VCU College of Engineering

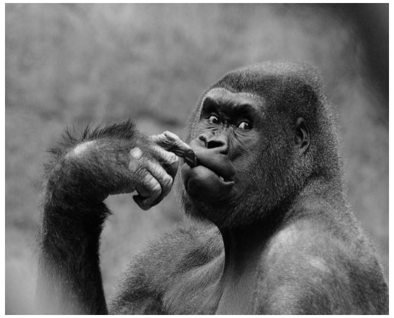
Team #01

Brianna Barrett, Eleanor Estwick, Cyaira Hughes, Savannah Nelson, Menelaos Sofroniou

Quality of Results

- Our model is biased toward classifying a given sequence as nonDRNA (high sensitivity 99.4, low specificity 21.0)
- This bias is reflected in the total accuracy of all labels
- The high total accuracy (90.98%) does not necessarily correlate to an accurate model
- The low average MCC (0.27) suggests a weak, positive relationship between predicted and actual classes

Design Motivation



Our three main motivations include:

★ A simple, easy to implement algorithm

★ A model that could produce results quickly

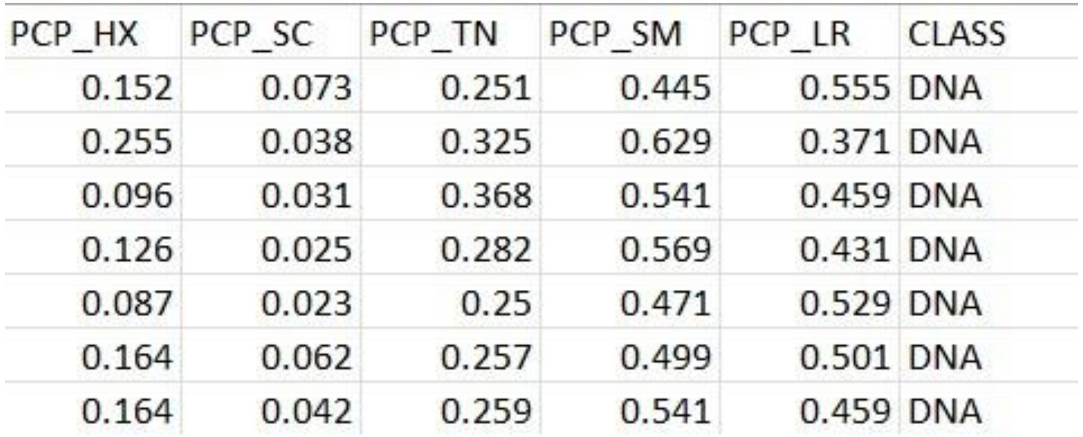
★ A design suitable for our data

Design Description

- ❑ **Feature Selection**
- ❑ **Design Creation**

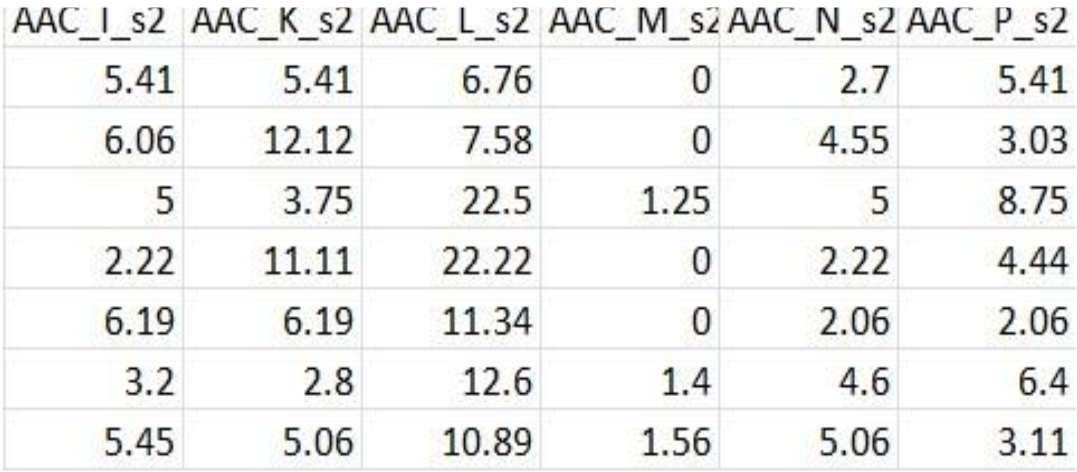
Feature Selection Sets Part (½)

- Physico-Chemical Properties Composition
  - Intrinsic physical and chemical characteristics.
  - Compositions of properties were selected from the whole sequence, the sequence split in half, and cutting 5 C and N terminals from each sequence.



