Protein Interaction Project Final Report

Virginia Commonwealth University

CMSC435 - Introduction to Data Science

November 22, 2024

Group 9 - Model Masters

Hoang Le, Sanie Fragata, Amaka Odidika, William Hunter, Stephen Nocera,

Sejal Pandeya, Femi Adepoju

Design Description

Features Generated & Selected From Input Sequences:

Basic Amino Acid Properties Features:

- Amino Acid Frequency (20 features):
    - Captures composition based on all 20 amino acids
- Alphatic (1 feature):
    - Measures open side chains to represent hydrophobicity
- Aromatic (1 feature):
    - Represents cyclic side chains with double bonds
- Charge Distribution (3 features):
    - Positive, negative, and uncharged properties

Compositional Features:

- CTDC - Composition:
    - Composition of various physicochemical properties of protein sequences including:
        - Hydrophobicity (21 features)
        - Van der Waals Volume (3 features)
        - Polarity (3 features)
        - Polarizability (3 features)
        - Charge (3 features)
        - Secondary Structure (3 features)
        - Solvent Accessibility (3 features)

Dipeptide-based Features:

- Dipeptide Composition (400 features):
    - Captures local order information through adjacent amino acid pair frequencies
- Grouped Dipeptide Composition (25 features):
    - Calculates the frequency of each consecutive amino acid group pair in the sequence based on physicochemical properties

Total Features: 489

Feature Data Processing & Generation:

1. Written Python script for amino acid frequency features
2. Written Python script to convert protein sequences txt file into FASTA format for iLearnPlus.
3. Utilized ML platform iLearnPlus for basic amino acid properties, protein sequence composition, and dipeptide features generation.
4. Combined all generated features in Excel, converted to a CSV file.

We have considered and tested multiple classification algorithms including: Decision Trees, Random Forest, SVM, k-NN, Neural Net, Gradient Boosted Trees, Generalized Linear Model, and LDA.

The 3 Algorithms & Parameters We Chose:

SVM Model (Design 1)

- Performance / Initial Design: Achieved 60.4% accuracy with 0.189 MCC score. While overall accuracy is poor, the model excels specifically at predicting DRNA and DNA classifications, making it the best performing algorithm for DNA/DRNA predictions. Runtime is approximately 1 hour, making it the slowest of the three models.
    - Parameters Chosen: Max Iterations = 10000, Balance Cost

k-NN Model (Design 2)

- Performance & Efficiency: Delivers second best performance with 91.32% accuracy and 0.297 MCC score. Excels at nonDRNA and RNA predictions while running efficiently (10-15 seconds). Key advantages include simple implementation, fewer parameters to tune, and balance of speed vs accuracy. Main limitation is the inability to predict DRNA.
    - Parameters Chosen: K = 8, Weighted, Numerical Measures → Manhattan Distance

Gradient Boosted Trees Model (Design 3, Best)

- Overall Best Performer: Shows the best performance we could get through trial and error with 91.76% accuracy and 0.328 MCC score. While slower than K-NN with 7-8 minute runtime, it provides better RNA predictions and highest overall accuracy. Like K-NN, it cannot predict DRNA, but the improved accuracy makes it the optimal choice when runtime isn't a critical priority.
    - Parameters Chosen: Number of Trees = 100, Maximal Depth = 10, Number of Bins = 5, Learning Rate = 0.02

Results

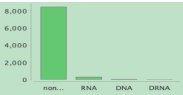
Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	6.9	58.3	10.2	9.7	9.7
	Specificity	99.3	78.8	99.7	99.9	99.9
	Accuracy	95.2	77.9	95.8	95.9	95.9
	MCC	**0.132**	**0.182**	**0.246**	**0.269**	**0.269**
RNA	*Sensitivity*	39.6	50.7	32.7	43.6	43.6
	Specificity	98.9	85.4	99.6	99.2	99.2
	Accuracy	95.3	83.3	95.6	95.9	95.9
	MCC	**0.501**	**0.228**	**0.505**	**0.563**	**0.563**
DRNA	*Sensitivity*	4.5	18.2	0.0	0.0	0.0
	Specificity	100.0	95.8	100.0	100.0	100.0
	Accuracy	99.7	95.5	99.7	99.7	99.7
	MCC	**0.122**	**0.034**	**-0.001**	**0.000**	**0.000**
nonDRNA	*Sensitivity*	98.6	61.3	91.8	99.3	99.3
	Specificity	29.8	87.9	86.0	30.7	30.7
	Accuracy	91.3	64.2	91.6	92.0	92.0
	MCC	**0.428**	**0.306**	**0.436**	**0.478**	**0.478**
averageMCC		**0.296**	**0.189**	**0.297**	**0.328**	**0.328**
accuracy4labels		90.8	60.4	91.3	91.8	91.8

