

Phylogenetic Cut-off Values for Species and Genera Discrimination of Yeast

Roumen Dimitrov*¹, Dilynora Gouliamova²

¹*Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria*

²*Institute of Microbiology, Bulgarian Academy of Sciences, Sofia, Bulgaria*

Abstract

In the article, we give a brief presentation of our new method based on a unified frame *Lodderomyces* and *Kazachstania* clades. Furthermore, a new approach has been applied to discriminate yeast species and genera in the *Starmerella* clade. An indication of this is numerous literature data showing low statistical support for basal branches of a phylogenetic tree of the *Starmerella* clade. Phylogenetic analysis shows that a tree of the *Starmerella* clade falls into 7 clusters with high statistical support of basal branches. Among these clusters, cluster 5 represents a new genus. At the border of cluster 5 and its neighbor cluster 4, the following characteristics change with a jump: nucleotide sequence identity (SI), evolutionary distance (ED), and physiological profiles (PP) between species belonging to clusters 4 and 5.

Keywords: molecular evolution, phylogenetics, physiologic profiles, yeast species and genera boundaries.

Резюме

В настоящата статия представяме кратка презентация на нашият нов метод, базиращ се на единен подход за дискриминация на видове и родове в дрождевите кладове *Lodderomyces* и *Kazachstania*. Освен това, новият подход е приложен за дискриминация на видове и родове в дрождевия клад *Starmerella*. Индикация за това са многобройни литературни данни, показващи ниска статистическа подкрепа на базалните разклонения на филогенетичното дърво на клада *Starmerella*. Анализът показва, че филогенетично дърво на клад *Starmerella* се разпада на 7 клъстера с висока статистическа подкрепа на базалните разклонения. Между тези клъстери, клъстер 5 представлява нов род. На границата на клъстер 5 и неговия съсед клъстер 4 със скок се променят следните характеристики: идентичност на нуклеотидните последователности (ИНП) еволюционното разстояние (ЕР) и физиологичните профили (ФП) между видове, принадлежащи към клъстери 4 и 5.

Introduction

Problem formulation

Current research on the taxonomic hierarchy is based on phylogenetic reconstruction. Phylogenetic trees are mainly inferred using DNA sequences, multiple sequence alignments, and ancestry DNA sequence reconstruction of nested monophyletic clades. Monophyletic clades are represented as groups of species that share common recent ancestors, and common phenotypic characteristics and follow predictive evolutionary processes (Lachance, 2018). Because in taxonomic hierarchy only species unit has direct biological meaning one of the major open questions is how to determine the boundary between higher than species phylogenetic units in the taxonomic hierarchy based on unified criteria. In the present article, we focus on unified criteria for the determination of boundaries

between species and genera (Fig. 1).

In the current classification of yeasts, three main problems remain unresolved (Fig. 2).

First, is the absence of universally applicable criteria for taxonomic discrimination of higher-level phylogenetic units. Second, higher-level phylogenetic units have poorly supported branching orders. Third, a phylogenetic tree based on single or multiple gene markers does not provide a complete picture of evolution such as - gene transfer, duplication, deletion, and functional innovation. The first problem can be resolved based on the observation of the existence of discontinuities in fundamental characteristics such as genomic sequence identity (SI), evolutionary distance (ED), and phenotypic clusters of physiological profiles (PP) (Gouliamo-

* Corresponding author: roumen.dimitrov@gmail

Phylogenetic tree and taxonomic hierarchy

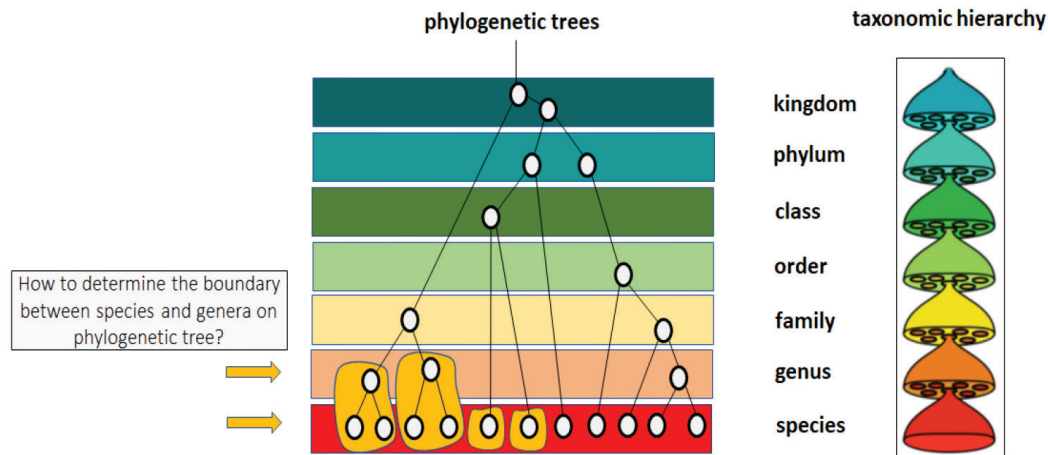


Fig. 1. Currently, taxonomic ranks are determined by phylogenetic trees that are derived primarily using DNA sequences of one or more markers i.e. barcodes. A monophyletic taxon includes a common ancestor of all species in the taxon and all species derived from it. Taxonomic hierarchy is represented as nested levels of taxonomic categories.

