DNA barcoding of green microalgae using *rbc*L and *nu*ITS2 markers reveals putative new species

Sámed I. I. A. Hadi[1,2]; Hugo Santana[1, 3]; Patrícia P. M. Brunale[1]; Taísa G. Gomes[1,4]; Márcia D. Oliveira[5]; Alexandre Matthiensen[6]; Marcos E. C. Oliveira[7]; Flávia C. P. Silva[1]; Bruno S. A. F. Brasil[1]*

Background

DNA barcoding is a method for species identification, which identifies specimens based on DNA sequence similarity against a sequence database of *a priori* defined species. DNA-based identification is particularly useful unveiling cryptic diversity at various taxonomic levels and identifying species where structural characters are few or difficult to observe (HEBERT et al., 2004). The green algae, Chlorophyta, comprise an ancient and taxonomic diverse lineage with approximately 8,000 described species. It is estimated that at least 5,000 species still remain undescribed, notably in tropical and subtropical areas (GUIRY, 2012). There is a growing interest in using green microalgae for biotechnological applications such as the production of fuels, chemicals, food and animal feed. This study aimed to identify neotropic green microalgae specimens isolated from Brazilian inland waters through the use of *rbc*L and *nu*ITS2 markers as DNA barcodes. Novel primers for neotropic specimens' *rbc*L gene amplification and sequencing are presented, as well as comparisons between *rbc*L and *nu*ITS2 markers variability, primers universality and databases accuracy and completeness.

Methods

The environments of collection included natural freshwater bodies within the Amazon rainforest, the Cerrado Savanna and the Pantanal flooded grasslands, as well as anthropogenic wastewater deposits from the sugarcane industry (vinasse), pisciculture ponds and swine farm derived wastewater. A total of 51 strains were isolated and deposited in the Collection of Photosynthetic Microorganisms for Agroenergy Research at Embrapa (Brasília/DF – Brazil). Total genomic DNA was isolated from 30 mg of fresh algal biomass using the CTAB DNA extraction protocol adapted by Doyle e Doyle (1987). Sequences were determined bi-directionally for at least two different amplicons using the BigDye

1 Embrapa Agroenergy, Brasília/DF – Brazil;
2 UFMG, Belo Horizonte/MG – Brazil;
3 UFBA, Vitória da Conquista/BA – Brazil;
4 UnB, Brasília/DF – Brazil;
5 Embrapa Pantanal, Corumbá/MS – Brazil;
6 Embrapa Swine and Poultry, Concórdia/SC – Brazil;
7 Embrapa Amazônia Oriental, Belém/PA – Brazil;
*s.ihadi@hotmail.com

Terminator v.3.1 Cycle Sequencing Kit on the automated DNA sequencer ABI 3130. The *nu*ITS2 sequences were annotated and trimmed using the *ITS2 Annotation* tool available at ITS2 Database (KOETSCHAN et al., 2012). For similarity searches, the *rbc*L sequences were submitted to the Barcode of Life Data Systems (BOLD systems) using the *Plant identification* tool, while *nu*ITS2 sequences were submitted to BLASTN for comparisons against sequences deposited at Genbank.

Results and Conclusions

The *nu*ITS2 region could be successfully sequenced from DNA samples extracted from 47 strains (94,12% sequencing success rate) by using a pair of universal primers. On the other hand, primers described in the literature for the *rbc*L gene does not achieved a satisfactory success rate, thus novel sets of primers were designed. Two of the newly designed primer pairs could successfully amplify and sequence 82,35% and 50,98% of the dataset. The combination of the sequencing results obtained from these two *rbc*L primer pairs allowed the construction of quality consensus sequences (QV>20) for 49 samples (96,08% sequencing success rate). DNA from all 51 Embrapa|LBA strains could be amplified and sequenced for at least one of the markers tested. A total of 23 distinct *nu*ITS2 genotypes and 26 distinct *rbc*L genotypes were obtained. Similar levels of diversity were observed among *nu*ITS2 (0.482) and *rbc*L (0.461) sequences. In order to perform the molecular identification of Embrapa|LBA strains, the *rbc*L sequences obtained were submitted to similarity searches against the DNA barcoding dedicated database, BOLD systems. The closest matches retrieved for *rbc*L sequences ranged from 90% to 99% of similarity. Currently, there are very few *nu*ITS2 sequences from chlorophytes deposited at taxonomically curated databases such as BOLD, therefore similarity searches were performed against the GenBank. The closest matches retrieved for *nu*ITS2 sequences ranged from 81% to 100% of similarity. Of all Embrapa|LBA *nu*ITS2 and *rbc*L sequences, only 8 *nu*ITS2 sequences retrieved matches with 100% similarity. Embrapa|LBA strains retrieved matches from species that belong to the Chlorophyceae and Trebouxiophyceae classes, especially to the orders Chlamydomonadales, Chlorococcales, Sphaeropleales and Chlorellales. Noteworthy, 66% of the *nu*ITS2 matches retrieved from GenBank lacked the Latin binomial that characterizes the complete species name, compared to 10% of the *rbc*L matches retrieved from BOLD. The results reported so far indicate that 17 and 0 Embrapa|LBA strains can be unambiguously identified to the species level based on *nu*ITS2 or *rbc*L sequences, respectively. Since the closest *nu*ITS2 matches retrieved for strains Embrapa|LBA#2-3, #26, #30, #32-36 and #42-44 lack the Latin binomial that characterizes the complete species name, phylogenetic analysis using reference sequences from currently accepted *Chlorella*, *Micractinium* and *Desmodemus* species were conducted to check if it is possible to assign a specific Latin name to this samples. The results demonstrate that *nu*ITS2 sequences from these strains group

together with their closest matches in monophyletic clades, suggesting that they probably belong to species not formally described. The results presented here indicate that a dual marker DNA barcoding pipeline based on *nu*ITS2 (primary marker) and *rbc*L (secondary marker) might be useful not only for green microalgae species identification, but also to assist new species discovery. Once available at its full potential, a DNA-based Chlorophyta identification method, together with its large-scale applications (e.g.: Metabarcoding), will bring considerable advances to areas such as biodiversity monitoring, water quality control and microalgae biotechnology.

Financial Support

Embrapa, CAPES and CNPq.

References

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, Irvine, v. 19, n. 1, p. 11-15, 1987.

GUIRY, M. D. How many species of algae are there? **Journal of Phycology**, New Jersey, v. 48, n. 5, p. 1057-1063, 2012.

HEBERT, P. D. N.; PENTON, E. H.; BURNS, J. M.; JANZEN, D. H.; HALLWACHS, W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. **Proceedings of the National Academy of Sciences**, Washington, v. 101, n. 41, p. 14812-14817, 2004.

KOETSCHAN C. HACKL, T.; MULLER, T.; WOLF, M.; FORSTER, F.; SCHULTZ, J. ITS2 DATABASE IV: Interactive taxon sampling for internal transcribed spacer 2 based phylogenies. **Molecular Phylogenetics and Evolution**, Copenhagen, v. 63, n. 3, p. 585-588, 2012.