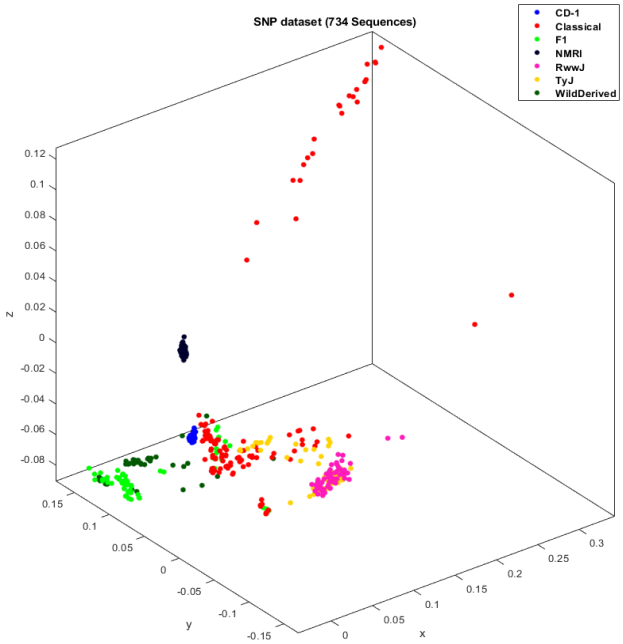
Results:
734 sequences (4932290 bp each).
Clusters: CD-1 (100 sequences), Classical (126 sequences), F1 (59 sequences), NMRI (287 sequences), RwwJ (57 sequences), TyJ (55 sequences), WildDerived (50 sequences)

Method: read the sequences as given (numerical series with "-1, 0, -1, 2"), applied Fourier transform to get magnitude spectra, Pearson correlation coefficient to get pairwise distance matrix, and classification algorithms.

ClassifierModel	Accuracy (%)
'LinearDiscriminant'	94
'LinearSVM'	90.7
'QuadraticSVM'	**94.1**
'FineKNN'	93.7
'SubspaceDiscriminant'	92.8
'SubspaceKNN'	92.9
'AverageAccuracy'	**93**



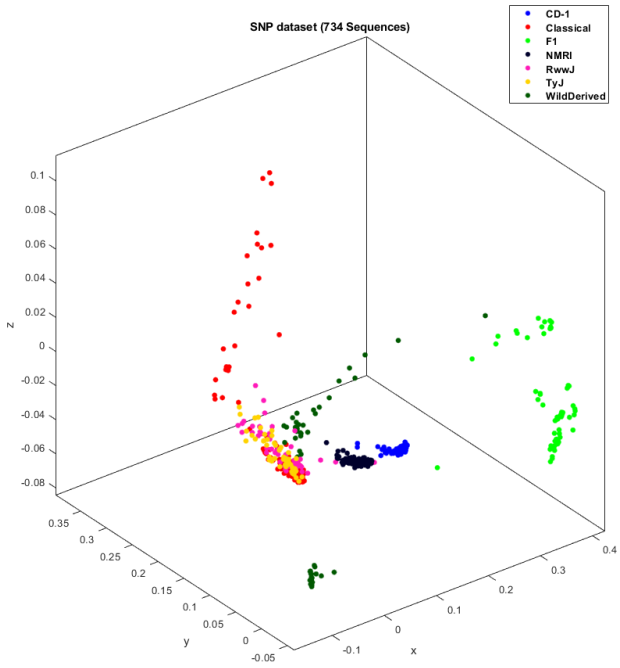
Confusion matrix (Quadratic SVM)

	CD-1	Classical	F1	NMRI	RwwJ	TyJ	WildDerived
CD-1	98	2	0	0	0	0	0
Classical	0	119	2	0	1	4	0
F1	0	4	45	0	0	0	10
NMRI	0	0	0	287	0	0	0
RwwJ	0	1	0	0	50	6	0
TyJ	0	3	0	0	6	46	0
WildDerived	0	2	2	0	0	0	46

Same dataset;
Method2: In every sequence, I replaced “-1, 0, 1, 2” with “T, C, A, G” (respectively) so that CGR can be generated.