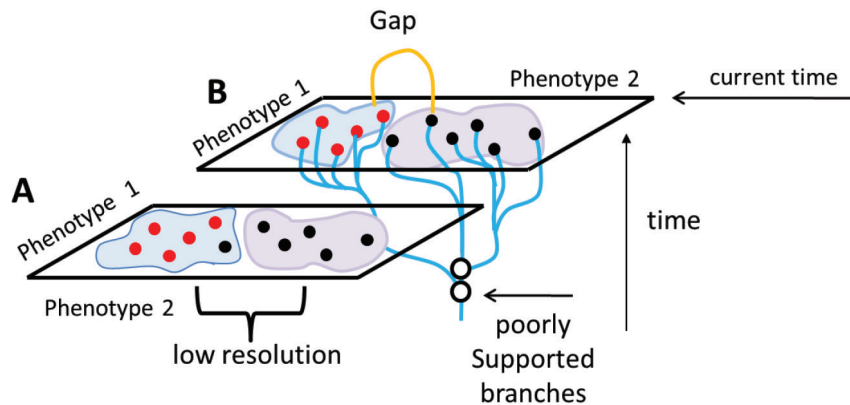


Fig. 2. Critical gaps remain in the current classification of yeasts: use of phenotypic characteristics with low resolution and absence of universally applicable criteria for taxonomic discrimination of higher-level phylogenetic units (A); phylogenetic trees with poorly supported branching order (B); single or multiple gene markers that do not provide a complete picture of evolution (C).

va *et al.*, 2016; Dimitrov and Gouliamova, 2019). The second problem can be resolved using a combined sequence analysis of ITS1+2 and LSU nrDNA markers for phylogenetic reconstruction (Fig. 3). We choose these markers because LSU nrDNA has the highest discriminatory power at the level of kingdom, phylum and it decreases towards a species and genus. ITS1+2 marker has the highest discriminatory power at the level of species, genus and it decreases above the genus (Dimitrov and Gouliamova, 2014). Thus, LSU nrDNA and ITS1+2 markers cover all spectra of phylogenetic units.

The third problem can be resolved based on a comparative analysis of yeast physiological profiles. They are unique in their ability to reflect functional innovation and evolutionary adaptation of species to new niches and food sources. When physiological profiles are mapped on the phyloge-

netic tree, it is possible to translate yeast diversity into discrete phenotypic clusters which preserve the topology and cluster separation of the phylogenetic tree (Fig. 4) (Dimitrov and Gouliamova, 2019). This approach clearly shows the presence of gaps in SI, ED, and PP on the boundaries between species and genera.

Distribution of morphological and/or physiological profiles on phylogenetic tree reveals novelty that distinguishes the new genus from ITS1+2 nearest neighboring clusters. There is a minimum evolutionary distance in terms of differences in markers sequence identities or between physiological profiles between two species that belong to the same or different clusters like clades or genera as well as discontinuities of these characteristics on the boundaries between genera on the phylogenetic tree.

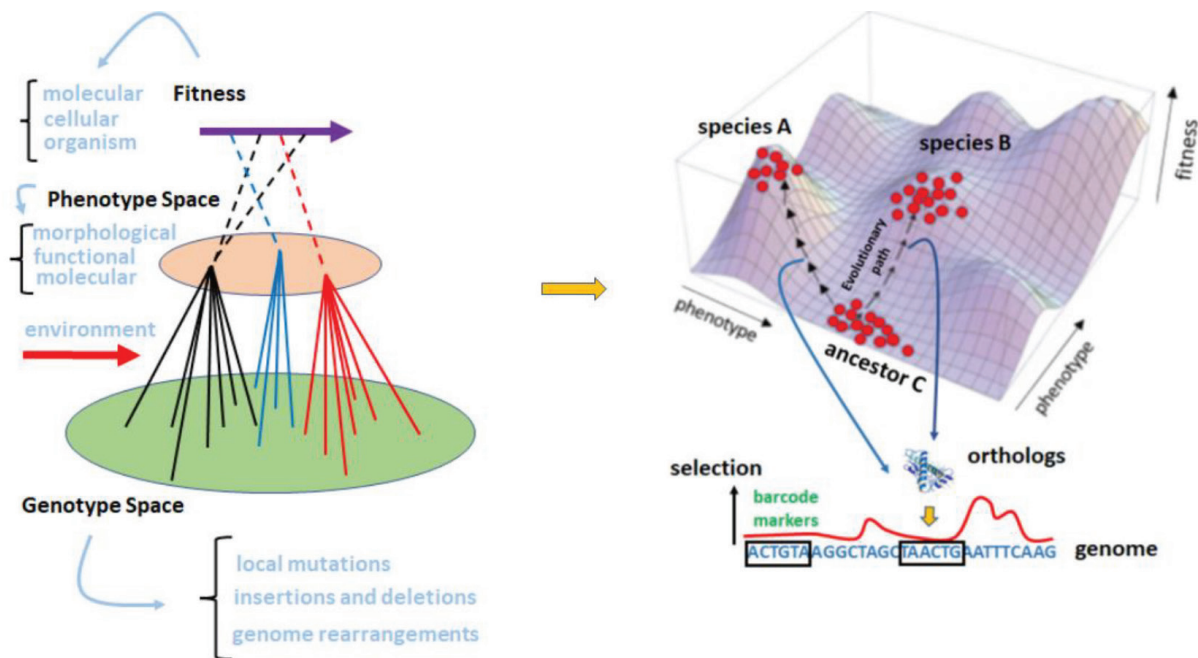


Fig. 3. Barcode marker selection. Genotype-phenotype mapping and fitness landscape of species evolution. Phylogenetic reconstructions should rely on neutral markers, which are not under selection, or at the very least not strong or directional selection.

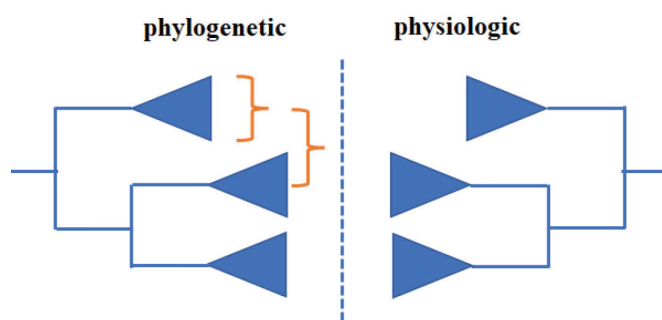


Fig. 4. Mapping of physiological profiles on phylogenetic trees preserves the topology and clusters of both trees. The mean distance between the new genus and nearest clusters will be bigger than the variability within each of them, thus confirming that they are well separated.

