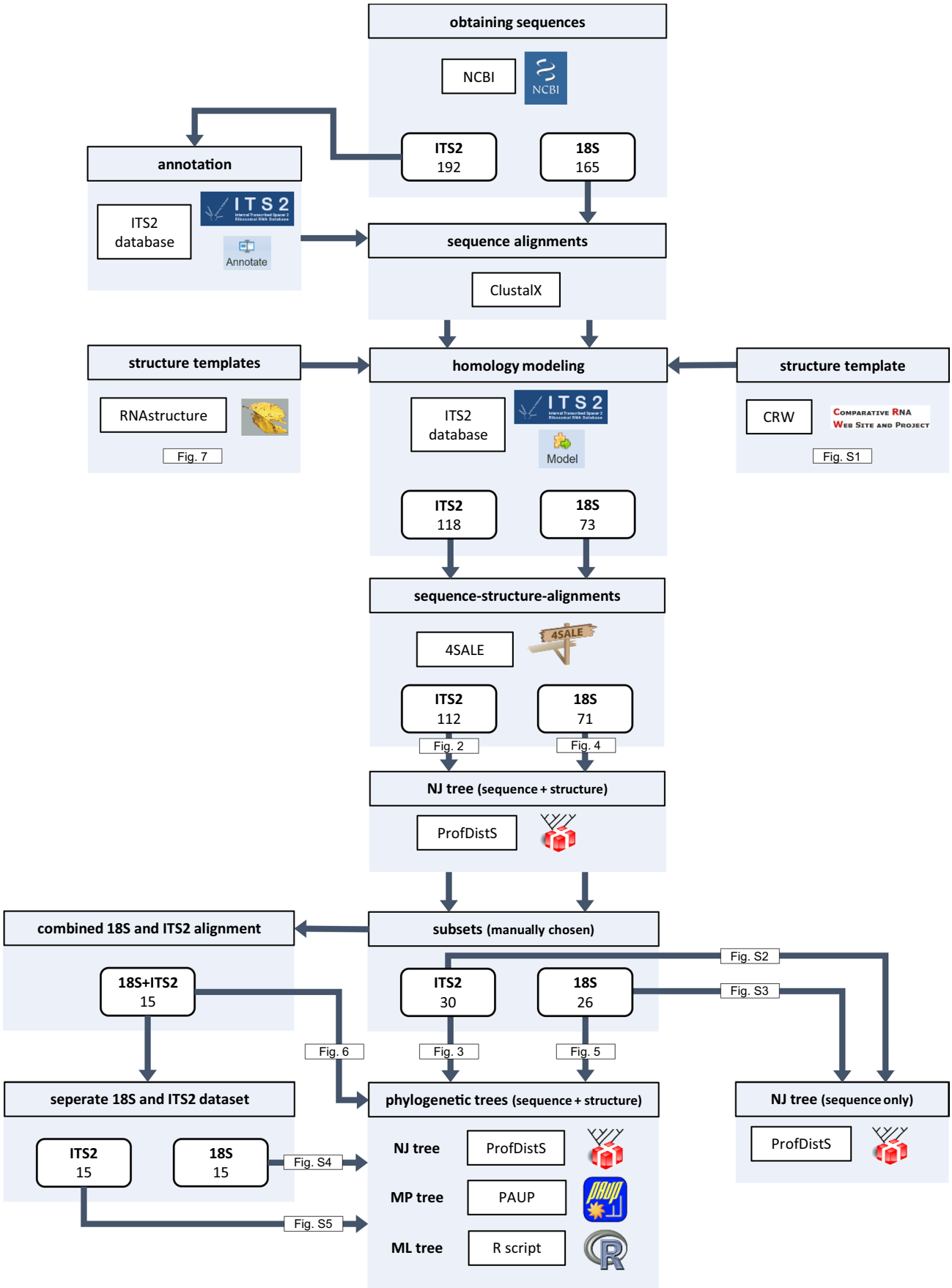
ITS2	Internal Transcribed Spacer 2
ML	Maximum Likelihood
NJ	Neighbor-Joining
MP	Maximum Parsimony

Introduction

According to Algaebase (Guiry and Guiry, 2021), organisms, colorless and apochlorotic, without chloroplasts and pyrenoids, currently classified as *Prototheca* W.Krüger, 1894 (Chlorophyta, Trebouxiophyceae) are widely distributed from temperate to tropical conditions in both fresh and marine waters. *Prototheca*, closely related to the *Chlorella* Beyerinck, 1890 species-complex, is polyphyletic with *Helicosporidium* D.Keilin, 1921 and *Auxenochlorella* (I. Shihira & R.W.Krauss) T.Kalina & M.Puncochárová, 1987 branching within clades of *Prototheca* species (e.g. Bakuła et al. 2020; Shave et al. 2021). *Prototheca* and *Chlorella* are the only known algal genera including disease-causing organisms in humans (Jagielski et al. 2019). *Prototheca* is associated with conditions termed protothecosis (Guiry and Guiry, 2021). Concerning diagnostics, taxonomic validity is important – in particular concerning the pathology associated taxa *P. wickerhamii* K.Tubaki & M.Soneda, 1959 and *P. zopfii* W.Krüger, 1894 (type species). Several species of *Prototheca* are rare opportunistic pathogens (Huerre et al. 1993) in humans (Lass-Flörl et al. 2007), other mammals (e.g. Möller et al. 2007; Marques et al. 2008) and fish (Jagielski et al. 2017). Protothecosis is an infectious disease, which often spreads through contact with contaminated water (Jagielski and Lagneau, 2007). The first case of protothecosis in humans was described in 1964 (Davies et al. 1964). The disease manifests in three clinical forms: cutaneous lesions, olecranon bursitis and disseminated or systemic infections (Leiman et al. 2004; Lass-Flörl

✉ Matthias Wolf
matthias.wolf@biozentrum.uni-wuerzburg.de

[1] Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany


obtaining sequences
NCBI
ITS2
192
18S
165
annotation
ITS2
database
Annotate
sequence alignments
ClustalX
structure templates
RNAstructure
Fig. 7
homology modeling
ITS2
database
Model
structure template
CRW
Comparative RNA Web Site and Project
Fig. S1
ITS2
118
18S
73
sequence-structure-alignments
4SALE
ITS2
112
18S
71
Fig. 2
Fig. 4
NJ tree (sequence + structure)
ProfDistS
combined 18S and ITS2 alignment
subsets (manually chosen)
Fig. S2
18S+ITS2
15
ITS2
30
18S
26
Fig. S3
Fig. 6
Fig. 3
Fig. 5
seperate 18S and ITS2 dataset
phylogenetic trees (sequence + structure)
NJ tree (sequence only)
ProfDistS
ITS2
15
18S
15
Fig. S4
NJ tree
ProfDistS
MP tree
PAUP
ML tree
R script
Fig. S5

◂**Fig. 1** Flowchart of all methods applied in this work. Sequences which could not be properly annotated or aligned were discarded, as well as *Prototheca* strains classified as "sp.". For alignment editing, Align (Hepperle et al. 2004) was used (not shown). After reconstruction of Neighbor-Joining (NJ) overview trees using ProfDistS (Wolf et al. 2008) subsets were manually chosen for Maximum-Likelihood (ML), Maximum-Parsimony (MP) and Neighbor-Joining (NJ) analysis. Further figures available with this manuscript are indicated

et al. 2007). Although protothecosis is considered a rare disease in humans with 105 of 211 published cases reported between 2000 and 2017 (Todd et al. 2018), it is likely that a large number of cases are unreported, undiagnosed or misdiagnosed (Masuda et al. 2021) since diagnostic methods as well as protocols for the treatment of protothecosis are not well established yet (Todd et al. 2018). Among other vertebrates, cattle are most commonly affected of protothecosis (eg. Bueno et al. 2006; Marques et al. 2008; Jagielski et al. 2019), where the infection typically manifests as mastitis (Shave et al. 2021).

Fifteen *Prototheca* species are currently accepted (Jagielski et al. 2019; Kunthiphun et al. 2019), split into two lineages, with a dominance of human- and cattle-associated species, respectively (Jagielski et al. 2019). In both lineages, altogether six species are known to cause infections in humans and other vertebrates: *P. blaschkeae* U.Roesler, A.Möller, A.Hensel, D.Baumann & U.Truyen, 2006, *P. bovis* Jagielski, 2019, *P. cutis* K.Satoh & Makimura, 2010, *P. miyajii* Masuda, Hirose, Ishikawa, Ikawa & Nishimura, 2016, *P. paracutis* Khunthiphun, Endoh, Takashima, Ohkuma, Tanasupawat & Savarajara, 2019 and *P. wickerhamii* (Masuda et al. 2021). Phylogenetic relationships among *Prototheca* and their affiliated species (Tables S1, S2) have been investigated using ribosomal DNA (rDNA) (Suzuki et al. 2018; Jagielski et al. 2018; Masuda et al. 2016; Kunthiphun et al. 2019) and/or partial mitochondrial cytochrome b (*cytb*) (Jagielski et al. 2019) sequence data. In this study, for the first time, using distance-, parsimony- and maximum likelihood-algorithms, we reconstruct the phylogeny of *Prototheca* and allies using rDNA (18S rDNA and ITS2) sequence- and secondary structure data simultaneously, an approach reviewed by Keller et al. (2010), Wolf et al. (2014) and Wolf (2015), increasing robustness and accuracy of phylogenetic tree estimation. The evolution of protothecean ITS2 secondary structures is discussed.

