SUPPLEMENTARY MATERIAL 1: Models

Bioinformatics: Application Note

Dragon PolyA Spotter: Predictor of poly(A) motifs within human genomic DNA sequences

Manal Kalkatawi[1,#], Farania Rangkuti[1,#], Michael Schramm[1,#], Boris R. Jankovic[1,#], Allan Kamau[1], Rajesh Chowdhary[2], John A.C. Archer[1], Vladimir B. Bajic[1,*]

[1] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

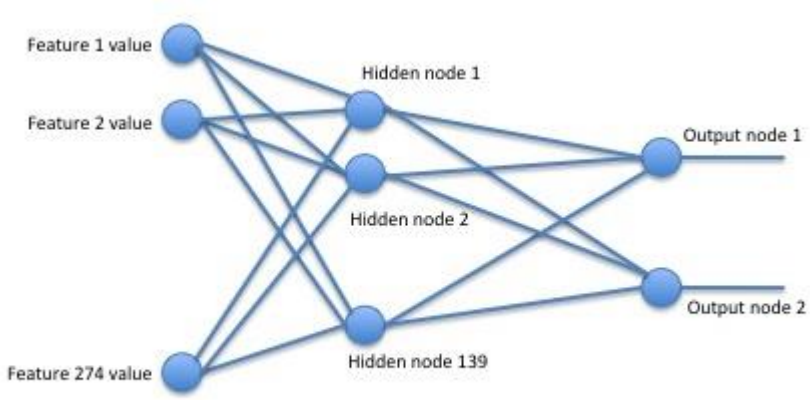
[2] Biomedical Informatics Research Center, MCRF, Marshfield Clinic, 1000 North Oak Avenue, Marshfield, WI 54449, USA

In this Supplementary Material, we briefly outline the structure of our prediction models as well as the process of deriving them. As mentioned in the main text, our Poly(A) motif prediction tool implements two types of prediction models, one based on Artificial Neural Networks (ANNs) and the other based on Random Forest (RF). User can select either of these methods for prediction purposes.

ANN models

All ANN models have the structure as depicted Supplementary Figure 1. The input layer accepts 274 features. The hidden layer contains [2+(Number of Features/2)] nodes. This structure was selected after some careful experimentation. The output layer contains two neurons where output signal of the first one corresponds to positive predictions and the output signal of the second one corresponds to negative ones. The higher signal determines the prediction. In order to avoid overfitting of the model, we deployed a variant of early stopping methodology described in (Zang and Yu, 2005) with a suitably selected value for epochs.

Supplementary Figure 1: Structure of the ANN used in our model



Random Forest Models

We trained RF type classifiers (Breiman, 2001) as implemented in WEKA (Hall at al., 2009). For each of the 12 variant datasets, we created a separate model using the same 274 features that we used in the case of ANN model. Once RF models were trained, we used these models as classifiers that are invoked by WEKA. We considered several models with respect to the number of trees evaluated their accuracy with cross-validation of different numbers of folds. The best result was achieved by growing 100 trees without restricting maximal depth using 9 random features at each nod, estimating it accuracy on 100 cross-validation.

Supplementary References:

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations, Vol.11, Iss.1

Breiman L. (2001) Random Forests. Machine Learning. 45(1): 5-32.

Zhang T, Yu B. (2005) Boosting with early stopping: Convergence and consistency, Ann. Statist. 33(4):1538-1579.