CMSC 435
Submitted:
Group: *Chain Gang*

Protein Classification Project

Submitted by:
Michael Fitzgerald, Jacob Unterman, Michael Poblacion, Gavin Alberghini

1) **Description of the design for our predictive models**
    a) **Pre-Processing Methods**

        i) **CharCount Method**

        We preprocessed the data into a count of the number of occurrences for each amino acid in each sequence.

        ii) **SubString Matching**

        iii) **N-grams Method**

        iv) **PROFEAT Method**

    b) **Model Selections**
        i) **kNN**
        ii) **SVM**
        iii) **Neural Net**
        iv) **Random Forest**
        v) **CHAID**
        Parameters used:
        vi) **Decision Tree**
        Parameters used:
            (1) Criterion: Gain Ratio
            (2) Maximal Depth: 20
            (3) Pruning: Yes
            (4) Confidence: 0.2
            (5) Pre Pruning: No

    c) **Design Selections**
        i) **Design 1**
        ii) **Design 2**
        iii) **Design 3**

iv) Best Design

2) Results of predictive models

Outcome	Quality Measure	Baseline Result	Design 1	Design 2	Design 3	Best Design
DNA	Sensitivity	6.9	18.84			
	Specificity	99.3	97.84			
	Predictive ACC	95.2	97.53			
	MCC	0.132	0.07			
RNA	Sensitivity	39.6	44.9			
	Specificity	98.9	97.6			
	Predictive ACC	95.3	96.86			
	MCC	0.501	0.29			
DRNA	Sensitivity	4.5	28.57			
	Specificity	100	99.88			
	Predictive ACC	99.7	99.86			
	MCC	0.122	0.16			
nonDRNA	Sensitivity	98.6	90.91			
	Specificity	29.8	89.02			
	Predictive ACC	91.3	90.64			
	MCC	0.428	0.69			
Average MCC		0.265	0.3			
Accuracy		90.8	89.02			

3) Conclusions