Local cut-off analysis for genera discrimination

We first applied our approach for the determination of boundaries of a new genus *Nematodospora* in the *Lodderomyces* clade. The analysis is based on the following considerations. First, individual species belonging to the same genus accumulate differences in some characteristics over time, while others will remain similar. It is supposed that species within a genus share more common characteristics with each other compared with species belonging to other genera. However, due to the uncertainty of evolutionary processes, it is impossible to predict which characteristics can be used to define the genus boundary. Moreover, it is even more difficult to define the boundary of a new genus when only one species is available. In this case, the only way to approximate the boundary of

a new genus is to compare locally genetic, phenotypic, and biogeographic data of the single species with its nearest neighbors on the phylogenetic tree (Fig. 5).

Second, the analysis of the biogeographic distributions of species showed that the biogeographic distribution of species within the nearest neighboring clusters and the new genus on the phylogenetic tree is different. Based on the phylogenetic, phenotypic (morphology of the ascospores), physiological profiling and biogeographic analysis of species distribution a new genus *Nematodospora valgi* gen. nov., sp. nov. was proposed in the *Lodderomyces* clade (Gouliamova *et al.*, 2016).

Global cut-off analysis for genera discrimination

Next, we applied our approach for the determination of boundaries of genera within the *Kazachstania* clade. *Kazachstania* clade has a long history of unsuccessful attempts to reconstruct a phylogenetic tree with stable topology. As a result, the *Kazachstania* clade has low statistical support for branching order and variable distribution of species and clades on the tree. This is mainly due to the high conservatism of LSU nrDNA sequences and uncertainty in the results obtained based on multigene analysis (Dimitrov and Gouliamova, 2019). Phylogenetic analysis of the alignment of combined ITS1+2 and LSU sequences was applied to resolve the problems of the *Kazachstania* clade. The analysis showed that the clade is unstable and is split into 5 Clusters (Fig. 6) that represent at least two distinct genera, the first genus comprises the

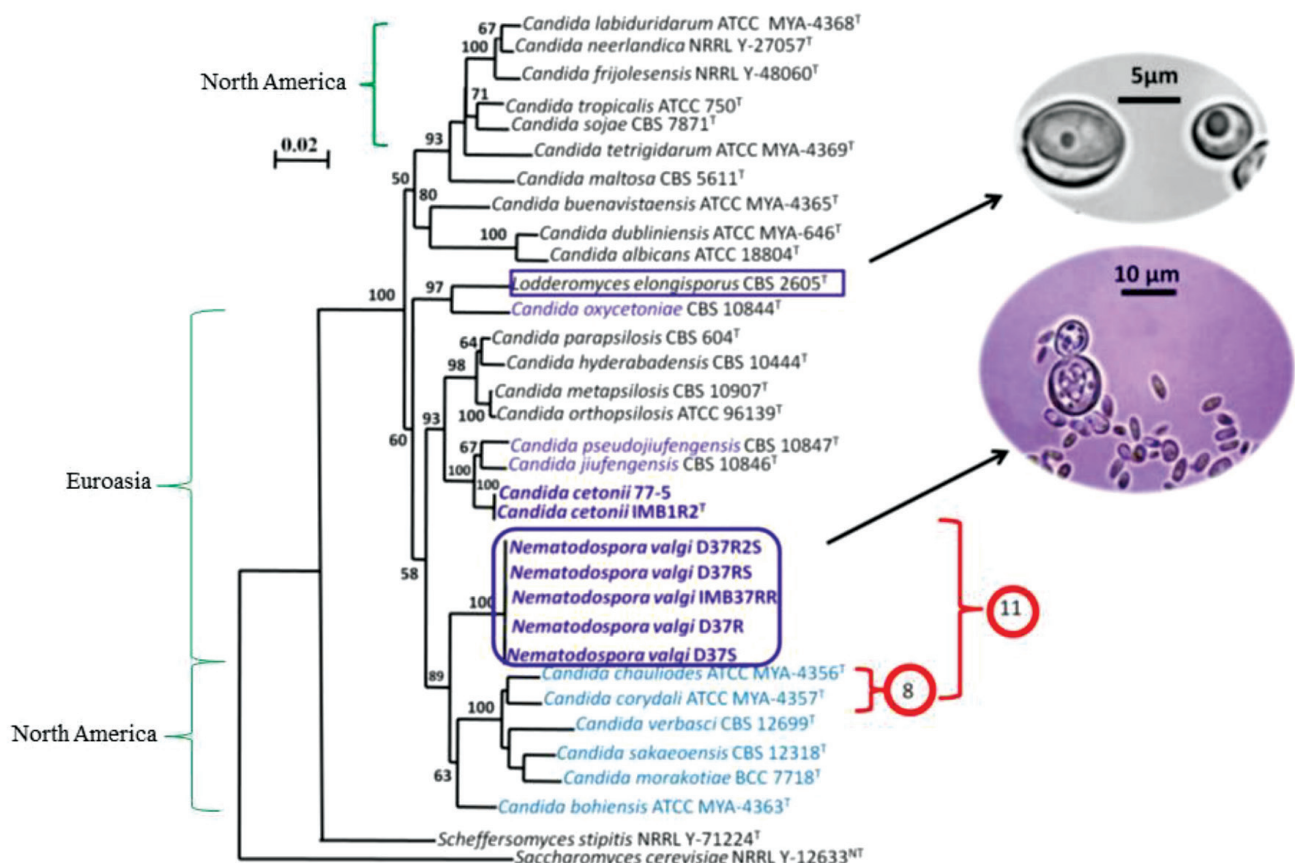


Fig. 5. Comparative analyses of DNA barcoding markers, the morphology of ascospores, physiological properties (in red), and biogeographic distribution (in blue) of yeast species of the *Lodderomyces* clade. Phylogenetic analysis using combined sequences of the ITS region and LSU nrDNA placed the novel species on a separate branch near the basal part of the *Lodderomyces* clade^{1,2}. The novel species has unique ascospore morphology distinct from the only ascospore-producing species *Lodderomyces elongisporus* and biogeographically well separated.