PCP_HX	PCP_SC	PCP_TN	PCP_SM	PCP_LR	CLASS
0.152	0.073	0.251	0.445	0.555	DNA
0.255	0.038	0.325	0.629	0.371	DNA
0.096	0.031	0.368	0.541	0.459	DNA
0.126	0.025	0.282	0.569	0.431	DNA
0.087	0.023	0.25	0.471	0.529	DNA
0.164	0.062	0.257	0.499	0.501	DNA
0.164	0.042	0.259	0.541	0.459	DNA

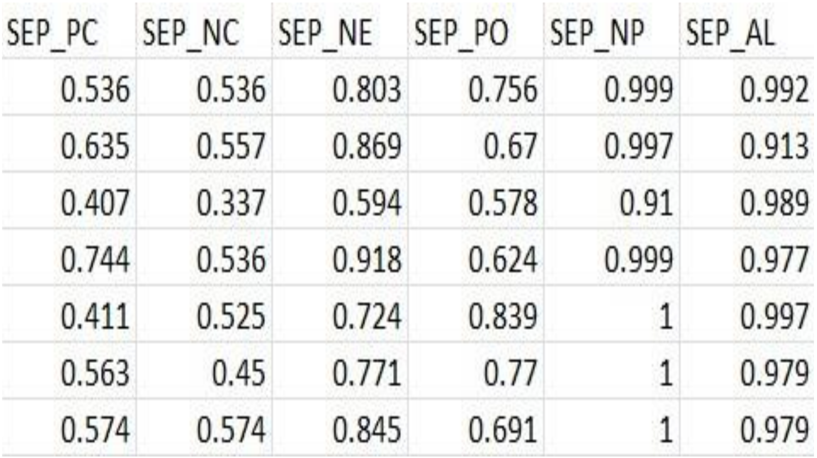
- Simple Amino Acid Composition
  - Simple calculation the percentage of each amino acid found in a sequence.
  - Amino acids sequences were selected from the entire sequence, the sequence split, and 5 C and N terminals cut from each sequence.



AAC_I_s2	AAC_K_s2	AAC_L_s2	AAC_M_s2	AAC_N_s2	AAC_P_s2
5.41	5.41	6.76	0	2.7	5.41
6.06	12.12	7.58	0	4.55	3.03
5	3.75	22.5	1.25	5	8.75
2.22	11.11	22.22	0	2.22	4.44
6.19	6.19	11.34	0	2.06	2.06
3.2	2.8	12.6	1.4	4.6	6.4
5.45	5.06	10.89	1.56	5.06	3.11

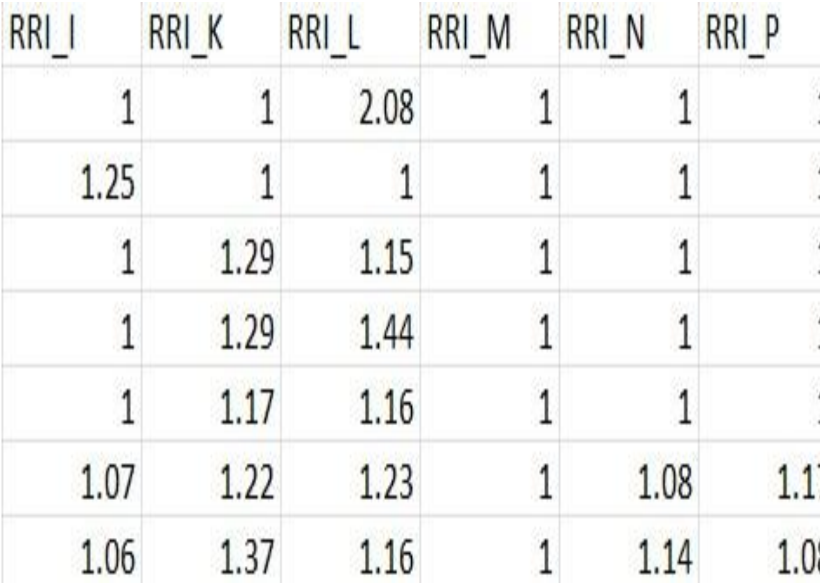
Feature Selection Sets Part (2/2)