CGR (k=6)

ClassifierModel	**Accuracy (%)**
'LinearDiscriminant'	**94.1**
'LinearSVM'	88.1
'QuadraticSVM'	90.7
'FineKNN'	87.9
'SubspaceDiscriminant'	93.1
'SubspaceKNN'	88.1
'AverageAccuracy'	**90.3**

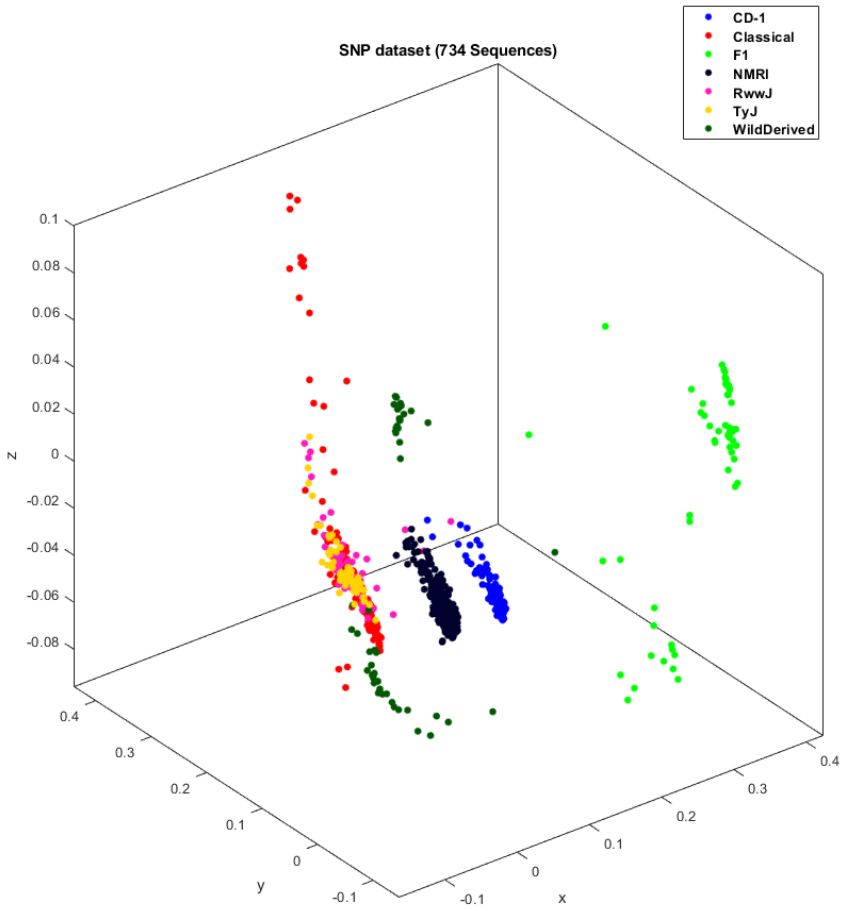


Confusion matrix (Linear Discriminant)

	CD-1	Classical	F1	NMRI	RwwJ	TyJ	WildDerived
CD-1	100	0	0	0	0	0	0
Classical	0	117	0	0	1	7	1
F1	0	0	58	0	0	0	1
NMRI	0	0	0	287	0	0	0
RwwJ	0	1	0	1	42	13	0
TyJ	0	5	0	0	11	39	0
WildDerived	0	1	1	0	0	0	48

CGR (k=9)

ClassifierModel	Accuracy (%)
'LinearDiscriminant'	**95.4**
'LinearSVM'	86.1
'QuadraticSVM'	90.7
'FineKNN'	88.4
'SubspaceDiscriminant'	94.1
'SubspaceKNN'	88.6
'AverageAccuracy'	**90.6**



Confusion matrix (Linear Discriminant)

	CD-1	Classical	F1	NMRI	RwwJ	TyJ	WildDerived
CD-1	100	0	0	0	0	0	0
Classical	0	118	0	0	3	5	0
F1	0	0	58	0	0	0	1
NMRI	0	0	0	287	0	0	0
RwwJ	0	1	0	0	46	10	0
TyJ	0	2	0	0	10	43	0
WildDerived	0	1	1	0	0	0	48

Things to try next:

1) Increasing value of k seems to improve accuracy, so we can try that.
2) I picked A, C, G, T randomly for "1, 0, -1, 2" (keeping complementary property in mind). We can try assigning letters in other order to test if something changes.
3) We have test set of 800 sequences that I want to test using these trained classifiers (will have to code a bit to clean that data).

Note: Please note that reading and cleaning (pre-processing) the data is time consuming (will take ~ 30 minutes with provided script on an average PC). I have provided ".mat" file containing already cleaned data (fixed typos in provided excel sheets, extracted sequences and stored into MATLAB variables for easy access) that can be loaded to complete the experiment under 1 minute.

Two scripts are provided:
genoTest.m - This file to be used to run the experiments with provided numerical sequences.
genoTestCGR.m – This file to be used for CGR representation.

Other resources to added to the same MATLAB directory:
Provided "SNP.mat" file to run the experiment under 1 minute.
Provided ".csv" files in case loading fails from the ".mat" file.
Most of the MLDSP scripts (can be obtained online).