Material and methods

For a material and methods workflow, see Fig. 1. ITS2 and 18S rDNA sequences of *Prototheca* and its affiliated species (Tables S1, S2) were obtained from NCBI Nucleotide database (retrieved on 2021–04-26) (Benson et al. 2009). ITS2 sequences were annotated using the "annotate" option implemented in the ITS2 database which uses Hidden Markov Models to annotate eukaryote ITS2 (Eddy, 1998; Keller et al. 2009; Schultz et al. 2006; Ankenbrand et al. 2015).

In ClustalX (Larkin et al. 2007), ITS2 as well as 18S rDNA sequences were aligned. Introns were removed from the 18S rDNA alignment with the help of the sequence editor Align (Hepperle et al. 2004).

Based on minimum free energy and constrained folding by using lower case letters, secondary structures of selected ITS2 (Tables S1, S2) sequences were predicted with RNAstructure (Reuter et al. 2010) which were then used as templates for homology modeling (Wolf et al. 2005; Selig et al. 2008) of the remaining secondary structures. Homology modeling was performed with the "model" option as implemented in the ITS2 database. Secondary structures of 18S rDNA sequences were also predicted via homology modeling using the ITS2 database. The template structure (*Jaagichlorella luteoviridis* (Chodat) Darienko & Pröschold, 2019) was obtained from the Comparative RNA Web Site (Cannone et al. 2002; Figure S1).

ITS2 and 18S rDNA sequence-structure datasets were each aligned using 4SALE (Seibel et al. 2006 and 2008). 4SALE uses a 12-letter-alphabet consisting of the four nucleotides and their structural states (unpaired, paired left, paired right) to encode sequence and structure information simultaneously. 4SALE was also used to visualize a consensus structure for *Prototheca* ITS2 sequences.

Sequence-structure alignments were exported from 4SALE for further analysis. Specifically, a sequence-structure Neighbor-Joining (NJ) (Saitou and Nei, 1987) tree was calculated based on both ITS2 and 18S rDNA sequence-structure alignments using ProfDistS (Friedrich et al. 2005; Wolf et al. 2008). For ITS2 sequence-structure data, Q_ITS2, a sequence-structure specific General Time Reversible correction model (cf. Lanave et al. 1984) as implemented in ProfDistS, was used, while for 18S rDNA sequence-structure data a sequence-structure specific JC model (Jukes and Cantor, 1969) was used as distance estimation method.

From each dataset, a subset with less taxa was manually chosen (Tables S1, S2). For each subset, a sequence-structure NJ tree was calculated using ProfDistS. Maximum-Parsimony (MP) and Maximum-Likelihood (ML) trees based on sequence-structure-data were calculated with PAUP (Swofford, 2002) (using one-letter encoded sequence-structure data) and R (R Core Team, 2018), respectively. The R-script is available at http://4sale.bioapps.biozentrum.uni-wuerzburg.de/mlseqstr.html. Additionally, ITS2 and 18S rDNA sequences were aligned in ClustalX and sequence-only NJ trees were calculated in ProfDistS. For all methods, due to the complexity of the sequence-structure approach, a bootstrap support (Felsenstein, 1985) was estimated based on 100 pseudo-replicates.

A third dataset was created consisting of combined ITS2 and 18S rDNA sequence-structure alignments. For this dataset, sequence-structure NJ, MP and ML trees were calculated with the programs and methods described above (for comparison reasons, additionally for this dataset, each marker was again handled separately, cf. Figures S4 and S5). All trees were rooted with *Chlorella vulgaris* Beyerinck, 1890 and *Parachlorella kessleri* L.Krieniz, E.H.Hegewald, D. Hepperle, V.A.R.Huss, T.Rohr & M.Wolf, 2004. All alignments are available on request.

Results and discussion

Taxon sampling

From NCBI, 192 ITS2 sequences of *Prototheca* and affiliated species could be obtained, as well as 165 18S rDNA sequences (Tables S1, S2). For ITS2 sequences, re-annotation was performed using the "annotate" tool in the ITS2 database with "Viridiplantae" as the model, inclusion of the proximal stem (last 25 nucleotides of 5.8S and first 25 nucleotides of 28S rDNA) and an E-Value of < 0.01 or < 0.1.

For ITS2 and 18S rDNA, sequence alignments were created using ClustalX. Sequences, which could not be annotated or aligned, were discarded. The ITS2 sequence of *Prototheca wickerhamii* was significantly longer than all other ITS2 sequences and could therefore not be properly aligned. The final alignments consist of 118 ITS2 sequences and 73 18S rDNA sequences (cf. Figure 1).

For ITS2 sequences, six secondary structure templates were created using RNAstructure (*P. blaschkeae*, *P. cutis*, *P. stagnorum* W.B.Cooke, 1968, *P. ulmea* R.S.Pore, 1986, *P. xanthoriae* Jagielski, 2019, *P. zopfii*). With these templates, structures of all other sequences could be predicted using the "model" tool of the ITS2 database with at least 70 percent transfer of the structure for most and 60 percent of the structure for three sequences (*P. tumulicola* Nagatsuka, Kiyuna, Kigawa & J.Sugiyama, 2016). Structures of *P. wickerhamii* could not be predicted with the templates described and showed a significantly longer and bifurcated fourth helix when modeled with RNAstructure. This taxon is therefore missing in further analyses. Phylogenetic trees were calculated on sequence-structure alignments generated in 4SALE, consisting of 112 taxa and a subset with 30 taxa.

For 18S rDNA sequences, a structure template was obtained from CRW (*Jaagichlorella luteoviridis,* X73998, Figure S1). All 73 18S rDNA sequences could be predicted with at least 70 percent transfer of the structures, for all structures except *Helicosporidium sp.* (67.82%). *Prototheca* sequences classified as "sp." were discarded. For 18S rDNA data, phylogenetic trees were calculated with 71 taxa and a subset of 26 taxa.

From ITS2 and 18S rDNA subsets, a combined sequence-structure alignment of 15 strains / sequences was created.

Phylogeny of *Prototheca* based on ITS2 sequence-structure data

A Neighbor-Joining tree was calculated based on 112 ITS2 sequence-structure pairs (Fig. 2). From the clades shown in this overview tree, 30 taxa were manually selected for NJ, MP and ML analysis (Fig. 3). Towards the root of this tree, a highly supported supergroup consisting of *P. miyajii*, *P. cutis* and *P. paracutis* finds itself with *Jaagichlorella luteoviridis* and *Auxenochlorella protothecoides*, showing the polyphyly of the *Prototheca* genus. Strains of *P. xanthoriae* form a sister clade to all other *Prototheca* strains in the ML tree, although its position differs in the trees based on NJ and MP algorithms. *P. moriformis* W.Krüger, 1894 is very highly supported to be a sister group to the remaining taxa, which then are further divided into a *P. tumulicola* /*P. stagnorum* clade and a second clade, a supergroup consisting of *P. zopfii* /*P. bovis*, *P. ciferrii* Negroni & Blaisten, 1941, *P. pringsheimii* Jagielski, 2019, *P. cerasi* Jagielski, 2019, *P. cookei* Jagielski, 2019, and *P. blaschkeae*. In this supergroup, *P. ciferrii* appears to be polyphyletic with *P. pringsheimii* sequences branching within the *P. ciferrii* clade. *P. ciferrii* /*P. pringsheimii* and their sister group *P. cerasi* appear to be a sister group to the *P. zopfii* /*P. bovis* clade. All of these strains form a sister group to *P. cookei*. *P. blaschkeae* appears to sister with just the *P. zopfii* /*P. bovis* clade in the overview NJ tree, but in trees calculated on the subset data the sister group also includes *P. ciferrii*, *P. pringsheimii*, *P. cerasi*, *P. cookei* and *P. moriformis* (only in the MP tree).

In general, the topology of the trees calculated on ITS2 sequence-structure-data show similar topology to the trees calculated by Masuda et al. (2016) and Hirose et al. (2018), with the additional taxa proposed by Jagielski et al. (2019) and Kunthiphun et al. (2019). *Auxenochlorella protothecoides* and *Jaagichlorella luteoviridis* are sister groups in our tree based on ITS2 sequence-structure data with a bootstrap support of 100 and both clade with *P. cutis* / *P. paracutis* / *P. miyajii* with a bootstrap support of 81. *Auxenochlorella protothecoides* branches with *Prototheca wickerhamii* (with a bootstrap support of 58) in the work of Hirose et al. (2018), and is sister group to all *Prototheca* sequences except *P. wickerhamii* in the tree proposed by Masuda et al. (2016) with a bootstrap support of 87. *P. ulmea* is poorly supported being a sister group to *P. zopfii*, *P. moriformis* and *P. blaschkeae* sequences in the same work whereas we show *P. tumulicola* / *P. stagnorum* as sister group to these species with a bootstrap support of 76.

