Beoung Hun Lee

CS 199

March 23, 2019

Edwin Solares

In this paper, the researchers introduce AUGUSTUS+, an upgraded program of AUGUSTUS, that utilizes expressed sequence tag(EST), protein sequence database and user-defined hints to predict a gene. AUGUSTUS+ combines both intrinsic method, which utilizes a generalized hidden Markov Model(GHMM) to predict gene, and extrinsic method, which utilizes existing information about EST or protein to predict gene. Based on the gene that is predicted by the GHMM, the gene is either rewarded with higher score or given penalty depending on the external hints. This method allows the program to filter out false positive while increasing a chance of finding the correct gene. External hints consist of information about start codon, stop codon, part of exon, entire exon, splice site or use-specified hints. AUGUSTUS+ can also combine hints to further confirm the prediction result.

AUGUSTUS+ was tested on two human data sets and the results were compared with other existing programs, including AUGUSTUS. Four different settings of AUGUSTUS+ was used, one that only uses EST information, one that only uses protein information, one that only uses combined hints, and one that uses all of the information. On the benchmark set sagl78, AUGUSTUS+ outperformed other programs in all setting. The accuracy of prediction increased as the more hints were provided, so the setting that uses all the hints had the highest accuracy. On more realistic chromosome 22, AUGUSTUS+ showed lower accuracy, but still outperformed other programs.

Reference

1. Stanke, M., Schoffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*. 7:62.