- <u>Shannon Entropy of Physico-Chemical Properties</u>
  - Measures the degree of randomness in a set of data
    - 0 indicates that this value is certain to appear within a particular set
    - 1 indicates complete randomness.
  - Shannon entropy physico-chemical properties in each amino acid sequence were selected from the whole sequence.



SEP_PC	SEP_NC	SEP_NE	SEP_PO	SEP_NP	SEP_AL
0.536	0.536	0.803	0.756	0.999	0.992
0.635	0.557	0.869	0.67	0.997	0.913
0.407	0.337	0.594	0.578	0.91	0.989
0.744	0.536	0.918	0.624	0.999	0.977
0.411	0.525	0.724	0.839	1	0.997
0.563	0.45	0.771	0.77	1	0.979
0.574	0.574	0.845	0.691	1	0.979

- <u>Repetitive Residue Information of Simple Amino Acids</u>
  - RRI values lie between 0 to the length of the sequence
    - 0 means that residue is not present
    - RRI equal to length of sequence means that all residues are same
      - Higher the value of RRI, higher the repetition of a particular residue.
  - The composition of RRI in each amino acid sequence was selected from the whole sequence.



RRI_I	RRI_K	RRI_L	RRI_M	RRI_N	RRI_P
1	1	2.08	1	1	[illegible]
1.25	1	1	1	1	[illegible]
1	1.29	1.15	1	1	[illegible]
1	1.29	1.44	1	1	[illegible]
1	1.17	1.16	1	1	[illegible]
1.07	1.22	1.23	1	1.08	1.1
1.06	1.37	1.16	1	1.14	1.0