The phylogenetic position of *P. xanthoriae* remains unresolved here as its position differs in all constructed

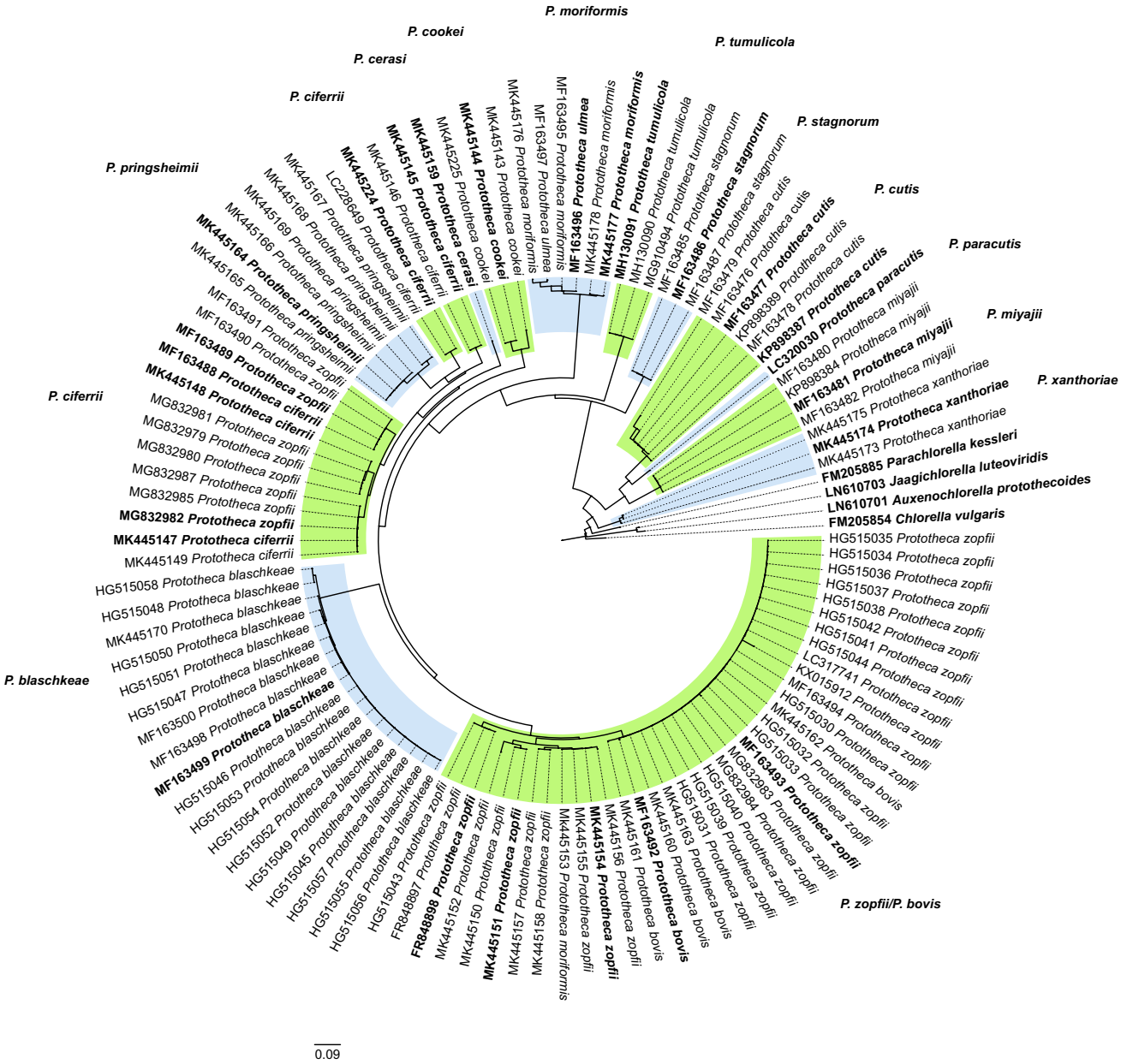


Fig. 2 ITS2 sequence-structure Neighbor-Joining tree obtained from ProfDistS (Wolf et al. 2008). An alignment of 112 sequence-structure-pairs (x.fasta format) of *Prototheca* and affiliated species was created using 4SALE (Seibel et al. 2006 and 2008) and encoded by a 12-letter alphabet (Wolf et al. 2014) for reconstruction of this tree. GenBank accession numbers accompany each taxon name. Clades are alternately marked green and blue and are additionally named alongside the tree in accordance with the clade names proposed in the phylogram by Jagielski et al. (2019). Taxa which were manually chosen for the subset are marked bold. The tree is rooted with *Chlorella vulgaris* FM205854

trees always with low bootstrap support. Several other relationships (e.g. the close relationship of *P. miyajii* + *P. cutis* /*P. paracutis*, *P. tumulicola* + *P. stagnorum* or *P. moriformis* + the supergroup consisting of *P. zopfii* /*P. bovis*, *P. ciferrii*, *P. pringsheimii*, *P. cerasi*, *P. cookei*, *P. blaschkeae*, *P. tumulicola,* and *P. stagnorum*) are very highly supported by bootstrap values > 90 in all (NJ, MP, ML) calculated trees. A single *P. moriformis* sequence (MK445153) was positioned within the *P. zopfii* /*P. bovis* clade in the ITS2 sequence-structure overview tree (Fig. 2). This strain (SAG 263–2) appears in the *P. moriformis* clade (cluster IX) in the phylogram based on the partial *cytb* sequences by Jagielski et al. (2019).

Comparing the sequence-structure tree to a tree based on sequence data only (Figure S2), it is apparent that the sequence-only tree is similar, sometimes lower supported