Best Design Confusion Matrix

accuracy: 91.76% +/- 0.30% (micro average: 91.76%)

	true nonDRNA	true RNA	true DNA	true DRNA	class precision
pred. nonDRNA	7804	293	335	21	92.32%
pred. RNA	48	228	18	0	77.55%
pred. DNA	7	2	38	1	79.17%
pred. DRNA	0	0	0	0	0.00%
class recall	99.30%	43.59%	9.72%	0.00%	

Blind Test Data Set Predictions Summary

Prediction **prediction(Class)**	Polynominal	0	 Open visualizations	Least DRNA (0)	Most nonDRNA (8459)	Values nonDRNA (8459), RNA (298), DNA (37), DRNA (0) Details...
Confidence_nonDRNA **confidence(nonDRNA)**	Real	0	Min 0.025	Max 0.944	Average 0.863	
Confidence_RNA **confidence(RNA)**	Real	0	Min 0.020	Max 0.924	Average 0.065	
Confidence_DNA **confidence(DNA)**	Real	0	Min 0.018	Max 0.881	Average 0.047	
Confidence_DRNA **confidence(DRNA)**	Real	0	Min 0.016	Max 0.174	Average 0.026	

The blind test predictions offered similar results to our best 5-fold cross validation result (design 3). It is unable to predict any DRNA, with a very low average confidence core of 0.026. Although the model is very confident in predicting the majority class, nonDRNA, it is average/poor at predicting the other classes. Before we ran the blind test, we predicted that the model would not be able to predict any DRNA, and we were proven correct as shown above.

Conclusions

Quality of Results:

- Our best model (Gradient Boosted Trees) achieved a 0.032 increase in averageMCC from the baseline of 0.296 to 0.328. This is due to a good improvement in DNA classification and a noticeable increase in RNA and nonDRNA classification. However, the model is very poor at DRNA classification with a large decrease in MCC score from the DRNA baseline of 0.122 to 0.0.
- The results also showed that our best model has similar sensitivity, specificity, accuracy scores across the board for all 4 classes when compared to the baseline.
    - However, our model has a sensitivity score of 0 for DRNA, which is an indicator of the model failing to classify any true positives for the class. This is definitely worse than the baseline sensitivity DRNA score of 4.5, as it cannot detect the class at all.

Experience with Project:

- Developing a predictive model by extracting features from the raw dataset was the most challenging aspect due to the need to balance relevance and complexity. We wanted to make sure that there weren't too many features, not too similar, and redundant, which can cause overfitting. Completing this project was rewarding, as it provided valuable hands-on practice in processing, generating, and analyzing real-world data. This project gave us a greater understanding of the Knowledge Discovery model in data science because instead of simply learning about it we also had the opportunity to apply it first hand. We also learned the importance of collaboration because we had to schedule meetings once a week to ensure everyone was on the same page. Ultimately, this project helped us build skills that will benefit our future careers in data related fields.

Advantages & Disadvantages of Our Method:

- Advantages
    - Improved averageMCC (0.328) and accuracy (91.8) compared to baseline result.
    - Relatively fast runtime, 7-8 minutes.
    - Great nonDRNA (majority class) classification performance.
- Disadvantages
    - The main disadvantage of our method is that the model is unable to classify and predict DRNA consistently.
    - Average to poor classification for DNA and RNA.
    - Choosing the correct amount of trees is imperative to avoid overfitting, however testing is difficult due to heavy computation.

Final Thoughts:

Overall, we think that the results produced from our best model are good but can be improved. Compared to our other models, Gradient Boosted Trees provided the best accuracy and MCC scores with a reasonable runtime. Gradient Boosted Trees is a better model than our previous two models simply because it is ever slightly more accurate than K-NN and despite not having the same potential to predict DRNA and DNA as SVM, the dataset still contains more nonDRNA and RNA which makes the Gradient Boosted Trees model more valuable in correctly predicting sequence classification.

There is a possibility that our model may struggle with a dataset with more DNA and DRNA as the majority class, our dataset may be overfitting due to the features we have extracted that heavily bias the majority class. One area of improvement would be doing more research focused on gaining a deeper understanding of the biology aspect of the features available to us. This can help us identify the features with the greatest impact on DNA and DRNA, which can lead to better classification and prediction for those poor performing classes from our model.