COEN 432 - Assignment#3 ML

I. Description

The HDV-LIG4_fitness_table spreadsheet provides data on the Genotype and fitness of an HDV ribozyme (a form of catalytic RNA). All ribozymes are of equal length and differ on only 14 locations within the following sequence, marked in red.

GGACCATTCGAMTCCCATTAGRCTGGKCCGCCTCCTSGCGGCGGGAGTTGSGCKAGGGA
GGAASAGYCTTYYCTAGRCTAASGMSCATCGATCCGGTTCGCCGGATCCAAATCGGGCTT
CGGTCCGGTTC

The fitness of a ribozyme is a real value from [0,1]. Once the red nucleotides are replaced by actual nucleotides (from the Genotype, and in the same order), the whole sequence will comprise of nucleotides from {A, C, T, G}. In addition to the sequence, it is possible to fold a sequence using an RNA folder (e.g., http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi), which will provide a 'secondary structure' comprising of {., (,)} symbols reflecting unpaired and paired nucleotides.

II. Inputs

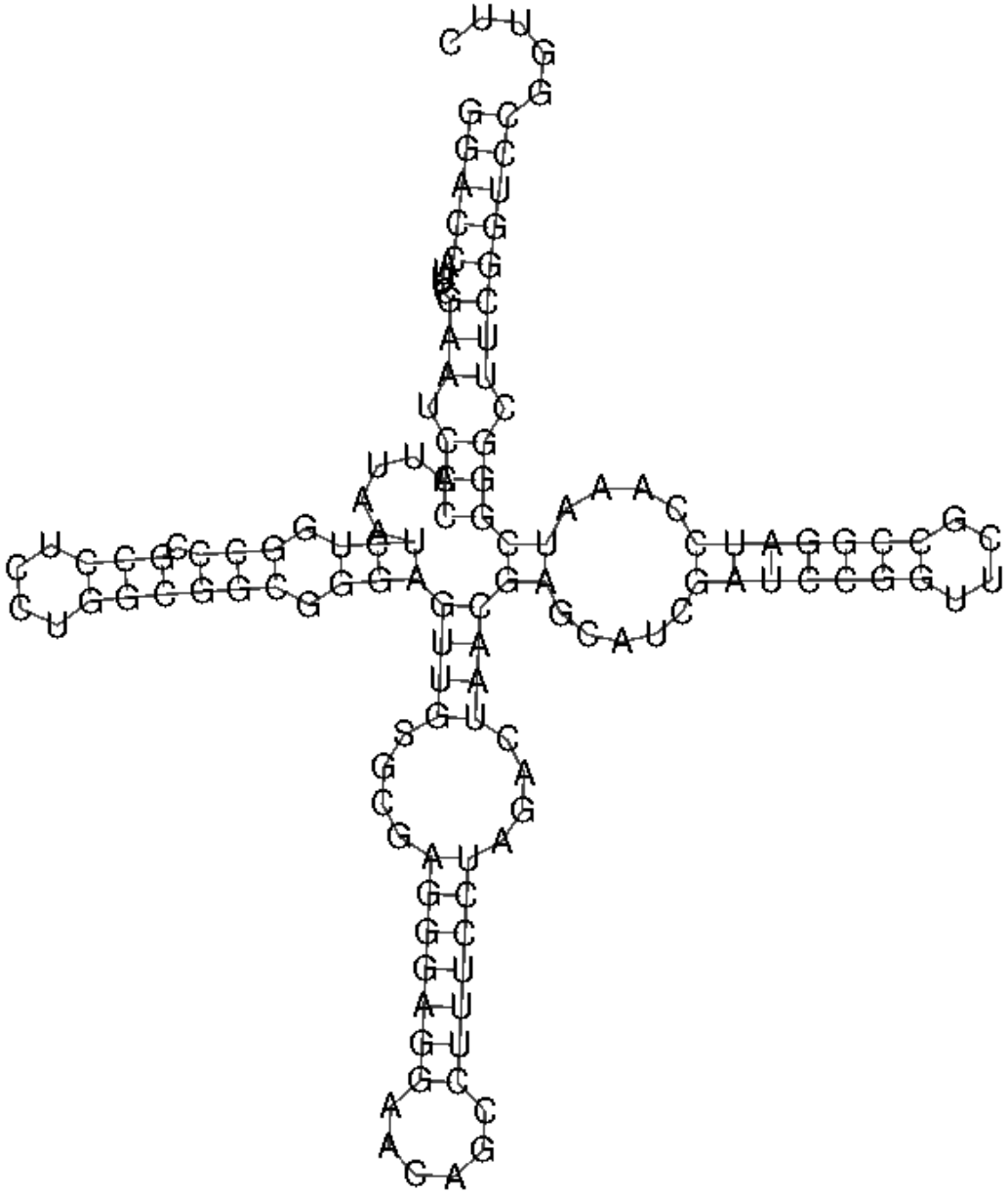
From a machine learning perspective, a single instant of an HDV can be written as (sequence, structure, fitness), such as the example below. In this case, the genotype AATCGGCCTCACAG was used to replace the red symbols.

GGACCATTCGAATCCCATTAATCTGGCCCGCCTCCTGGCGGCGGGAGTTGSGCGAGGGA
GGAACAGCCTTTCCTAGACTAACGAGCATCGATCCGGTTCGCCGGATCCAAATCGGGCTT
CGGTCCGGTTC

Given a folding algorithm, a possible secondary structure for this sequence (assuming certain conditions of temperature and ionic concentrations) is:

(((((....(((.(((.....(((.(((.(((....)))))).)))((((....(((((((......)))))))....))))((.....(((((((....)))))))....)))))..)))))).....

Visually, this can be represented as:



III. Output:

The challenge for you is to use a learning algorithm that would, given the sequence (and a possible structure) of an HDV, predict the fitness of that HDV. The samples provided can be used (e.g., by splitting the dataset 80:20 *or* via 10-fold cross-validation) for training and testing. No matter what model you decide to use and how you decide to use it, your objective is to achieve the best possible prediction of accuracy.