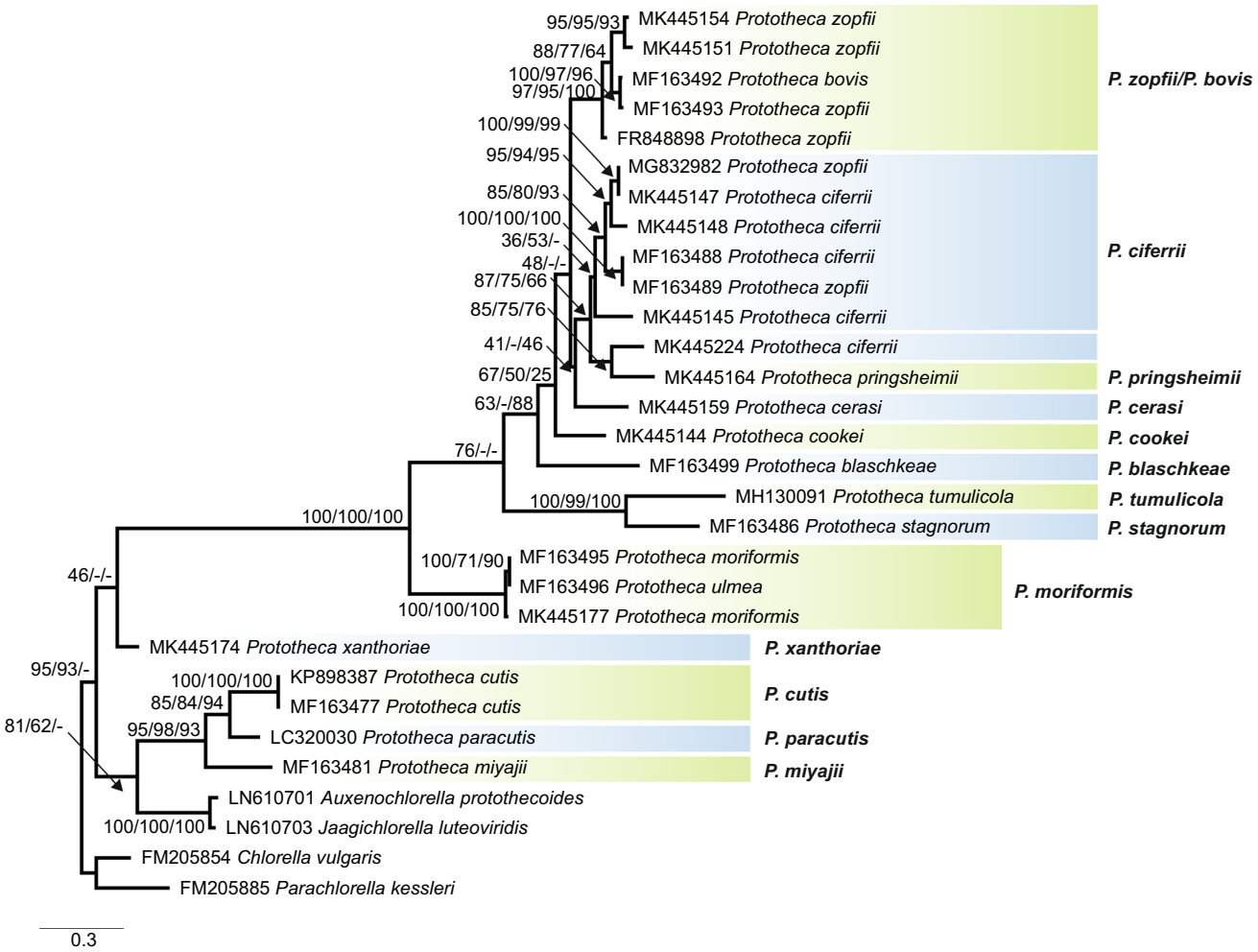


Fig. 3 ITS2 sequence-structure Maximum-Likelihood tree calculated with R (R Core Team, 2018) including a representative subset of 30 sequence-structure pairs from *Prototheca* and its affiliated species which were manually selected from Fig. 2. Bootstrap values from 100 pseudo-replicates mapped at the internodes are from Maximum-Likelihood (ML), Maximum-Parsimony (MP, obtained from PAUP (Swofford, 2002)) and Neighbor-Joining (NJ, obtained from ProfDistS (Wolf et al. 2008)) analyses. For NJ tree reconstruction the global multiple sequence-structure alignment (.xfasta format) as derived by 4SALE (Seibel et al. 2006 and 2008) was automatically encoded by a 12-letter alphabet (Wolf et al. 2014). For ML and MP tree reconstruction the "one letter encoded" fasta format (12-letter alphabet) as derived by 4SALE (Seibel et al. 2006 and 2008) was used. GenBank accession numbers accompany each taxon name. Clades are alternately marked green and blue and are additionally named alongside the tree in accordance with the clade names proposed in the phylogram by Jagielski et al. (2019). The tree is rooted with *Chlorella vulgaris* FM205854 and *Parachlorella kessleri* FM205885

than the tree based on the sequence-structure alignment and shows several differences in the topology, e.g. the positions of the *P. tumulicola* / *P. stagnorum* clade or the *P. moriformis* clade. *Auxenochlorella protothecoides* and *Jaagichlorella luteoviridis* branch inside the *P. cutis* / *P. paracutis* / *P. miyajii* clade in the sequence-only tree. This latter clade without *A. protothecoides* and *J. luteoviridis* is highly supported in the sequence-structure tree with a bootstrap value of 93.

Phylogeny of *Prototheca* based on 18S rDNA sequence-structure data

Using the Neighbor-Joining algorithm, an overview tree based on 71 18S rDNA sequence-structure pairs was created (Fig. 4). Here, as in several other trees based on rDNA data (e.g. Masuda et al. 2016; Hirose et al. 2018; Shave et al. 2021), *P. wickerhamii* appears to be polyphyletic with two strains (X56099, X74003) branching outside of the *P. wickerhamii* clade. Jagielski et al. (2019), reclassified these taxa as a new species, *Prototheca xanthoriae*. From the NJ overview tree, a subset was created by manual selection of

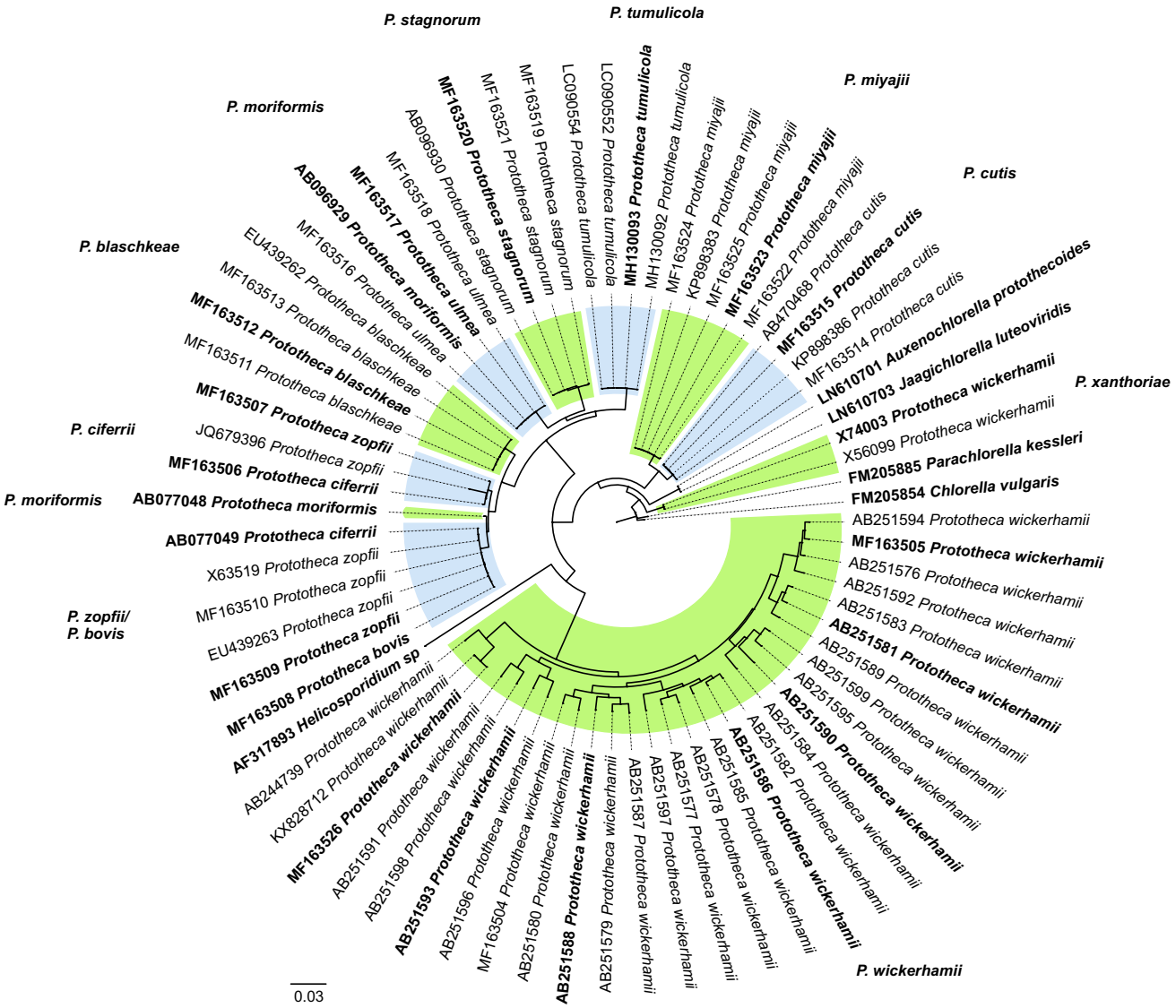


Fig. 4 18S rDNA sequence-structure Neighbor-Joining tree obtained from ProfDistS (Wolf et al. 2008). An alignment of 71 sequence-structure-pairs (x.fasta format) of *Prototheca* and its affiliated species was created using 4SALE (Seibel et al. 2006 and 2008) and encoded by a 12-letter alphabet (Wolf et al. 2014) for reconstruction of this tree. GenBank accession numbers accompany each taxon name. Clades are alternately marked green and blue and are additionally named alongside the tree in accordance with the clade names proposed in the phylogram by Jagielski et al. 2019. Taxa which were manually chosen for the subset are marked bold. The tree is rooted with *Chlorella vulgaris* FM205854 and *Parachlorella kessleri* FM205885

26 sequence-structure pairs. Trees calculated on this subset data (Fig. 5) show *P. xanthoriae* as the sister group to all strains except the outgroup. *Auxenochlorella protothecoides* and *Jaagichlorella luteoviridis* are sister group to all remaining taxa, which are then further divided into a *P. miyajii* /*P. cutis* clade and a second clade, in which *Helicosporidium* sisters with *P. wickerhamii* and another supergroup of several *Prototheca* species. This supergroup forms two clades, the first being *P. ulmea* /*P. moriformis* and their sister group *P. tumulicola* /*P. stagnorum*, the second divided into a *P. blaschkeae* clade and a second clade consisting of *P. ciferrii*, *P. moriformis* and *P. zopfii* /*P. bovis*. Bootstrap support of this tree is generally high as all but one external nodes are supported by a bootstrap value > 65. The trees calculated on 18S rDNA sequence-structure data show similar topology to the trees based on LSU rDNA data proposed in literature (e.g. Masuda et al. 2016; Hirose et al. 2018), but differ from the phylogram based on partial *cytb* sequences by Jagielski et al. (2019) in several aspects. *P. stagnorum*, *P. tumulicola* and *P. moriformis* appear towards the root of the tree in the *cytb* sequence based phylogram, while our tree shows all three species distant to the root and forming the sister