¹Fig. 5 is reproduced from Fungal Biology, v. 120, D. E. Gouliamova, R. A. Dimitrov, M. Th. Smith, M. Groenewald, M.M. Stoilova-Disheva, B.V. Georgiev, T. Boekhout. DNA barcoding revealed *Nematodospora valgi* gen. nov., sp. nova in the *Lodderomyces* clade, pages 179-190, Copyright Elsevier. ²Picture of *Lodderomyces elongisporus* ascospore is reproduced from (Kurtzman, 2011).

species of Clusters 1, 2, and 3 and the second genus includes the species of Cluster 4. Clusters 1 and 2 (83% bootstrap support) include thirty yeast species. Cluster 3 (100% bootstrap support) is comprised of twelve species. Cluster 4 (100% bootstrap support) includes four yeast species. Cluster 5 is comprised of an outgroup yeast species *Kluyveromyces marxianus*.

Mapping physiological profiles on a phylogenetic tree allow for analyzing the way SI and PP values of the markers are distributed between species and higher order phylogenetic clusters. We found out that there is a minimum of ED in terms of differences in marker nucleotide SI and between phenotypic profiles between two species that belong to the same or different clusters like clades or genera as well as discontinuities of these characteristics on the boundaries between genera on the phylogenetic tree. Indeed, a newly defined higher-level phyloge-

netic unit has unique combinations of phenotypic characteristics distinguishing it from those of other related phylogenetic units. Data from Dimitrov and Gouliamova (2019) in Fig. 7 show that there is a strong correlation between yeast molecular evolutionary distances and combined ITS1+2 and LSU rDNA SIV (coefficient of determination – 0.84).

There is a correlation between combined ITS1+2 and LSU rDNA SIV and differences between their PP (coefficient of determination – 0.5). There is a strong correlation between the average profile distances and the average SI between the clusters and there is also a correlation between the phylogenetic and physiological trees. Also, there is a discontinuity in PP and SI at the boundary between Cluster 4 and Cluster 3. This indicated that cluster 4 represents a new genus. Obtained cut-off values for the new genus *Grigorovia* are shown in Fig. 7.

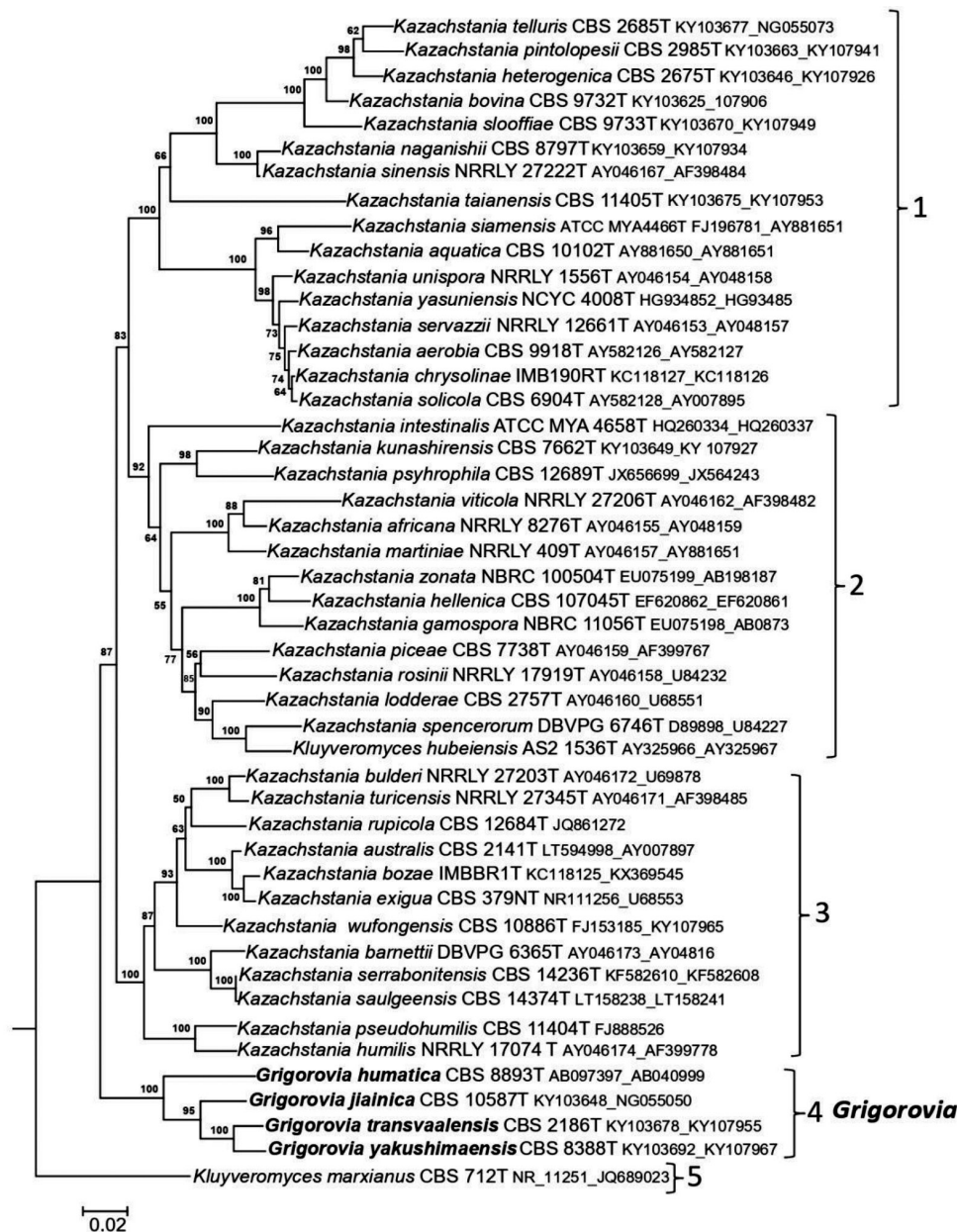


Fig. 6. Phylogenetic relationships of members of the *Kazachstania* clade inferred using neighbor-joining analysis (Kimura 2 parameter model) of combined sequences of the LSU rDNA and ITS1+2 markers. *Kluyveromyces marxianus* represents the outgroup. The scale represents 0,02 nucleotide substitutions per site (Dimitrov and Gouliamova, 2019; Gouliamova and Dimitrov, 2020).

Application of the global cut-off analysis for discrimination of the new genus in the *Starmerella* clade

In the present article, we applied our approach to the analysis of possible genera in the *Starmerella* clade because previously obtained data show low statistical support for basal branches of *Starmerella* clade (Dimitrov and Gouliamova, 2021). Our present article aims to reveal how SI and ED between genomic sequences (combined ITS1+2 and LSU nrDNA), as well as phenotypic clusters (PP) correspond to gaps at the boundaries between yeast species on the phylogenetic tree of the *Starmerella* clade.

Materials and Methods

Phylogenetic analysis

We follow the methods applied in our previous articles (Gouliamova *et al.*, 2016; Dimitrov and Gouliamova, 2019). To avoid large gaps in the alignment the 3'-overhangs of the ITS1+2 and 5' -overhangs of the LSU nrDNA sequences were removed so that all sequences ended and started at the same position. Sequences of the ITS1+2 and LSU nrDNA markers were concatenated and realigned. Phylogenetic relationships of yeasts were determined using the neighbor-joining method (Saitou and Nei, 1987) with Kimura 2 parameter model (Kimura, 1980) with a gamma distribution of rates

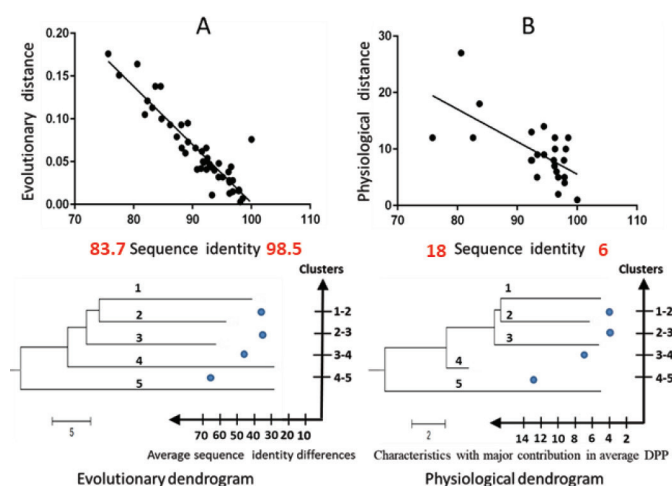


Fig. 7. Cut-off values for the new genus *Grigoriavia* (Dimitrov and Gouliamova, 2019). Correlation between evolutionary (A) and physiological (B) profile distances of type strains of species of the *Kazachstania* clade relative to sequence identity values. Comparative analysis of the phylogenetic and physiological dendrograms. There is a discontinuity in PP and SI at the boundary between Cluster 4 and cluster 3.

among sites with parameter 5, substitutions include transitions+transversions and pairwise deletion of gaps.

Maximum parsimony analysis was done with random addition of 10 initial trees, MP search method subtree-running-regrafting (SPR), MP search level 1, and max number of the tree to retain 100. All analyses were done using MEGA 7 software (Tamura *et al.*, 2021). Confidence in the branching order of phylogenetic trees was assessed by bootstrap analysis (1000 replicates) (Felsenstein, 1985).

Distance-based analysis of physiological profiles

All physiological characteristics are treated independently and each species profile is represented as a string of 1s, 0s, or ns. Thus, if a given species has a physiological characteristic there is 1 into the corresponding cross of species profile column *i* and physiological characteristic row *k*, otherwise there is 0 or n. Physiological characteristics were treated as follows: positive – 1 (+, (–/+), w, s, v, d); negative – 0 (–). When two profiles are compared – 1 ((1,0), (0,1)); – 0((1,1), (0,0)); – no data ((1, n), (n, 1), (0, n), (n, 0), (n, n)). Total differences between two physiological profiles d_{ij} were considered as a sum of compared positions.

Unweighted pair group method

SI distances d_{ij} are used to construct rooted identity trees (dendrograms) that reflect the structure of relationships present between species and preserve the topology of the phylogenetic tree. This

structure is revealed using hierarchical clustering based on the neighbor-joining method. This is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees (Tamura *et al.*, 2011).

Results and Discussion

Cut-off values of combined ITS1+2 and LSU nrDNA sequences for yeast species and genera discrimination of the Starmerella clade

Phylogenetic analysis of the alignment of combined ITS1+2 and LSU nrDNA sequences of yeast species of the *Starmerella* clade and two new Bulgarian yeast species, 42 (0) and 91WR, placed all species in 7 clusters with more than 90% bootstrap support of the major tree branches (Fig. 8).

Cluster 1 (60% bootstrap support) – 9 yeast species, cluster 2 (86% bootstrap support) – 8 yeast species, cluster 3 (83% bootstrap support) – 9 yeast species, cluster 4 (99% bootstrap support) – 4 yeast species, cluster 5 (99% bootstrap support) – 3 yeast species, cluster 6 (88% bootstrap support) – 15 yeast species and outgroup cluster 7 (100% bootstrap support) – 3 yeast species. Maximum likelihood and parsimony analyses reproduced the results of the neighbor-joining analysis with slightly lower statistical support for the major branches (data are not presented). All near-optimal trees obtained have the same tree topology (Fig. 9).

