In this classification project, I used Vector Space Model, specifically "Term Frequency – Inverse Document Frequency" (TF_IDF) and cosine similarity analyses to classify the sequences. Since there are training data provided, this is a supervised learning classification case.

Term frequency measures how important a word may be in a document by counting its occurrences. Inverse Document Frequency decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents. This can be combined with term frequency to calculate TF-IDF, which is the frequency of a term adjusted for how rarely it is used.

Given a sentence, we can transform it into a numeric vector based on the frequency of terms exist in the sentence. Then cosine of the angle formed between two vectors will indicate how closely related the two documents are. Cosine ranges from -1 to 1, but we are only interested in the range of 0 to 1. If the cosine is close to 1, this indicates the two vectors and hence the two documents are very similar to each other. And if the cosine is close to 0, this indicates that the two vectors are perpendicular to each other (dot product is 0), and thus the two documents are different from each other.

I have given careful step-by-step details on how to use the model in the python file "FinalProj.py". I used Python 3.5. The main module in this project is "sklearn", providing functions for the TF-IDF model and cosine similarity analysis.

There are two features that I look at: k-mers and repeats. The selection for these features locates from line 73 to line 85 of the python file. Specifically, to choose k-mers feature, uncomment line 80 and comment out line 83. If choosing repeats feature, uncomment line 83 and comment out line 80.

For k-mers feature, when I chose k upto 3 (k_max = 3), line 48, I got the convergent result shown below, meaning that the same result will be obtained with k_max > 3. For repeats feature, I also got the same result for the 200 sequences testing data set, with minimum word = 4 (line 49), and minimum repeat = 2 (line 50). In both cases, cross-validation always returns perfect classification. The result for the testing data set is: For Type 1: there are 100 sequences, from Sequence0 to Sequence99; For Type 2: no sequences; For Type 3: 50 sequences, from Sequence100 to Sequence 149; For Type 4: 50 sequences, from Sequence150 to Sequence199. I got results that are very close to the convergent result when using k_max = 2 (for k-mers feature); and minimum word = 1 to 3 (for repeats feature). This shows that the TF-IDF and cosine similarity analyses for the test data converge very quickly.

In my opinion, the repeats feature is the better feature of the two. It captures more specific and (maybe) more important patterns in a sequence. We do not have to look at every k-mers but only those that repeat themselves, which can be biologically meaningful if we are dealing with real biological sequences. The repeats feature thus may represent a "parsimonious" model.

=====TRUE vs PREDICTION VALUES ON CROSS-VALIDATION DATA =====

True Type 1:
['Sequence11' 'Sequence12' 'Sequence14' 'Sequence19' 'Sequence3'
'Sequence37' 'Sequence39' 'Sequence4' 'Sequence46' 'Sequence8']
True Type 2:
['Sequence11' 'Sequence12' 'Sequence13' 'Sequence14' 'Sequence22'
'Sequence23' 'Sequence25' 'Sequence27' 'Sequence34' 'Sequence39']
True Type 3:
['Sequence16' 'Sequence22' 'Sequence26' 'Sequence27' 'Sequence31'
'Sequence33' 'Sequence38' 'Sequence45' 'Sequence46' 'Sequence48']
True Type 4:
['Sequence10' 'Sequence2' 'Sequence21' 'Sequence23' 'Sequence3'
'Sequence33' 'Sequence38' 'Sequence42' 'Sequence8' 'Sequence9']

==========

There are 10 Predicted Type 1:
['Sequence11' 'Sequence12' 'Sequence14' 'Sequence19' 'Sequence3'
'Sequence37' 'Sequence39' 'Sequence4' 'Sequence46' 'Sequence8']
There are 10 Predicted Type 2:
['Sequence11' 'Sequence12' 'Sequence13' 'Sequence14' 'Sequence22'
'Sequence23' 'Sequence25' 'Sequence27' 'Sequence34' 'Sequence39']
There are 10 Predicted Type 3:
['Sequence16' 'Sequence22' 'Sequence26' 'Sequence27' 'Sequence31'
'Sequence33' 'Sequence38' 'Sequence45' 'Sequence46' 'Sequence48']
There are 10 Predicted Type 4:
['Sequence10' 'Sequence2' 'Sequence21' 'Sequence23' 'Sequence3'
'Sequence33' 'Sequence38' 'Sequence42' 'Sequence8' 'Sequence9']