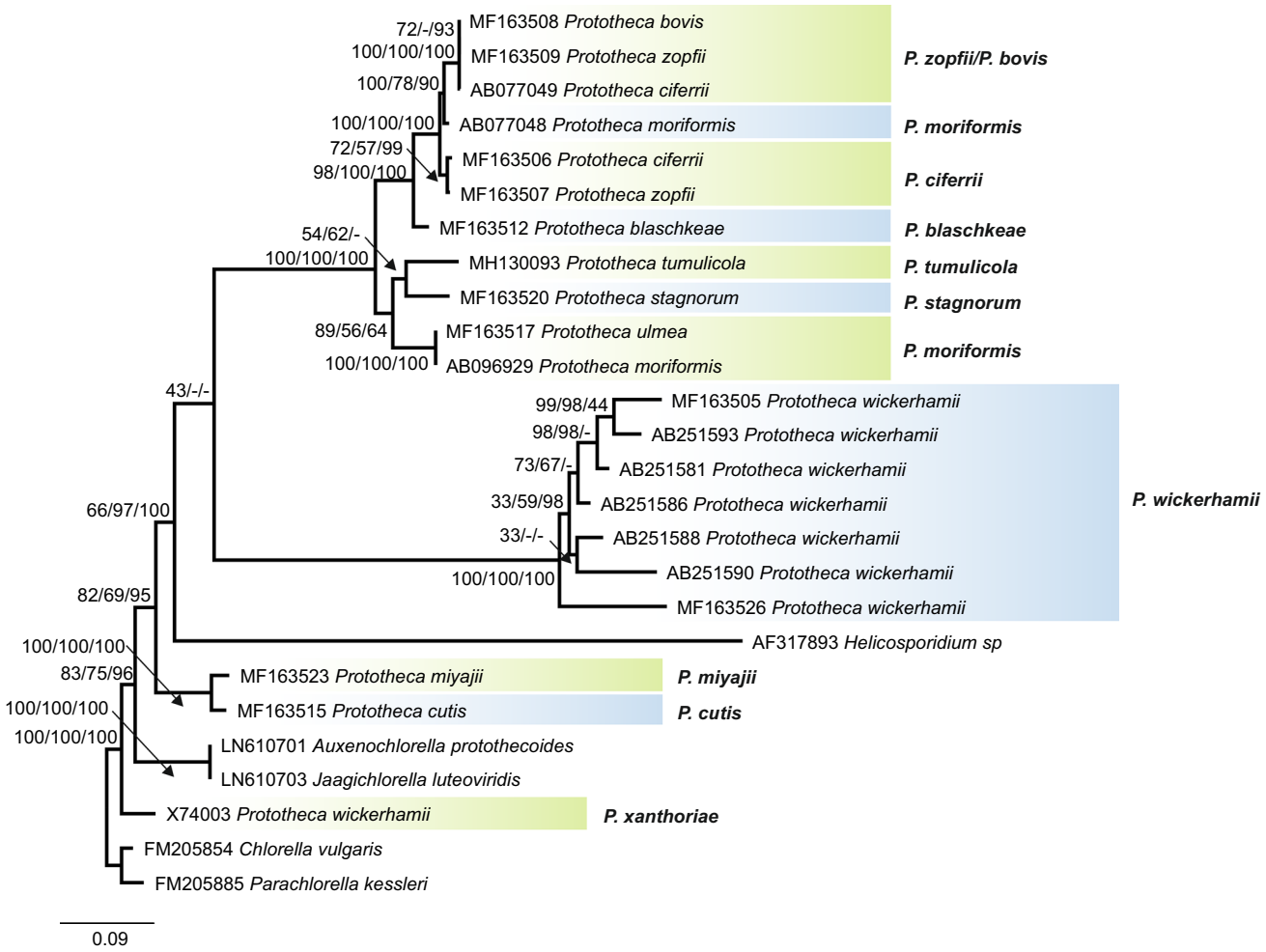


Fig. 5 18S rDNA sequence-structure Maximum-Likelihood tree calculated with R (R Core Team, 2018) including a representative subset of 26 sequence-structure pairs from *Prototheca* and its affiliated species which were manually selected from Fig. 4. Bootstrap values from 100 pseudo-replicates mapped at the internodes are from Maximum-Likelihood (ML), Maximum-Parsimony (MP, obtained from PAUP (Swofford, 2002)) and Neighbor-Joining (NJ, obtained from ProfDistS (Wolf et al. 2008)) analyses. For NJ tree reconstruction the global multiple sequence-structure alignment (.xfasta format) as derived by 4SALE (Seibel et al. 2006 and 2008) was automatically encoded by a 12-letter alphabet (Wolf et al. 2014). For ML and MP tree reconstruction the "one letter encoded" fasta format (12-letter alphabet) as derived by 4SALE (Seibel et al. 2006 and 2008) was used. GenBank accession numbers accompany each taxon name. Clades are alternately marked green and blue and are additionally named alongside the tree in accordance with the clade names proposed in the phylogram by Jagielski et al. (2019). The tree is rooted with *Chlorella vulgaris* FM205854 and *Parachlorella kessleri* FM205885

group to *P. zopfii* / *P. bovis*, *P. moriformis*, *P. ciferrii* and *P. blaschkeae*. In the *cytb* sequenced phylogram, a multifurcation occurs including *P. wickerhamii*, the sister groups *P. miyajii* and *P. cutis* as well as a clade consisting of *P. xanthoriae*, *Helicosporidium sp*. and *Auxenochlorella protothecoides*. *P. wickerhamii* is shown to be the sister group of *P. zopfii* / *P. bovis*, *P. ciferrii*, *P. blaschkeae*, *P. tumulicola*, *P. stagnorum* and *P. moriformis* in our tree, although with low bootstrap support. *Helicosporidium sp*. is sister group to all of these species (with moderate bootstrap support) and *P. miyajii* / *P. cutis* sister with these species including *Helicosporidium sp*. with bootstrap values > 70 for all methods applied (NJ, MP, ML).

Comparing the sequence-structure tree to a tree based on sequence data only (Figure S3), it is apparent that the bootstrap support is mostly higher although sometimes similar in the sequence-structure tree. The topology differs slightly, e.g. in the *P. zopfii* /*P. bovis* + *P. ciferrii* + *P. moriformis* clade, *P. tumulicola* + *P. stagnorum* clade or within the *P. wickerhamii* clade.

Phylogeny of *Prototheca* based on combined ITS2 and 18S rDNA sequence-structure data

A combined ITS2 and 18S rDNA sequence-structure alignment was created from strains which appeared in both the ITS2 and 18S rDNA subset. NJ, MP and ML trees were calculated on this 15 taxa sequence-structure alignment (Fig. 6). This tree is highly supported by bootstrap values ≥ 70 at all nodes throughout the whole tree with most of them being > 95. *P. cutis* and *P. miyajii* form a clade outside all other *Prototheca* clades, which are then further divided into a *P. moriformis* clade and the sister group consisting of *P. stagnorum*, *P. tumulicola*, *P. blaschkeae*, *P. ciferrii* and *P. zopfii* /*P. bovis*. In this supergroup, *P. stagnorum* and *P. tumulicola* find themselves together against the remaining taxa which then have *P. blaschkeae* as sister group to *P. ciferrii* and *P. zopfii* /*P. bovis* clades.

Comparing the ITS2 and 18S rDNA subset trees to the tree based on the combined alignment, the trees show similar topology despite several species missing in the combined alignment. In all three trees, *P. zopfii* / *P. bovis* (and one *P. moriformis* strain in the 18S rDNA tree) is the sister group to *P. ciferrii*, forming a supergroup which then is sister group to *P. blaschkeae*.

While *P. moriformis* and *P. tumulicola* / *P. stagnorum* are sister groups in the 18S rDNA tree, *P. moriformis* is sister group to several more species in the ITS2 and the combined tree. *P. cutis* / *P. miyajii* form a sister group to all other *Prototheca* strains in the combined and the 18S rDNA tree (except *P. xanthoriae*, which sisters with all species except the outgroup in the 18S rDNA tree) but are more closely related to *Auxenochlorella protothecoides* and *Jaagichlorella luteoviridis* in the tree based on ITS2 sequence-structure data. Accordingly, these nodes are the nodes showing a relatively low bootstrap support in the very highly supported tree based on the combined 18S rDNA and ITS2 alignment.

