Possible bias in predicting essential genes

Zandora Celeste Hastenreiter Ferreira Nunes, Francisco Pereira Lobo, Giovanni Marques de Castro

Universidade Federal de Minas Gerais

Abstract

The ever-growing amount of complete genomes calls for the creation of computational strategies in order to extract biologically meaningful information from this data. The prediction of essential genes is one of the areas where there is a growing effort to create such strategies, since they are attractive targets for intervention in parasites, vectors and pests. The use of machine learning can help detecting such genes and reduce the burden for experimental validation, aiding the choosing of interesting candidates. For this work, we used a Random Forest algorithm to predict essentiality in insect genes, as this information would potentially help in the development of specific, low impact bioinsecticides. Most of the currently available software for this task relies on information based on expensive experimental data, such as gene annotation, gene expression profiles and interaction networks. Our approach is homology-independent in the sense that the features are calculated from its nucleotide and peptide sequence data alone, such as amino acid/nucleotide frequencies and sequence entropy. Initially, we recovered alleles of Drosophila melanogaster from FlyBase classified as loss of function, amorphic and hypomorphic. For the essential genes, we recovered the alleles annotated as having a lethal phenotype and, for the nonessential genes, we recovered those that are not annotated as lethal. We converted the allele ids to gene ids and retrieved the longest isoform nucleotide sequence for that gene. The training dataset consisted of 1256 essential genes and 636 nonessential genes, and the test set consisted of 47 essential and 88 nonessential genes. The genes present in the test dataset were curated from published works. The model consisted of 1000 trees and 10 fold cross-validation, repeated 5 times. We obtained an AUC of 0.7238. We found this result to be caused by a bias introduced by a small group of evolutionary old essential genes and young nonessential genes; when we withdraw these groups from the test set, the AUC dropped to 0.578. Surprisingly, when we use only this biased group as the test set, the AUC obtained was 0.995, suggesting our classifier learned to discriminate between old and young genes. As future work, in order to further evaluate our model, we will retrieve orthologs present only in D. melanogaster and D. simulans and consider these as young genes; orthologs also present in D. mojavensis and D. virilis will considered old genes. The new test set will then comprise essential young genes and old nonessential genes.

Funding: CAPES, CNPq, PPG Genética - UFMG