VCU College of Engineering

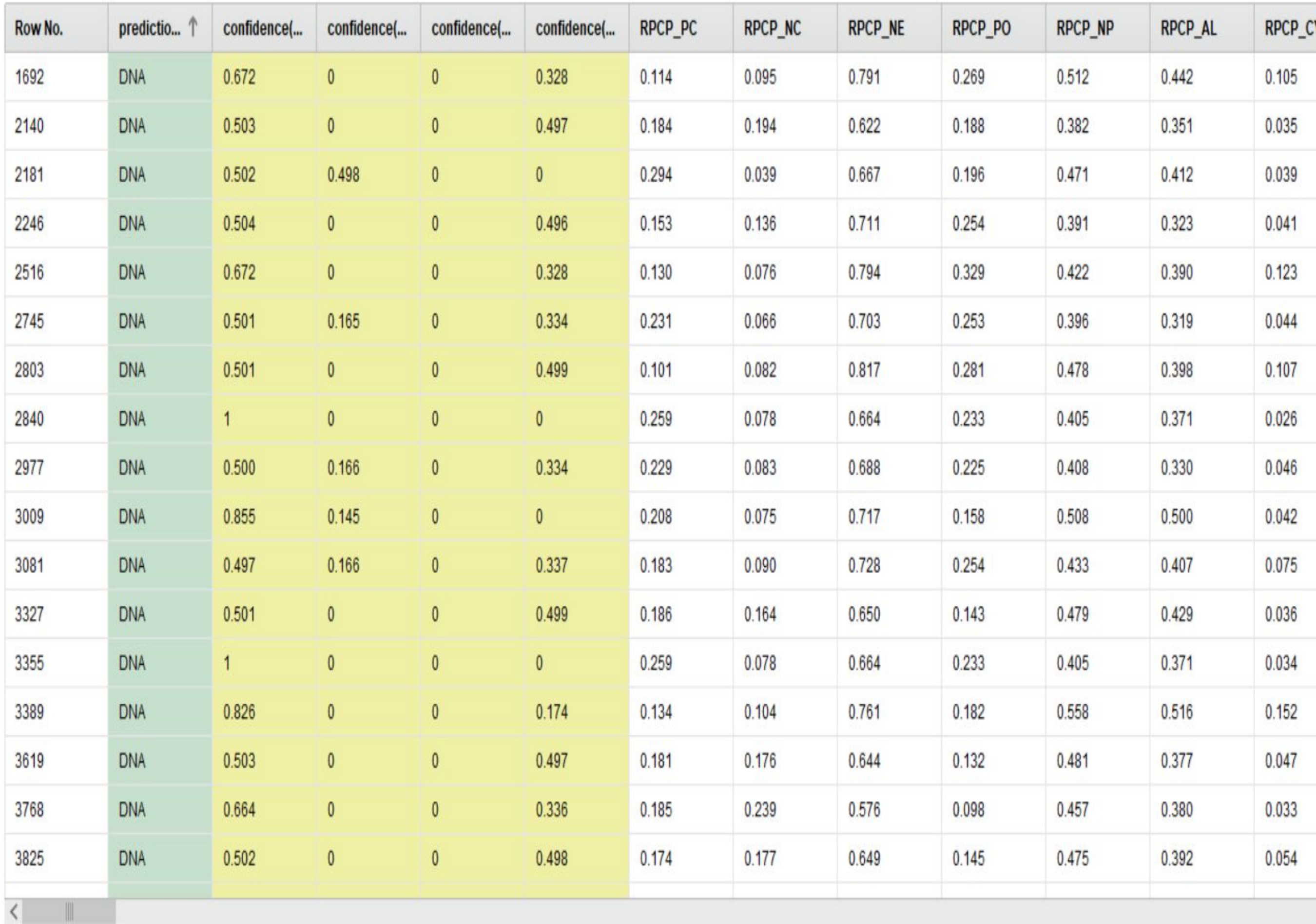
Design Creation

K-Nearest Neighbors (kNN)

- ★ uses ‘feature similarity’
- ★ supports multi-label classification

1. The distance is calculated
2. The closest k data points are chosen
3. These neighbors help calculate the final prediction of the new points

