Technical take home exercise

Background:

We're interested in extracting knowledge from proprietary biological datasets and then using it to make predictions.

Your task:

Proteins can be described using strings of letters (e.g. *EWFSPPFSWC*). You have been provided with a training dataset of 1,000,000 protein sequences (see *TRAINING_SET.csv*). Every protein sequence in the training dataset has been given a fitness score of either 1 or 0. Using the training set, your task is to predict the fitness scores (either 1 or 0) of the 50,000 sequences in the file *PROBLEM_SET.csv* . You may adopt any approach you like.

Files provided:

Filename	**Description**
TRAINING_SET.csv	1,000,000 **scored** protein sequences
PROBLEM_SET.csv	50,000 **unscored** protein sequences

Expected output:

Please return

(a) A brief outline of your strategy for tackling this problem.
(b) The file *PROBLEM_SET.csv* with your fitness score predictions contained in an adjacent column (i.e. in same format as *TRAINING_SET.csv*).
(c) Any code that you wrote.

Please submit a zipped folder containing your file(s) to ***james@labgeni.us***, within 3 hr 10 mins of receipt of this document.

Good luck!