An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference

Guillermin Agüero-Chapin[a,b,c], Aminael Sánchez-Rodríguez[d], Pedro I. Hidalgo-Yanes[b,e], Yunierkis Pérez-Castillo[b], Reinaldo Molina-Ruiz[b], Kathleen Marchal[d], Vítor Vasconcelos[a,c] and Agostinho Antunes[a,c*]

[a] *CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas, 177, 4050-123 Porto, Portugal*

[b] *Molecular Simulation and Drug Design (CBQ), Universidad Central ¨Marta Abreu¨ de Las Villas (UCLV), Santa Clara, 54830, Cuba*

[c] *Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Portugal*

[d] *CMPG, Department of Microbial and Molecular Systems, KU Leuven, Kasteelpark Arenberg 20, B-3001 Leuven, Belgium*

[e] *Area of Microbiology, University of León, 24071 León, Spain*

***Corresponding author:** Antunes, A., E-mail: aantunes@ciimar.up.pt or aantunes777@gmail.com

Abstract

The ITS2 gene class shows a high sequence divergence among its members that have complicated its annotation and its use for reconstructing phylogenies at a higher taxonomical level (beyond species and genus). Several alignment strategies have been implemented to improve the ITS2 annotation quality and its use for phylogenetic inferences. Although, alignment based methods have been exploited to the top of its complexity to tackle both issues, no alignment-free approach have been able to successfully address both topics. By contrast, the use of simple alignment-free classifiers, like the topological indices (TIs) containing information about the sequence and structure of ITS2, may reveal to be a useful approach for the gene prediction and for assessing the phylogenetic relationships of the ITS2 class in eukaryotes. Thus, we used the **TI2BioP** (**T**opological **I**ndices **to** **BioP**olymers) methodology, freely available at http://ti2biop.sourceforge.net/ to calculate two different TIs. One class was derived from the ITS2 artificial 2D structures generated from DNA strings and the other from the secondary structure inferred from RNA

folding algorithms. Two alignment-free models based on Artificial Neural Networks were developed for the ITS2 class prediction using the two classes of TIs referred above. Both models showed similar performances on the training and the test sets reaching values above 95% in the overall classification. Due to the importance of the ITS2 region for fungi identification, a novel ITS2 genomic sequence was isolated from *Petrakia* sp. This sequence and the test set were used to comparatively evaluate the conventional classification models based on multiple sequence alignments like Hidden Markov based approaches, revealing the success of our models to identify novel ITS2 members. The isolated sequence was assessed using traditional and alignment-free based techniques applied to phylogenetic inference to complement the taxonomy of the *Petrakia* sp. fungal isolate.