Calculating trees on ITS2 and 18S rDNA sequence-structure data of the 15 chosen taxa for the combined alignment

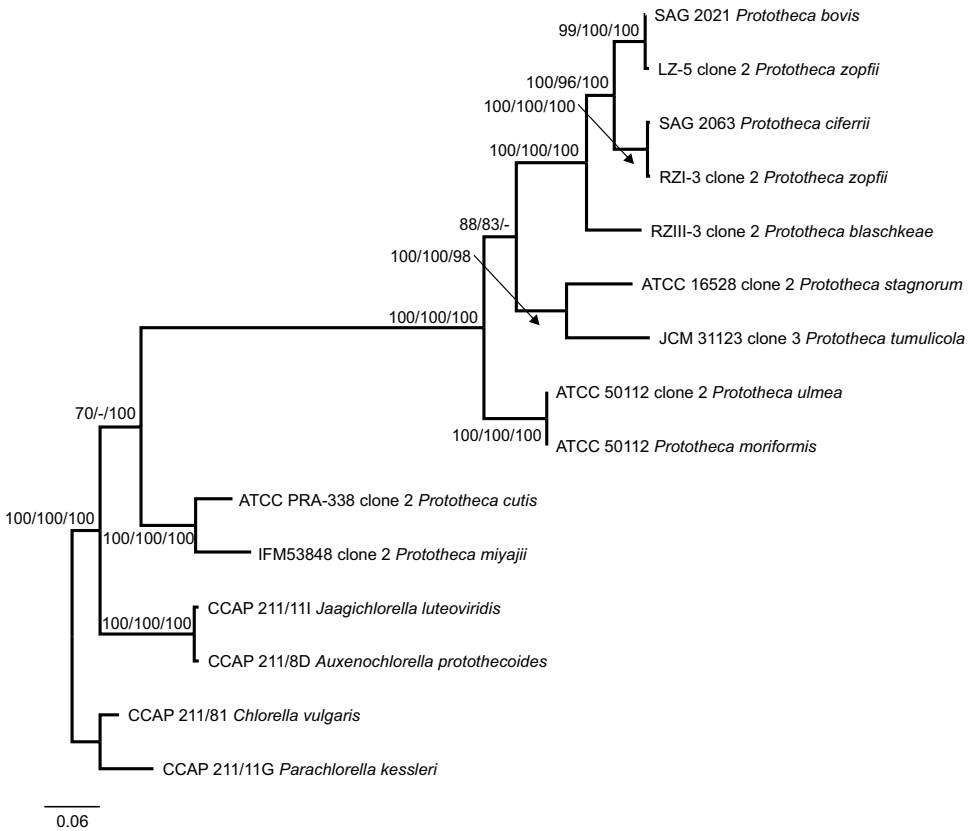


Fig. 6 Combined 18S and ITS2 sequence-structure Maximum-Likelihood tree calculated with R (R Core Team, 2018) including 15 sequence-structure pairs from *Prototheca* and its affiliated species. Bootstrap values from 100 pseudo-replicates mapped at the internodes are from Maximum-Likelihood (ML), Maximum-Parsimony (MP, obtained from PAUP (Swofford, 2002)) and Neighbor-Joining (NJ, obtained from ProfDistS (Wolf et al. 2008)) analyses. For NJ tree reconstruction the global multiple sequence-structure alignment (.xfasta format) as derived by 4SALE (Seibel et al. 2006 and 2008) was automatically encoded by a 12-letter alphabet (Wolf et al. 2014). For ML and MP tree reconstruction the "one letter encoded" fasta format (12-letter alphabet) as derived by 4SALE (Seibel et al. 2006 and 2008) was used. Strain numbers accompany each taxon name. The tree is rooted with *Chlorella vulgaris* CCAP 211/81 and *Parachlorella kessleri* CCAP 211/11G

separately, it is apparent that the combined ML tree shows more similarity to the 18S rDNA tree (Figure S4) than the ITS2 tree (Figure S5). Both separate trees show the supergroup consisting of *P. moriformis*, *P. tumulicola*, *P. stagnorum*, *P. blaschkeae*, *P. ciferrii* and *P. zopfii* / *P. bovis* with a bootstrap value of 100. The relationships between the *Prototheca* strains in this supergroup varies; however the combined tree and the 18S rDNA tree show *P. blaschkeae* being a sister group to *P. ciferri* and *P. zopfii* / *P. bovis* while in the ITS2 tree, *P. blaschkeae* is related to *P. zopfii* / *P. bovis*, but with low bootstrap support. The ITS2 tree shows *P. miyajii* / *P. cutis* being sister group to *Auxenochlorella protothecoides* and *Jaagichlorella luteoviridis*, whereas this relationship doesn't appear in the 18S rDNA or combined ML tree. The bootstrap support of the 18S rDNA tree is overall high with all but one bootstrap value > 60. In the ITS2 tree, two bootstrap values are lower than 50 with the accompanying nodes being the ones where the ITS2 tree doesn't show the same topology as either the 18S rDNA or combined tree.

Finally, if we compare ITS2 and 18S rDNA trees, we must not forget that we cannot include *P. wickerhamii* in the comparison. In order to deduce the phylogeny of the entire genus *Prototheca*, i.e., including *P. wickerhamii*, one always needs at least one additional marker gene beside ITS2.

Evolution of protothecean ITS2 secondary structures

ITS2 secondary structures of six *Prototheca* sequences were constructed using RNAstructure (Fig. 7). In general, these structures folded into the common core structure known for eukaryotes with four helices (Schultz et al. 2005). Protothecean ITS sequences are known to vary in length (Marques et al. 2015). *Prototheca* sequences in this work were between 269 (all three *P. tumulicola* strains) and 543 / 544 bp (*P. moriformis* MF163495 /*P. ulmea* MF163497) long. ITS2 sequences of *P. wickerhamii* were significantly longer (1171–1358 bp). ITS2 structures from *P. blaschkeae* and *P. cutis* showed an exceptionally large fourth helix, while helix IV of *P. stagnorum* and *P. zopfii* was rather short. The third helix of *P. ulmea* appears to be bifurcated. Figure 8 visualizes the sequence-structure alignment of all *Prototheca* strains in the subset by a 51% consensus structure. A few bindings in helix II, between helix II and III and at the end of helix III are shown to be 80% conserved where known ITS2 structure motifs (the U-U mismatch in helix II, the triple A between helix II and III and the UGGU motif in helix III) are generally located.

Despite the differences in length in the *Prototheca* ITS2 sequences, homology modeling of the secondary structures was possible with just three templates (*P. zopfii*, *P. cutis*, *P. stagnorum*) at 50% consensus level for all *Prototheca* sequences except *P. blaschkeae* and *P. wickerhamii*. The *P. zopfii* template could be used to model other *P. zopfii* structures and those of *P. bovis*, *P. cerasi*, *P. ciferrii*, *P. cookei*, one *P. moriformis* strain and *P. pringsheimii*. These species also form a supergroup in the ML sequence-structure tree (Fig. 3). With the *P. cutis* template, all strains of the *P. cutis* /*P. paracutis* + *P. miyajii* clade could be predicted. The *P. stagnorum* template could be used for prediction of the secondary structures of other *P. stagnorum* and the *P. tumulicola* sequences as well as *P. ulmea*, *P. xanthoriae* and *P. moriformis* sequences with a lower consensus. Therefore, three additional templates were created (*P. blaschkeae*, *P. ulmea* and *P. xanthoriae*).

Given their distant relationship in the pylogenetic trees based on ITS2 sequence-structure data, elongation of helix IV of the ITS2 in *P. blaschkeae* and *P. cutis* seems to have occurred independently in the course of evolution.

ITS2 is one of the most effective phylogenetic markers. The high variability allows to study closely related organisms, the conserved structure reveals larger relationships. In most cases, the secondary structure helps to better align variable sequences. Sometimes, however, the length variations and differences even within a genus are already so large that alignments (whether based only on sequence or on sequence-structure information) should be viewed with caution. *Prototheca* is such an example. Homology is difficult to discern and individual sequences are even impossible to align at all. On the other hand, if you take out only a few sequences (e.g. those with an extremely elongated fourth helix), the alignment quickly becomes much more compact. With this study, we reconstructed phylogenetic trees on extremely diverse *Prototheca* sequences—whose sequence-structure information was encoded into a new alphabet; and indeed the results show robust trees similar to those based on other markers (e.g. 18S, LSU or *cytb*). We encourage the commuity to draw on additional markers and, by comparison and/or concatination, to better and better understand the phylogeny of *Prototheca* and related taxa.

To understand ITS2 length differences further research is needed. Compared to other genera, in terms of extreme sequence differences, it seems possible to discover additional species in the *Prototheca* species complex. Such species will then put the sequence differences into perspective and/or significantly advance our understanding of length variation (e.g. by expansion, duplication, and/or alternative splicing), or more generally, our understanding of RNA sequence-structure evolution.