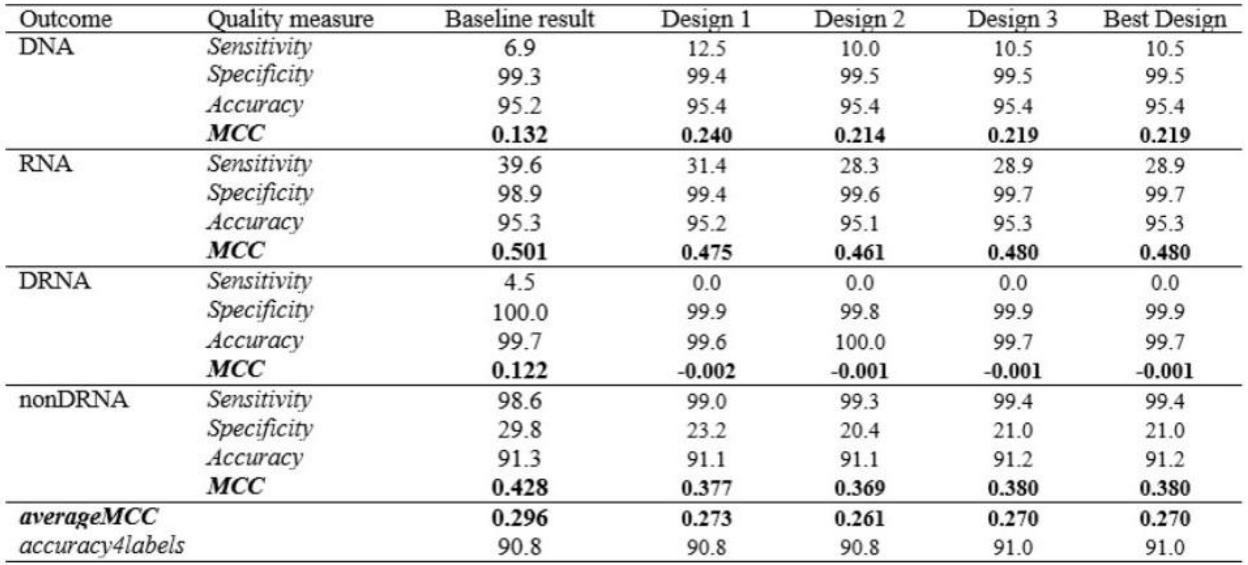
Model production on the blind dataset



Row No.	predictio... ↑	confidence(...	confidence(...	confidence(...	confidence(...	RPCP_PC	RPCP_NC	RPCP_NE	RPCP_PO	RPCP_NP	RPCP_AL	RPCP_C
1692	DNA	0.672	0	0	0.328	0.114	0.095	0.791	0.269	0.512	0.442	0.105
2140	DNA	0.503	0	0	0.497	0.184	0.194	0.622	0.188	0.382	0.351	0.035
2181	DNA	0.502	0.498	0	0	0.294	0.039	0.667	0.196	0.471	0.412	0.039
2246	DNA	0.504	0	0	0.496	0.153	0.136	0.711	0.254	0.391	0.323	0.041
2516	DNA	0.672	0	0	0.328	0.130	0.076	0.794	0.329	0.422	0.390	0.123
2745	DNA	0.501	0.165	0	0.334	0.231	0.066	0.703	0.253	0.396	0.319	0.044
2803	DNA	0.501	0	0	0.499	0.101	0.082	0.817	0.281	0.478	0.398	0.107
2840	DNA	1	0	0	0	0.259	0.078	0.664	0.233	0.405	0.371	0.026
2977	DNA	0.500	0.166	0	0.334	0.229	0.083	0.688	0.225	0.408	0.330	0.046
3009	DNA	0.855	0.145	0	0	0.208	0.075	0.717	0.158	0.508	0.500	0.042
3081	DNA	0.497	0.166	0	0.337	0.183	0.090	0.728	0.254	0.433	0.407	0.075
3327	DNA	0.501	0	0	0.499	0.186	0.164	0.650	0.143	0.479	0.429	0.036
3355	DNA	1	0	0	0	0.259	0.078	0.664	0.233	0.405	0.371	0.034
3389	DNA	0.826	0	0	0.174	0.134	0.104	0.761	0.182	0.558	0.516	0.152
3619	DNA	0.503	0	0	0.497	0.181	0.176	0.644	0.132	0.481	0.377	0.047
3768	DNA	0.664	0	0	0.336	0.185	0.239	0.576	0.098	0.457	0.380	0.033
3825	DNA	0.502	0	0	0.498	0.174	0.177	0.649	0.145	0.475	0.392	0.054

Quality Comparison

Table 1



Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	6.9	12.5	10.0	10.5	10.5
	Specificity	99.3	99.4	99.5	99.5	99.5
	Accuracy	95.2	95.4	95.4	95.4	95.4
	MCC	**0.132**	**0.240**	**0.214**	**0.219**	**0.219**
RNA	*Sensitivity*	39.6	31.4	28.3	28.9	28.9
	Specificity	98.9	99.4	99.6	99.7	99.7
	Accuracy	95.3	95.2	95.1	95.3	95.3
	MCC	**0.501**	**0.475**	**0.461**	**0.480**	**0.480**
DRNA	*Sensitivity*	4.5	0.0	0.0	0.0	0.0
	Specificity	100.0	99.9	99.8	99.9	99.9
	Accuracy	99.7	99.6	100.0	99.7	99.7
	MCC	**0.122**	**-0.002**	**-0.001**	**-0.001**	**-0.001**
nonDRNA	*Sensitivity*	98.6	99.0	99.3	99.4	99.4
	Specificity	29.8	23.2	20.4	21.0	21.0
	Accuracy	91.3	91.1	91.1	91.2	91.2
	MCC	**0.428**	**0.377**	**0.369**	**0.380**	**0.380**
averageMCC		**0.296**	**0.273**	**0.261**	**0.270**	**0.270**
accuracy4labels		90.8	90.8	90.8	91.0	91.0

Quality Comparison

- ➔ **DNA** - improved specificity, sensitivity weakened but closer to baseline
- ➔ **RNA** - improved specificity and accuracy, weakened sensitivity
- ➔ **DRNA** - outcomes consistent across designs
- ➔ **nonDRNA** - improved accuracy and sensitivity, weakened specificity