==========

========== PREDICTION ON 200 SEQUENCES TEST DATA ==========

There are 100 Type 1 Sequences:
['Sequence0' 'Sequence1' 'Sequence10' 'Sequence11' 'Sequence12'
'Sequence13' 'Sequence14' 'Sequence15' 'Sequence16' 'Sequence17'
'Sequence18' 'Sequence19' 'Sequence2' 'Sequence20' 'Sequence21'
'Sequence22' 'Sequence23' 'Sequence24' 'Sequence25' 'Sequence26'
'Sequence27' 'Sequence28' 'Sequence29' 'Sequence3' 'Sequence30'
'Sequence31' 'Sequence32' 'Sequence33' 'Sequence34' 'Sequence35'
'Sequence36' 'Sequence37' 'Sequence38' 'Sequence39' 'Sequence4'
'Sequence40' 'Sequence41' 'Sequence42' 'Sequence43' 'Sequence44'
'Sequence45' 'Sequence46' 'Sequence47' 'Sequence48' 'Sequence49'
'Sequence5' 'Sequence50' 'Sequence51' 'Sequence52' 'Sequence53'
'Sequence54' 'Sequence55' 'Sequence56' 'Sequence57' 'Sequence58'
'Sequence59' 'Sequence6' 'Sequence60' 'Sequence61' 'Sequence62'
'Sequence63' 'Sequence64' 'Sequence65' 'Sequence66' 'Sequence67'
'Sequence68' 'Sequence69' 'Sequence7' 'Sequence70' 'Sequence71'
'Sequence72' 'Sequence73' 'Sequence74' 'Sequence75' 'Sequence76'
'Sequence77' 'Sequence78' 'Sequence79' 'Sequence8' 'Sequence80'
'Sequence81' 'Sequence82' 'Sequence83' 'Sequence84' 'Sequence85'
'Sequence86' 'Sequence87' 'Sequence88' 'Sequence89' 'Sequence9'
'Sequence90' 'Sequence91' 'Sequence92' 'Sequence93' 'Sequence94'
'Sequence95' 'Sequence96' 'Sequence97' 'Sequence98' 'Sequence99']
There are 0 Type 2 Sequences:
[]
There are 50 Type 3 Sequences:
['Sequence100' 'Sequence101' 'Sequence102' 'Sequence103' 'Sequence104'
'Sequence105' 'Sequence106' 'Sequence107' 'Sequence108' 'Sequence109'
'Sequence110' 'Sequence111' 'Sequence112' 'Sequence113' 'Sequence114'
'Sequence115' 'Sequence116' 'Sequence117' 'Sequence118' 'Sequence119'
'Sequence120' 'Sequence121' 'Sequence122' 'Sequence123' 'Sequence124'
'Sequence125' 'Sequence126' 'Sequence127' 'Sequence128' 'Sequence129'
'Sequence130' 'Sequence131' 'Sequence132' 'Sequence133' 'Sequence134'
'Sequence135' 'Sequence136' 'Sequence137' 'Sequence138' 'Sequence139'
'Sequence140' 'Sequence141' 'Sequence142' 'Sequence143' 'Sequence144'
'Sequence145' 'Sequence146' 'Sequence147' 'Sequence148' 'Sequence149']

There are 50 Type 4 Sequences:
['Sequence150' 'Sequence151' 'Sequence152' 'Sequence153' 'Sequence154'
'Sequence155' 'Sequence156' 'Sequence157' 'Sequence158' 'Sequence159'
'Sequence160' 'Sequence161' 'Sequence162' 'Sequence163' 'Sequence164'
'Sequence165' 'Sequence166' 'Sequence167' 'Sequence168' 'Sequence169'
'Sequence170' 'Sequence171' 'Sequence172' 'Sequence173' 'Sequence174'
'Sequence175' 'Sequence176' 'Sequence177' 'Sequence178' 'Sequence179'
'Sequence180' 'Sequence181' 'Sequence182' 'Sequence183' 'Sequence184'
'Sequence185' 'Sequence186' 'Sequence187' 'Sequence188' 'Sequence189'
'Sequence190' 'Sequence191' 'Sequence192' 'Sequence193' 'Sequence194'
'Sequence195' 'Sequence196' 'Sequence197' 'Sequence198' 'Sequence199']