Conclusion

In this work, using sequence-structure information simultaneously, for two phylogenetic markers (ITS2 and 18S rDNA), we reconstructed generally well-supported phylogenetic trees that are in overall agreement with the trees based

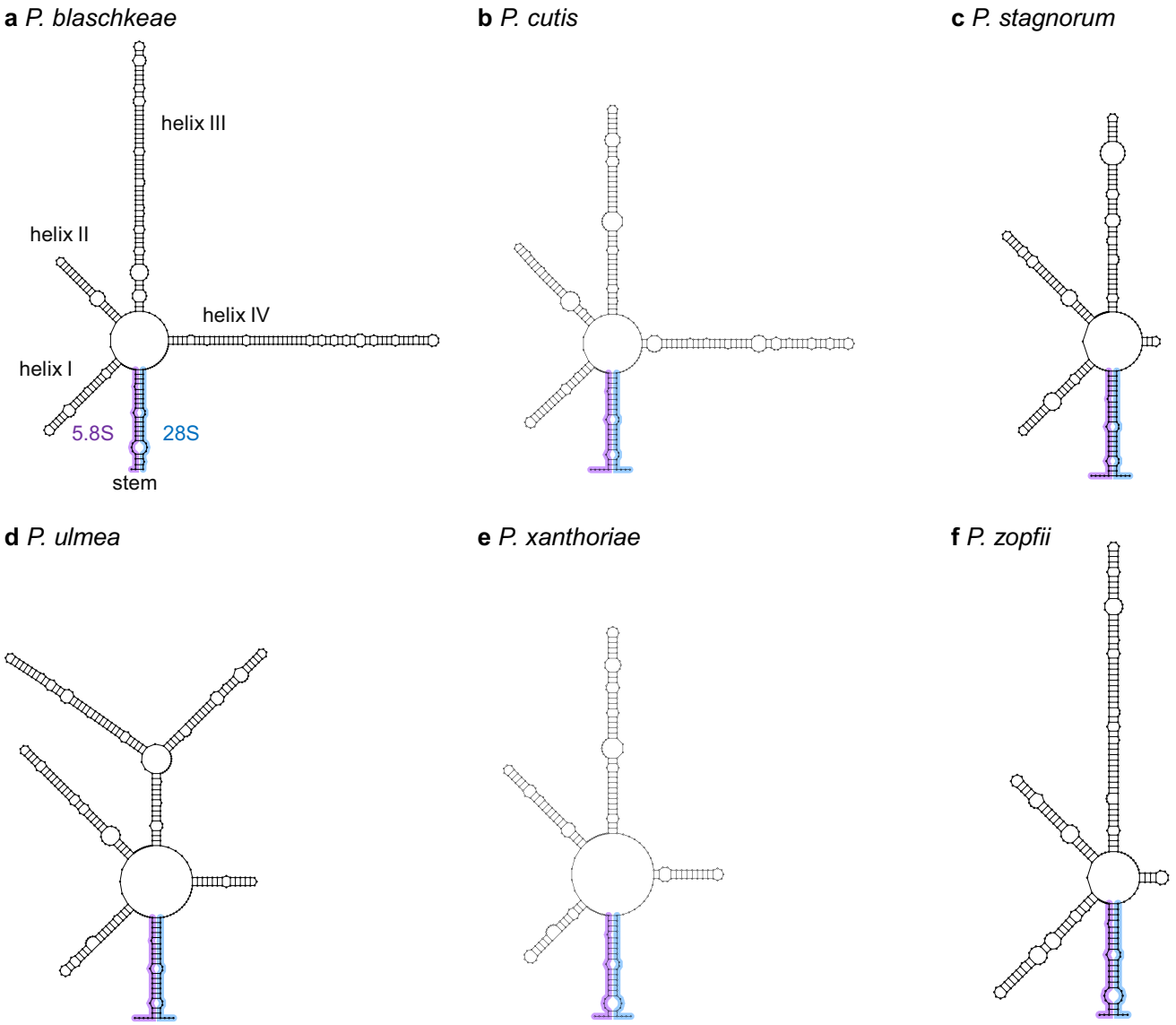


Fig. 7 ITS2 secondary structure templates used for homology modeling of *Prototheca* sequences in the ITS2 database (Schultz et al. 2006; Ankenbrand et al. 2015). Templates were created in RNAstructure (Reuter et al. 2010) based on minimum free energy and constrained folding. The stem, consisting of the last 25 nucleotides of the 5.8S and the first 25 nucleotides of the 28S rDNA is highlighted in purple (5.8S) and blue (28S) using Varna (Darty et al. 2009)

on rDNA sequences (mainly LSU data) proposed in literature but show several topological differences to trees calculated on *cytb* sequences. *Prototheca wickerhamii*, the main causative for human protothecosis, could not be included in analysis based on ITS2 data since its ITS2 sequences were exceptionally long and could therefore not be aligned with other *Prototheca* sequences. The phylogenetic trees calculated on sequence-structure alignments of our subset data show Maximum-Likelihood support (> 50) for all but three branches in both the ITS2 and the 18S rDNA tree. Bootstrap support values are generally higher than those from sequence-only analyses (in this study or in the available literature using RNA and/or protein data).

The ITS2 of *Prototheca* is known to vary in length. Our study shows that out of the *Prototheca* ITS2 structures we reconstructed, *P. blaschkeae* and *P. cutis* displayed an elongated fourth helix. Helix III of *P. moriformis* (formerly *P. ulmea*) appears to be bifurcated. Despite the differences in length, a 51% consensus structure showing all but the fourth helix could be visualized with some nucleotide bonds being 80% conserved throughout all examined *Prototheca* structures.

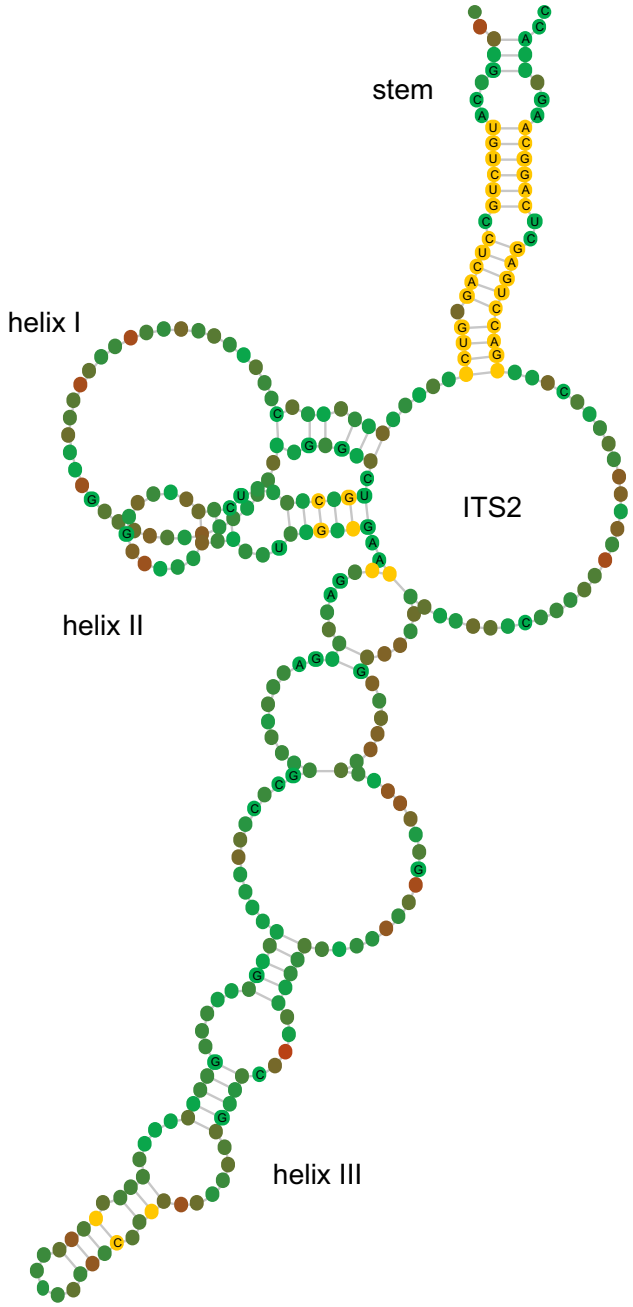


Fig. 8 Visualization of the subset *Prototheca* sequence-structure alignment without gaps (outgroups, *Jaagichlorella* and *Auxenochlorella* species were excluded) by a 51% consensus structure created in 4SALE (Seibel et al. 2006 and 2008). Nucleotide bonds that are at least 80% conserved are marked in yellow. Conservation of the sequence is indicated by red (low conservation) to green (high conservation) color. Nucleotides which are 100% conserved in all sequences are written as A, U, G or C

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11756-021-00971-y.

Acknowledgements We would like to thank Noelle Pina for proof reading English.

