The challenge of finding genes in sugarcane

Kashiwabara[1], AY. Durham, AM[1].
1- Universidade de São Paulo, Brazil

Introduction

Ab initio gene finders are very important to quickly produce evidences of different protein-coding gene structures and play a important role in novel genome sequences. The accuracy of gene finders depends on a set of factors, including the use of a high quality training set. Predicting genes in sugarcane presents a challenge due to the limited coverage of transcript data and the lack of a large dataset of confirmed genes. The solution adopted to far in Sugarcane sequencing is the use of *Z. mays* data to train the gene predictor Augustus, which is then applied on the available BAC data. The task of manually validating this prediction, however, is challenging, error prone and extremely time-consuming. We present here a study on gene prediction in sugarcane that automatically evaluates the current prediction, evaluates the SAS sequence assembly, show a better gene predictor for Sugarcane, and produces a set of 366 gene predictions that have at least partially validated by EST data, and 857 novel predictions.

Results and Conclusions

We used three gene predictors to discover genes in Sugarcane: PASA, Augustus and MYOP. PASA creates consensus from mRNA mappings. The strict quality evaluation of the mappings guarantees that PASA is very accurate, albeit dependent on the coverage of transcript sequences. With PASA we used mappings from two different datasets, SUCEST (Sugarcane ESTs) and SAS (Sugarcane Assembled Sequences) sequences. Augustus is one of the most successful gene predictors available and has been used in the initial annotation of the sugarcane BACs sequenced so far. MYOP is a gene prediction platform that is being developed by our group. We trained MYOP using a *Z. mays* dataset and downloaded *Z. mays* tables from the Augustus (version 2.5.5).

In a second step, we used PASA's SUCEST results for a series of tasks: validating SAS assemblies, evaluating the quality of Augustus and MYOP as gene predictors for sugarcane, and producing a consensus using the data of all gene predictions.

As a result of the entire process, we have established MYOP superiority when using the *Z. mays* training set, with an increase in sensitivity (34% vs. 31%) and a small decrease in positive prediction value (92% vs. 93%). More important, MYOP correctly detected over 7% more confirmed gene nucleotides. In addition, MYOP has produced predictions of new coding regions totaling 478.699nt, of which our most conservative measurements indicate 80% should be correct.

PASA results using the SUCEST and SAS sequences that have consistent mapping on the BACs, indicate that the SAS assembly produced many artifacts: of the 258.644nt mapped by SAS, 49.271nt (~20%) were not

confirmed by EST data (it should be 100%). Also EST mapping was able to detect an extra 213.467nt of mapped regions.

The last step of this work was use PASA to generate consensus predictions from the results of MYOP, Augustus and PASA-SUCEST. This produced, for the 56 BACs analyzed, 508 genes, which should provide a more reliable gene set for future annotation.

However there is still a way to go to find all sugarcane genes. A total of 213.467nt of coding regions mapped by ESTs were not detected by any of the gene predictors. We will now use the new gene dataset found here to train MYOP, in an effort to increase gene discovery.

Partially Supported by: CNPq 141069/2007

Author publications

Machado-Lima A, Kashiwabara AY, Durham AM (2010)*. Decreasing the number of false positives in sequence classification. BMC Genomics. 2010 Dec 22;11 Suppl 5:S10.*

Araújo FM, Machado-Lima A, Durham AM, Teixeira R, Oliveira G. (2009). *Sequence and structural analysis of the 5' noncoding region of hepatitis C virus in patients with chronic infection*. J Med Virol. 2009 Jul;81(7):1212-9.