2.

1. I chose decision trees and Support Vector Machines. I chose decision trees because I had familiarity with them after we went over them in class. I read about support vector machines from assignment 1 and became interested in them and it seemed like an algorithm with a lot of parameters I could change. They are also both good at classification problems. SVMs categorize text into separate groups and decision trees tell you which group a piece of data falls into
2.

Reported information	Test type	k-NN	Logistic regression	Decision Trees	Support Vector Machine
Accuracy with default parameters	Entire dataset	92.37%	91.30%	91.39%	90.71%
	50%	90.77%	90.79%	90.58%	90.83%
	Cross-validation	90.85%	91.19%	91.29%	90.71%
Accuracy with best parameters	Entire dataset	100%	91.30%	99.13%	90.71%
	50%	91.29%	90.79%	96.74%	90.83%
	Cross-validation	91.21%	91.19%	97.15%	90.71%
List names of parameters		K, weighted vote, measure types, mixed measure	Solver: Auto standardized	Criterion, maximal depth, apply pruning, confidence, apply prepruning, minimal gain, minimal leaf size	Selected best results already
List selected best values of parameters (in the same order as in the list of names)	Entire dataset	1, weighted vote, measure type: mixed measures, mixed measure: mixedEuclidianDistance	Solver: Auto Standardized	Criterion: gini_index	
	50%	21, weighted vote, measure type: mixed measures, mixed measure: mixedEuclidianDistance	Solver: Auto Standardized	Criterion: gini_index	
	Cross-validation	21, weighted vote, measure type: mixed measures, mixed measure: mixedEuclidianDistance	Solver: Auto standardized	Criterion: gini_index	

3. K-NN produced 100% but I think that that's because it is overfit with the data it was given and the extremely low value for k which was 1 in this case. Also considering the fact that there are a lot of other features making this dataset noisy the 100% is definitely not a good result.
4. This predictor would be better suited to predict proteins that do not interact with nucleic acids because the level of precision is very high for predicting no values and the recall is very high as well.

accuracy: 90.85% +/- 0.37% (micro average: 90.85%)

	true No	true Yes	class precision
pred. No	7723	669	92.03%
pred. Yes	136	267	66.25%
class recall	98.27%	28.53%	

5. Yes it is an improvement because this way we can actually predict values. It's also able to be exposed to new data and still produce accurate results.