Following our previous article (Dimitrov and Gouliamova, 2019) we used combined ITS1+2 and LSU nrDNA sequence identity (SI) to determine those neighbors on the tree that under all possible rotations around nodes of the tree will have maximum SI. Identity values between all pairs of sequences in multiple alignments were determined based on the exact number of matched nucleotides relative to a shorter sequence. Secondly, we chose a few species from the boundaries of each clade, and for all possible pair boundaries belonging to different clades, we determined their average SI. Thus, all sequences that represent the nearest neighbors on the tree were determined (Fig. 10).

Determination of the species neighbors on the phylogenetic tree which under all possible rotations around nodes on the tree will have maximum sequence identity values SI of their markers. Mapping of physiological profiles on the phylogenetic tree at the global maximum on the bootstrap fitness landscape, preserves the topology and clusters of both phylogenetic and physiological trees Fig. 4 and Fig. 9.

The correlation between ED and PP of nearest

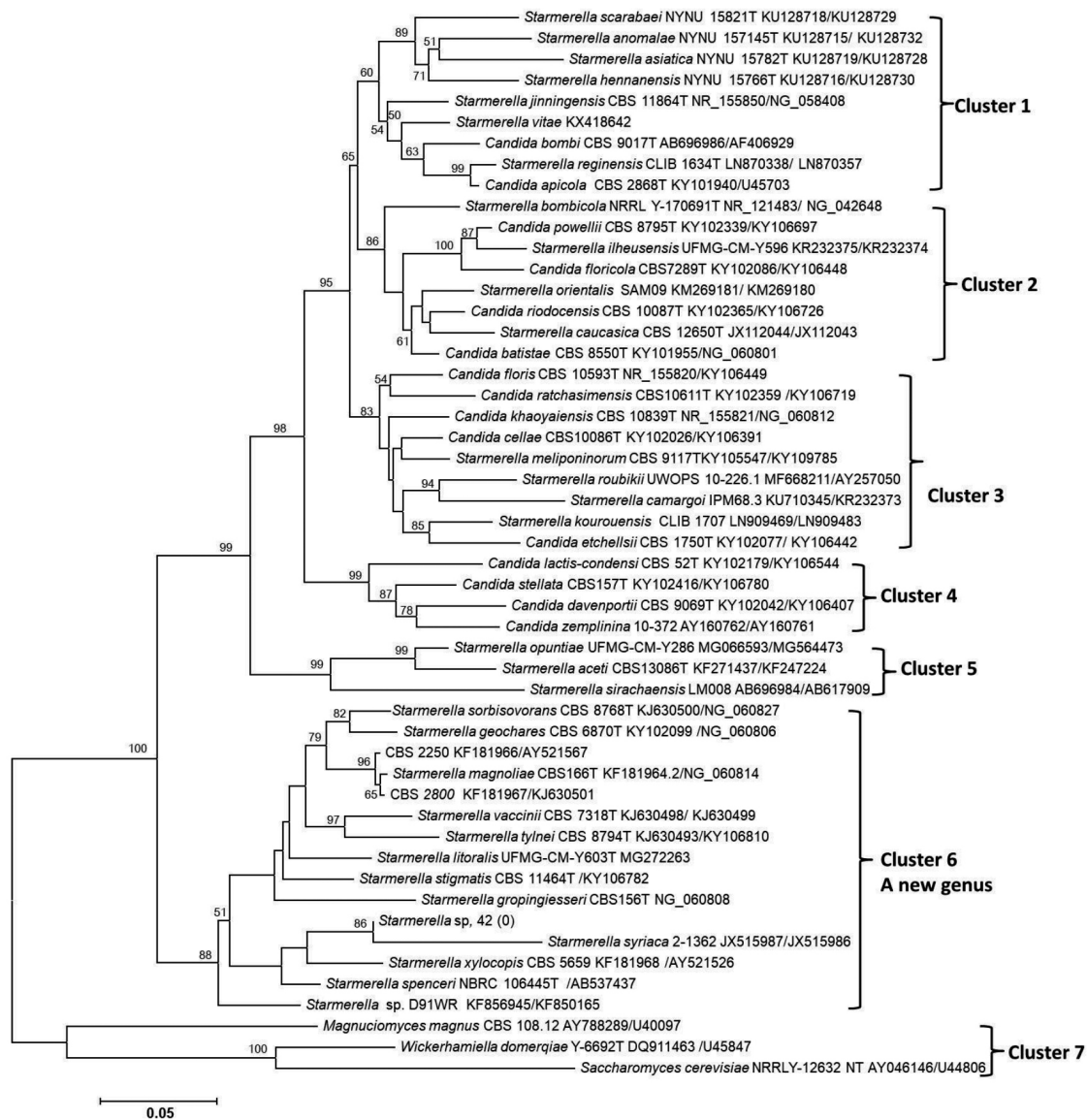


Fig. 8. Phylogenetic tree of yeast species belonging to the *Starmerella* clade, new Bulgarian yeast strains, 42 (0) and 91WR, reconstructed using combined sequences of ITS1+2 and LSU nrDNA markers. Scale nucleotide substitution per site.

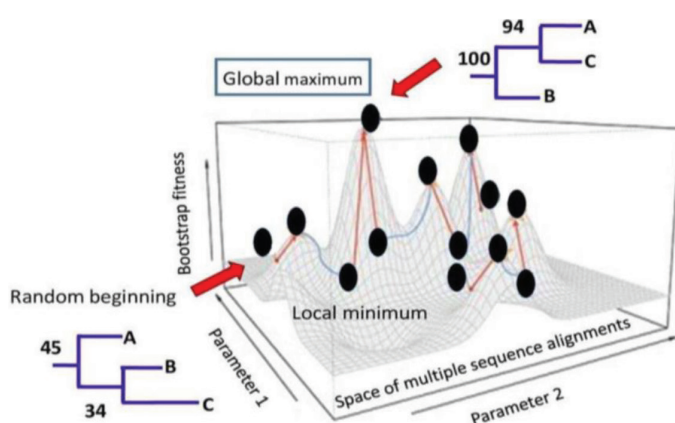


Fig. 9. In search of the global maximum for the phylogenetic tree on the bootstrap fitness landscape. Changing the parameters of the multiple alignments leads to the movement of the phylogenetic tree on the surface of the bootstrap fitness landscape. This movement leads to large or small changes in the topology of the phylogenetic tree as well as the length and bootstrap support for its branches.

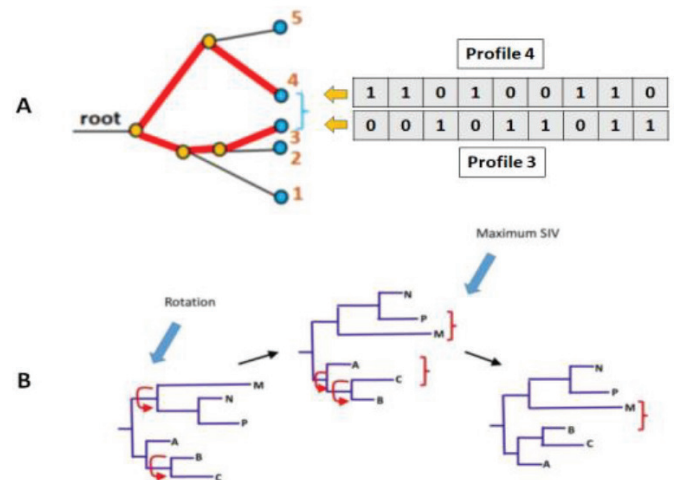


Fig. 10. A) Physiological profiles to phylogenetic tree mapping and the node pathways (marked with red) of the neutral model of marker molecular evolution from the common ancestor to two descendant species 3 and 4 are shown. B) The probabilities of evolutionary pathways are independent, which allows for rotations around phylogenetic nodes.

neighbor species of the *Starmerella* clade relative to SI as well as a comparative analysis of discontinuity of SI and PP on evolutionary dendrogram are presented in Fig. 11 (A, B).

The correlation between ED and SI is strong with the coefficient of determination $R^2 = 0.81$. In agreement with our previous results (Dimitrov and Gouliamova, 2019) when SI tends to 100% ED tends to 0 (Fig. 11A). As previously, we can approximate cut-off values using the scatter point which has the lowest ED and maximum SI. Cut-off values for discrimination of species of the *Starmerella* clade are as follows: a) SI of combined ITS1+2 and LSU nrDNA the Cut-off value is 99.27% and the cut-off value for the ED between them is 0.005. In Fig. 11A the scatter point of distances belonging to the boundary sequences of Clusters 5, 6, and 7 are well separated and located at the very beginning of the regression line with minimum SI and maximum evolutionary distance. Clusters 6 and 7 form separate outgroup clades and there is a discontinuity jump in sequences identity at the boundary between Cluster 6 and Cluster 5 (Fig. 11C). Cluster 7 is represented by *Saccharomyces cerevisiae*, *Magnusiomyces magnus*, and *Wickerhamiella domerqiae* representing the outgroup for all *Starmerella* species. The boundary values of Cluster 6 relative to Cluster 5 could be considered as the cut-off value for discrimination of separate yeast genera. Therefore, based on the average pair-boundary SI and ED we propose 72.91% as the cut-off value for SI and

0.271 as the cut-off value for ED for discrimination of genera of the *Starmerella* clade.

Cut-off values of physiological profiles for species and genera discrimination of the *Starmerella* clade

Following our previous article (Dimitrov and Gouliamova, 2019) the above procedure for determination of cut-off value for SI was also applied to obtain cut-off values for discrimination of species of the *Starmerella* clade using distances between their PP (Fig. 11B). In agreement with our previous results (Dimitrov and Gouliamova, 2019) the correlation between distances of PP and SI is valid for sequences with identities of more than 90% with a coefficient of determination of their linear correlation curve $R^2 = 0.3$. When SI is below 90% correlation is observed only between close neighbors on the tree. In terms of distances between PP and the observation that when SI tends to 100% the distances between PP tend to 7, the cut-off value for discrimination of two closely related species is determined to be 7. As mentioned in our previous article (Dimitrov and Gouliamova, 2019), values below 7 correspond to intraspecies variabilities. In Fig. 11B the scatter point of PP belonging to the boundary species of Clusters 5, 6, and 7 are well separated and located at the very beginning of the regression line with minimum SI and maximum PP. There is a gap in PP at the boundary between Cluster 5 and Cluster 6 (Fig. 11C). Based on average PP values between pair-boundary species we propose 17.5 as the cut-off value for PP for discrimination of sepa-

