

Abstract from PubMed

J Theor Biol. 2010 May 20. [Epub ahead of print]

A classification-based prediction model of messenger RNA polyadenylation sites.

Ji G, Wu X, Shen Y, Huang J, Q. Quinn Li

Department of Automation, Haiyun Keyan Building 2, Room 212, Xiamen University, Xiamen, 361000, China.

Abstract

Messenger RNA polyadenylation is one of the essential processing steps during eukaryotic gene expression. The site of polyadenylation [(poly(A) site] marks the end of a transcript, which is also the end of a gene. A computation program that is able to recognize poly(A) sites would not only prove useful for genome annotation in finding genes ends, but also for predicting alternative poly(A) sites. Features that define the poly(A) sites can now be extracted from poly(A) site datasets to build such predictive models. Using methods, including K-gram pattern, Z-curve, position-specific scoring matrix and first-order inhomogeneous Markov sub-model, numerous features were generated and placed in an original feature-space. To select the most useful features, attribute selection algorithms, such as Information Gain and Entropy, were employed. A training model was then built based on Bayesian Network to determine a subset of the optimal features. Test models corresponding to the training models were built to predict poly(A) sites in Arabidopsis and rice. Thus, a prediction model, termed Poly(A) site Classifier, or PAC, was constructed. The uniqueness of the model lies in its structure in that each sub-model can be replaced or expanded, while feature generation, selection and classification are all independent processes. Its modular design makes it easily adaptable to different species or datasets. The algorithm's high specificity and sensitivity were demonstrated by testing several datasets, and, at the best combinations, they both reached 95%. The software package may be used for genome annotation and optimizing transgene structure. Copyright © 2010 Elsevier Ltd. All rights reserved.
PMID: 20546757 [PubMed - as supplied by publisher]