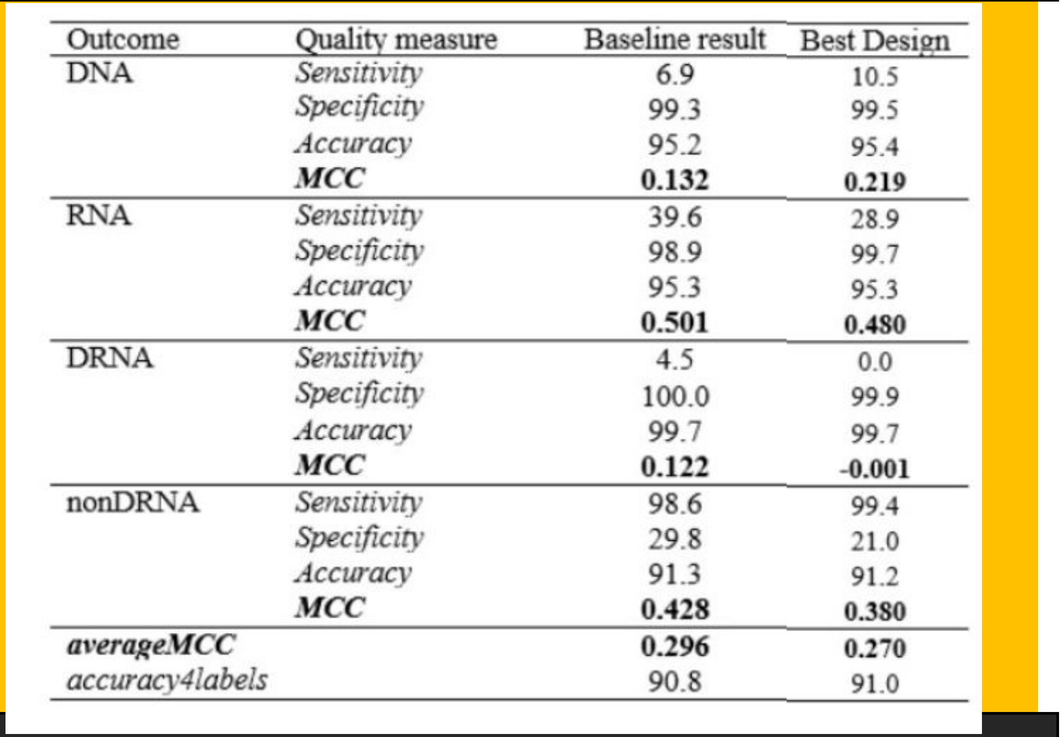
VCU College of Engineering

Conclusions

1. Quality of Results
2. Baseline Comparison
3. Advantages/Disadvantages
4. Experience

Baseline Comparison

Table 1



Outcome	Quality measure	Baseline result	Best Design
DNA	*Sensitivity*	6.9	10.5
	Specificity	99.3	99.5
	Accuracy	95.2	95.4
	MCC	**0.132**	**0.219**
RNA	*Sensitivity*	39.6	28.9
	Specificity	98.9	99.7
	Accuracy	95.3	95.3
	MCC	**0.501**	**0.480**
DRNA	*Sensitivity*	4.5	0.0
	Specificity	100.0	99.9
	Accuracy	99.7	99.7
	MCC	**0.122**	**-0.001**
nonDRNA	*Sensitivity*	98.6	99.4
	Specificity	29.8	21.0
	Accuracy	91.3	91.2
	MCC	**0.428**	**0.380**
averageMCC		**0.296**	**0.270**
accuracy4labels		90.8	91.0

Comparison

- Biggest differences were in Sensitivity
  - RNA: -10.7
  - DRNA: -4.5
  - DNA: +3.6
  - nonDRNA: +0.8
- Differences for Specificity
  - nonDRNA: -8.8
  - RNA: +0.8
  - DNA: +0.2
  - DRNA: -0.1

VCU College of Engineering

Advantages and Disadvantages of Model

Advantages

- Improved accuracy of predicting DNA
- Provided closest MCC to the baseline for RNA, nonDRNA, and nearly DNA

Disadvantages

- Model could not capture DRNA strains

Experience

- Positive learning experience
- Rotated group member responsibilities
- Consistent communication
- Constant collaboration

Questions?