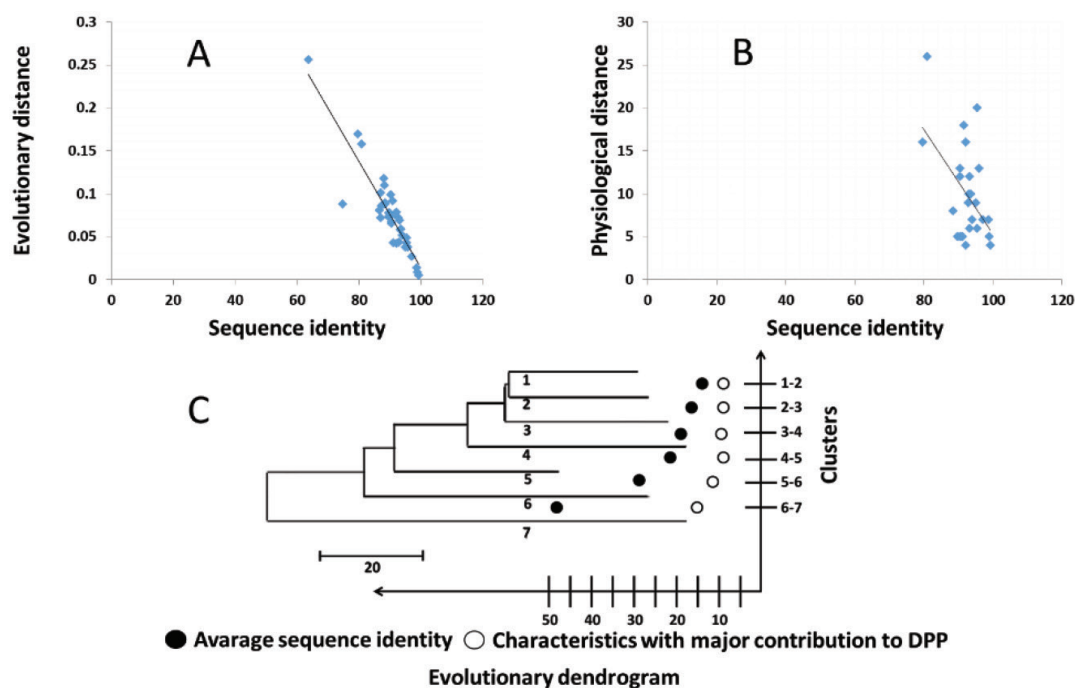


Fig. 11. Correlation between evolutionary (A) and physiological (B) profile distances of species of the *Starmerella* clade relative to SI. Comparative analysis of the average SI and physiological characteristics with a major contribution to PP (C).

rate genera in the *Starmerella* clade. As in our previous article (Dimitrov and Gouliamova, 2019), the correlation between PP and SI is not as strong. This is due to the low number of available physiological data. For the *Starmerella* clade, the missing physiological characteristics are comparable with those in the *Kazachstania* clade. For example, in the *Starmerella* clade, the minimum and maximum missing physiological characteristics for species were 1 (*C. floricola*) and 45 (*S. opuntiae*). While in the *Kazachstania* clade they were 12 (*C. pseudohumilis*) and 40 (*K. yasuniensis*) (Westerdijk Fungal Biodiversity Institute database –<http://wi.knaw.nl>).

Based on gene markers, the evolutionary model of marker sequence mutation, and outgroups, yeasts species cluster in nested monophyletic groups on phylogenetic trees based on the identity of shared sequences between them and between higher-order phylogenetic clusters. Following our previous approaches, after mapping physiological profiles on the phylogenetic tree we analyzed the distribution of marker SI and PP between species and higher order phylogenetic clusters. We expected that there should be a minimum of ED in terms of differences in marker nucleotide SI or phenotypic profiles between two species that belong to the same or different clusters like clades or genera as well as gaps in these characteristics on the boundaries between genera on the phylogenetic tree.

In search for such combinations of phenotypic characteristics that exhibit gaps at the boundaries of monophyletic genera we previously proposed a complex approach based on the comparative analyses of genetic, phenotypic, and biogeographic data (Gouliamova *et al.*, 2016). In agreement with our previous observation, the results in this article show that there is a strong correlation between yeast molecular ED and combined ITS1+2 and LSU nrDNA SIV (coefficient of determination – 0.84). There is a lower correlation between combined ITS1+2 and LSU nrDNA SI and differences between their PP (coefficient of determination – 0.3). There is a strong correlation between the average profile distances and the average SI between the clusters. Finally, mapping of physiological profiles on the phylogenetic tree at the global maximum on the bootstrap fitness landscape preserves the topology and clusters of both phylogenetic and physiological trees. A formal description of the new genus and new Bulgarian species will be presented in our next article.

Conclusions and Future Perspectives

The results presented above confirm further that the approach proposed in our previous articles and reviewed in the introduction of the present article represents a reliable method for defining the boundaries between different species and genera on a phylogenetic tree based on unified criteria. The next step is based on whole genome sequencing and RNA-seq analysis of coding and non-coding gene expression to analyze differences in morphological and physiological properties and how they change at the boundaries between yeast species and genera.

Acknowledgements:

Our research was financially supported by the grant KP-06-N31/19 from the Bulgarian Science Fund.

References

- Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783-791.
- Dimitrov, R., D. Gouliamova (2014). New method for sequence alignment based on probabilities of nucleotide correspondances. *Biotechnol. Biotech. Equip.* **26**: 218-223.
- Dimitrov, R., D. Gouliamova (2019). Genetic and phenotypic cut-off values for species and genera discrimination of the *Kazachstania* clade. *C. R. Acad. Bulg. Sci.* **72**: 350-357.
- Gouliamova, D., R. Dimitrov (2020). *Kazachstania chrysolineae* and *Kazachstania bozae* two new yeast species of the genus *Kazachstania*. Transfer of four *Kazachstania* species to *Grigorovia* gen. nov. as new combinations. *C. R. Acad. Bulg. Sci.* **73**: 48-57.
- Dimitrov, R. A., D. E. Gouliamova (2020). Phylogenetic cut-off values for species and genera discrimination of yeasts. Poster presentation at EMBL Conference *Molecular Mechanisms in Evolution and Ecology*, Heidelberg, Germany, 30 September – 2 October.
- Lachance, M. A. (2018). C. P. Kurtzman's evolving concepts of species, genus and higher categories. *FEMS Yeast Res.* **18**: foy103. [10.1093/femsyr/foy103](https://doi.org/10.1093/femsyr/foy103).
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111-120.
- Kurtzman, C. P. (2011). *Lodderomyces* van der Walt (1971). In: Kurtzman, C. P., J. W. Fell, T. Boekhout (Eds.), *The Yeast: A Taxonomic Study*. Elsevier, London, pp. 561-563.
- Saitou, N., M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406-425.
- Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Mol. Biol. Evol.* **289**: 2731-2739.
- Tamura, K., G. Stecher, S. Kumar (2021). MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**: 3022-3027.