1.1)

a) Stochastic gradient descent could be taken as an optimization of gradient descent approach.
b) To better fit a set of training examples, we get an objective (cost) function Q. Gradient descent is used to maximize/minimize the Q function: iterating training examples, and gradually adjusting Q's parameters based on derivatives. Intuitively, each iteration on training data helps to move Q to a local optima.
c) For traditional gradient descent approach, parameters are only updated after all training examples are processed once. This could be a problem when processing high-dimensional or large amount of data. However, stochastic gradient descent randomly process training examples and update parameters for each example.
d) Traditional gradient descent is not as efficient as stochastic gradient descent, but it would always converge on a global optima since all parameters are adjusted according to training examples.

1.2)

There are often multiple layers in DNN; to update parameters of the DNN, we often need to multiply derivatives of following layers. While sigmoid tend to generate smaller gradient when input is big. That means, it would be difficult to update parameters based on gradient of sigmoid function. (check discussions here [1])

2.1)

State: {A, T, C, G}

Initial Probability: P(A) = 7/30 = 0.233; P(T) = 8/30 = 0.267; P(C) = 9/30 = 0.300; P(G) = 6/30 = 0.200

Transition Probabilities:

	A	T	C	G
A	0.143 (1/7)	0.714 (5/7)	0 (0/7)	0.143 (1/7)
T	0.143 (1/7)	0 (0/7)	0.857 (6/7)	0 (0/7)
C	0.25 (2/8)	0.125 (1/8)	0.125 (1/8)	0.5 (4/8)
G	0 (0/3)	0.67 (2/3)	0.33 (1/3)	0 (0/3)

2.2) P(ATCG) = P(A) P(T|A) P(C|T) P(G|C) = 0.233 * 0.714 * 0.857 * 0.5 = 0.071

3.1) CpG island: there are many DNA regions that have a high frequency of CG sites. Predicting such CpG island is importance, for instance, they are related with the start of the gene (promoter). One problem with CpG island is how to check whether a certain DNA sequence comes from a CpG island. It could be solved by a Markov model: given a collection of sequences from CpG island and a collection from non-CpG island, we can build two Markov models: CpG model (M+) and non-CpG model (M-). For a new sequence S, we can check whether S is a CpG island by comparing P(S | M+) and P(S | M-).

3.2) Motif finding: motif refers to an important nucleotide/amino-acid sequence pattern. Finding motifs is a significant task, for instance, it may indicate potential binding sites. An approach to solve the problem is EM algorithm: finding all k-mers from the a collection of sequences; EM algorithm could help to derive two models (motif model M and background model B) and maximize the probability of assigning k-mers to these two models.

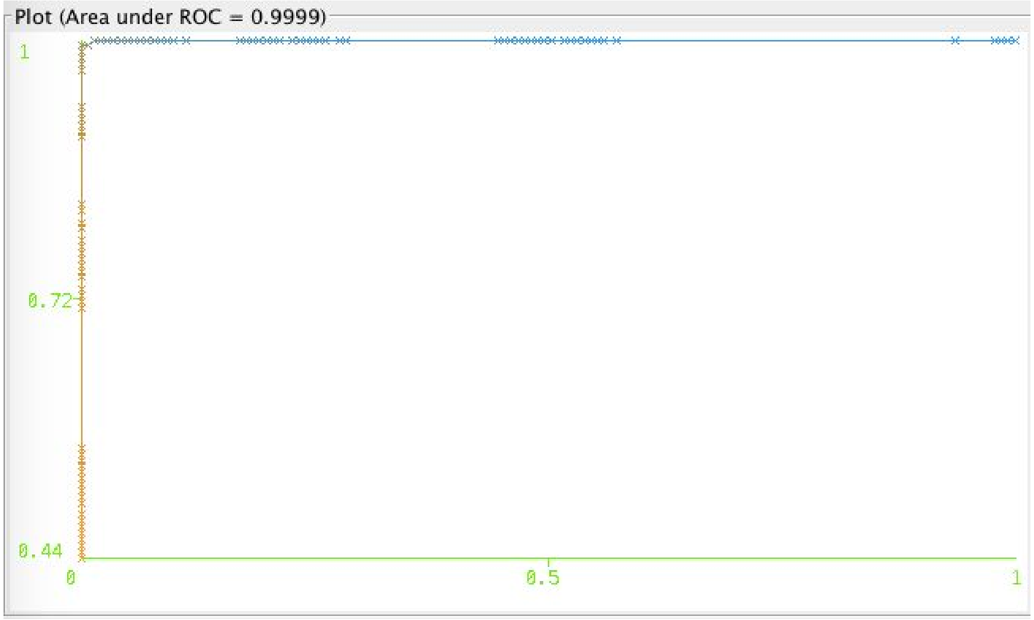
3.3) Predicting DNA- and RNA-binding protein specificity: given a sequence s, determine whether it's related with a binding site. As discussed in this course, DNN could be used to come up with a binding score for each sequence, as conducted in DeepBind. Essentially, a DNN is trained from a set of sequences, which are labeled a binding score experimentally.

4) Specifically,
   a) Since this dataset is of arff format, we use ML implementations of "weka.jar";
   b) The ML approach is random forest;
   c) We do 10-fold cross-validation with weka-supported API. Basically, the program randomly splits dataset into 10 parts; and each time 9 parts are used to do training and the left part is used to do validation;
   d) We evaluate the random forest model with the metrics provided by weka: confusion matrix, f-measure. And we also implemented ROC curve.
   e) As noted in RF.readme file, the program is compiled by "javac -cp ".:weka.jar" RF.java" and executed with "java -cp ".:weka.jar" RF";
   f) Our random forest is composed of 10 trees and we do 10-fold cross validation. The result is as follows:

	NO (classified as)	YES (classified as)
NO	150	1
YES	0	141

The Precision, Recall, F-Measure are as follows:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	**Class**
0.993	0	1	0.993	0.997	1	**NO**
1	0.007	0.993	1	0.996	1	**YES**
0.997	0.003	0.997	0.997	0.997	1	**Weighted average**



[1] What are the advantages of ReLU over sigmoid function in deep neural networks? https://stats.stackexchange.com/questions/126238/what-are-the-advantages-of-relu-over-sigmoid-function-in-deep-neural-networks?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa