Supervised DNA Barcodes species classification: analysis, comparisons and results

Background

Specific fragments, coming from short portions of DNA (e.g., mitochondrial, nuclear, and plastid sequences), have been defined as DNA Barcode and can be used as markers for organisms of the main life kingdoms. Species classification with DNA Barcode sequences has been proven effective on different organisms. Indeed, specific gene regions have been identified as Barcode: *COI* in animals, *rbcL* and *matK* in plants, and *ITS* in fungi. The classification problem assigns an unknown specimen to a known species by analyzing its Barcode. This task has to be supported with reliable methods and algorithms.

Methods

In this work the efficacy of supervised machine learning methods to classify species with DNA Barcode sequences is shown. The Weka software suite, which includes a collection of supervised classification methods, is adopted to address the task of DNA Barcode analysis. Classifier families are tested on synthetic and empirical datasets belonging to the animal, fungus, and plant kingdoms. In particular, the function-based method Support Vector Machines (SVM), the rule-based RIPPER, the decision tree C4.5, and the Naïve Bayes method are considered. Additionally, the classification results are compared with respect to *ad-hoc* and well-established DNA Barcode classification methods.

Results

A software that converts the DNA Barcode FASTA sequences to the Weka format is released, to adapt different input formats and to allow the execution of the classification procedure. The analysis of results on synthetic and real datasets shows that SVM and Naïve Bayes outperform on average the other considered classifiers, although they do not provide a human interpretable classification model. Rule-based methods have slightly inferior classification performances, but deliver the species specific positions and nucleotide assignments. On synthetic data the supervised machine learning methods obtain superior classification performances with respect to the traditional DNA Barcode classification methods. On empirical data their classification performances are at a comparable level to the other methods.

Conclusions

The classification analysis shows that supervised machine learning methods are promising candidates for handling with success the DNA Barcoding species classification problem, obtaining excellent performances. To conclude, a powerful tool to perform species identification is now available to the DNA Barcoding community.