Feature Selection for Genome Classification
Elaine Garbarine and Gail Rosen

Motivation
Metagenomics
• A large portion of existing bacteria cannot be cultured in a lab environment and thus sequencing them is impossible with traditional approaches.
• Metagenomics has evolved to solve this problem by taking environmental samples containing many organisms and sequencing the samples.

Next Generation Sequencing
• New sequencing methods such as Roche 454 and Solexa can sequencing a great deal of DNA quickly.
• These methods produce short sequences anywhere from 35 base pairs to 450 base pairs which make classification very difficult.

Goals
• Develop a method to select a number of smaller words, called Nmers, that can be used as features for a Support Vector Machine Classifier.
• Use the SVM and these features to classify bacterial genomes.

Feature Selection Methods
TF-IDF:
• Used in text mining to evaluate word occurrences in a document.
• Can be viewed as the amount of information of a term (specific word) weighted by its occurrence probability.

\[
\text{TF-IDF}_t = \sum_{m=1}^{M} \left( \frac{d_{mt}}{\sum_{l=1}^{L} d_{ml}} \log \left( \frac{M}{\sum_{m=1}^{M} d_{ml}} \right) \right)
\]

mRMR:
• Minimum-Redundancy Maximum-Relevance.
• Selects features subject to the constraint that they are mutually as dissimilar to each other as possible while also being marginally as similar to the classification variable as possible.
• Has been used on Microarray data sets for genes related to different types of cancers.

Data Set and Frequency Profiles
Data Set:
• 100 bacterial genomes.
• Cover 3 phyla, 14 Genera, 64 Species.
• 6 Genera belong to Firmicutes phyla, 6 to Proteobacteria phyla and 2 to the Cyanobacteria phyla.

Creating Frequency Profiles
• Use short sequences of a specific length called Nmers.
• For this work Nmers of length N=6 and 9.
• The frequency of all possible Nmers are counted in each genome sequence.
• These frequency count profiles for each genome are used for the probe selection.

Feature Selection Methods
Kullback-Leibler Divergence/Mutual Information
• Kullback-Leibler (\(D_{KL}\)): Measure of the difference between two probability distributions, in this case frequency profiles for two genomes A and B.

\[
D_{KL}(A(m), B(m)) = p_{A}(m) \log_2 \frac{p_{A}(m)}{p_{B}(m)} + p_{B}(m) \log_2 \frac{p_{B}(m)}{p_{A}(m)}
\]

• Mutual Information (I(\(X(m), C\))): measures how much information is shared between a word and a genome.

\[
\arg \max_{X(m)} I(X(m), C) = \sum_{c \in C} p(X_c(m), c) \log_2 \frac{p(X_c(m), c)}{p(X_c(m))p(c)}
\]

C is the set of genomes A and B.

Conclusions and Future Work
Conclusions:
• 6mers selected using the feature selection techniques mostly outperform 9mers.
• 9mers are problematic due to the large number of possible features (4096 vs. 26214).

Future Work:
• Investigate 7 and 8mers for genome classification.
• Attempt to improve classifier results for 9mers.
• Test other classifiers to determine if better performance can be achieved.