Authors' contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Tanja Plieger and Matthias Wolf. The first draft of the manuscript was written by Tanja Plieger and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of interest/Competing interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

CRediT taxonomy Conceptualization: Matthias Wolf; Methodology: Matthias Wolf; Formal analysis and investigation: Tanja Plieger, Matthias Wolf; Writing—original draft preparation: Tanja Plieger; Writing—review and editing: Matthias Wolf; Supervision: Matthias Wolf.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F (2015) ITS2 Database V: Twice as Much. Mol Biol Evol 32(11):3030–3032. https://doi.org/10.1093/molbev/msv174

Bakuła Z, Gromadka R, Gawor J, Siedlecki P, Pomorski JJ, Maciszewski K, Gromadka A, Karnkowska A, Jagielski T (2020) Sequencing and analysis of the complete organellar genomes of *Prototheca wickerhamii*. Front Plant Sci 11:1296. https://doi.org/10.3389/fpls.2020.01296

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res 41:D36–D42. https://doi.org/10.1093/nar/gks1195

Bueno VF, de Mesquita AJ, Neves RB, de Souza MA, Ribeiro AR, Nicolau ES, de Oliveira AN (2006) Epidemiological and clinical aspects of the first outbreak of bovine mastitis caused by Prototheca zopfii in Goiás State. Brazil Mycopathologia 161(3):141–145. https://doi.org/10.1007/s11046-005-0145-8

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, Gutell RR (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformat

3:2. Erratum. In: BMC Bioinformat 3(1):15. https://doi.org/10.1186/1471-2105-3-2
Darty K, Denise A, Ponty Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25(15):1974–1975. https://doi.org/10.1093/bioinformatics/btp250
Davies RR, Spencer H, Wakelin PO (1964) A case of human protothecosis. Trans R Soc Trop Med Hyg 58:448–451. https://doi.org/10.1016/0035-9203(64)90094-x
Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763. https://doi.org/10.1093/bioinformatics/14.9.755
Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39(4):783–791. https://doi.org/10.1111/j.1558-5646.1985.tb00420.x
Friedrich J, Dandekar T, Wolf M, Müller T (2005) ProfDist: a tool for the construction of large phylogenetic trees based on profile distances. Bioinformatics 21(9):2108–2109. https://doi.org/10.1093/bioinformatics/bti289
Guiry MD, Guiry GM (2021) AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. https://www.algaebase.org. Accessed 1 Sept 2021
Hepperle D (2004) Align Ver.07/04©. Multisequence alignment-editor and preparation/manipulation of phylogenetic datasets. Win32-Version. http://www.sequentix.de. Accessed 1 Sept 2021
Hirose N, Hua Z, Kato Y, Zhang Q, Li R, Nishimura K, Masuda M (2018) Molecular Characterization of Prototheca strains isolated in China revealed the first cases of protothecosis associated with Prototheca zopfii genotype 1. Med Mycol 56(3):279–287. https://doi.org/10.1093/mmy/myx039
Huerre M, Ravisse P, Solomon H, Ave P, Briquelet N, Maurin S, Wuscher N (1993) Protothécoses humaines et environnement [Human protothecosis and environment]. Bull Soc Pathol Exot 86(5 Pt 2):484–488
Jagielski T, Bakuła Z, Gawor J, Maciszewski K, Kusber WH, Dyląg M, Nowakowska J, Gromadka R, Karnkowska A (2019) The genus Prototheca (Trebouxiophyceae, Chlorophyta) revisited: Implications from molecular taxonomic studies. Algal Research 43:101639
Jagielski T, Dyląg M, Roesler U, Murugaiyan J (2017) Isolation of infectious microalga Prototheca wickerhamii from a carp (Cyprinus carpio) - a first confirmed case report of protothecosis in a fish. J Fish Dis 40(10):1417–1421. https://doi.org/10.1111/jfd.12614
Jagielski T, Gawor J, Bakuła Z, Decewicz P, Maciszewski K, Karnkowska A (2018) cytb as a New Genetic Marker for Differentiation of Prototheca Species. J Clin Microbiol 56(10):e00584-e618. https://doi.org/10.1128/JCM.00584-18
Jagielski T, Lagneau PE (2007) Protothecosis. A Pseudofungal Infection J Mycol Med 17:261–270. https://doi.org/10.1016/j.mycmed.2007.08.003
Jukes TH, Cantor CR (1969) Evolution of Protein Molecules. Munro, H.N., Ed., Mammalian Protein Metabolism, Academic Press, New York, 21–132. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7
Kano R (2020) Emergence of Fungal-Like Organisms: Prototheca. Mycopathologia 185(5):747–754. https://doi.org/10.1007/s11046-019-00365-4
Keller A, Förster F, Müller T, Dandekar T, Schultz J, Wolf M (2010) Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. Biol Direct 5:4. https://doi.org/10.1186/1745-6150-5-4
Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M (2009) 5.8S–28S rRNA interaction and HMM-based ITS2 annotation. Gene 430(1–2):50–57. https://doi.org/10.1016/j.gene.2008.10.012
Kunthiphun S, Endoh R, Takashima M, Ohkuma M, Savarajara A (2019) Prototheca paracutis sp. nov., a novel oleaginous achlorophyllous microalga isolated from a mangrove forest. Mycoscience 60(3):165–169
Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20(1):86–93. https://doi.org/10.1007/BF02101990
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948. https://doi.org/10.1093/bioinformatics/btm404
Lass-Flörl C, Mayr A (2007) Human protothecosis. Clin Microbiol Rev 20(2):230–242. https://doi.org/10.1128/CMR.00032-06
Leimann BC, Monteiro PC, Lazéra M, Candanoza ER, Wanke B (2004) Protothecosis. Med Mycol 42(2):95–106. https://doi.org/10.1080/13695780310001653653
Marques S, Huss VA, Pfisterer K, Grosse C, Thompson G (2015) Internal transcribed spacer sequence-based rapid molecular identification of Prototheca zopfii and Prototheca blaschkeae directly from milk of infected cows. J Dairy Sci 98(5):3001–3009. https://doi.org/10.3168/jds.2014-9271
Marques S, Silva E, Kraft C, Carvalheira J, Videira A, Huss VA, Thompson G (2008) Bovine mastitis associated with Prototheca blaschkeae. J Clin Microbiol 46(6):1941–1945. https://doi.org/10.1128/JCM.00323-08
Masuda M, Hirose N, Ishikawa T, Ikawa Y, Nishimura K (2016) Prototheca miyajii sp. nov., isolated from a patient with systemic protothecosis. Int J Syst Evol Microbiol 66(3):1510–1520. https://doi.org/10.1099/ijsem.0.000911
Masuda M, Jagielski T, Danesi P, Falcaro C, Bertola M, Krockenberger M, Malik R, Kano R (2021) Protothecosis in Dogs and Cats-New Research Directions. Mycopathologia 186(1):143–152. https://doi.org/10.1007/s11046-020-00508-y
Möller A, Truyen U, Roesler U (2007) Prototheca zopfii genotype 2: the causative agent of bovine protothecal mastitis? Vet Microbiol 120(3–4):370–374. https://doi.org/10.1016/j.vetmic.2006.10.039
R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.r-project.org/. Accessed 1 Sept 2021
Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformat 11:129. https://doi.org/10.1186/1471-2105-11-129
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454
Schultz J, Maisel S, Gerlach D, Müller T, Wolf M (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. RNA 11(4):361–364. https://doi.org/10.1261/rna.7204505
Schultz J, Müller T, Achtziger M, Seibel PN, Dandekar T, Wolf M (2006) The internal transcribed spacer 2 database–a web server for (not only) low level phylogenetic analyses. Nucleic Acids Res 34:W704–W707. https://doi.org/10.1093/nar/gkl129
Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M (2006) 4SALE–a tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinformat 7:498. https://doi.org/10.1186/1471-2105-7-498
Seibel PN, Müller T, Dandekar T, Wolf M (2008) Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. BMC Res Notes 1:91. https://doi.org/10.1186/1756-0500-1-91
Selig C, Wolf M, Müller T, Dandekar T, Schultz J (2008) The ITS2 Database II: homology modelling RNA structure for molecular systematics. Nucleic Acids Res 36:D377–D380. https://doi.org/10.1093/nar/gkm827

Shave CD, Millyard L, May RC (2021) Now for something completely different: Prototheca, pathogenic algae. PLoS Pathog 17(4):e1009362. https://doi.org/10.1371/journal.ppat.1009362

Suzuki S, Endoh R, Manabe RI, Ohkuma M, Hirakawa Y (2018) Multiple losses of photosynthesis and convergent reductive genome evolution in the colourless green algae Prototheca. Sci Rep 8(1):940. https://doi.org/10.1038/s41598-017-18378-8

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods) Version 4.0b10. Sinauer Associates Sunderland, Massachusetts

Todd JR, Matsumoto T, Ueno R, Murugaiyan J, Britten A, King JW, Odaka Y, Oberle A, Weise C, Roesler U, Pore RS (2018) Medical phycology 2017. Med Mycol 56(suppl_1):S188–S204. https://doi.org/10.1093/mmy/myx162

Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T (2005) Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. RNA 11(11):1616–1623. https://doi.org/10.1261/rna.2144205

Wolf M, Ruderisch B, Dandekar T, Schultz J, Müller T (2008) ProfDistS: (profile-) distance based phylogeny on sequence–structure alignments. Bioinformatics 24(20):2401–2402. https://doi.org/10.1093/bioinformatics/btn453

Wolf M, Koetschan C, Müller T (2014) ITS2, 18S, 16S or any other RNA - simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. Gene 546(2):145–149. https://doi.org/10.1016/j.gene.2014.05.065

Wolf M (2015) ITS so much more. Trends Genet 31(4):175–176. https://doi.org/10.1016/j.tig.2015.02.005