Application of Supervised Learning Algorithms for Rapid Identification of MRSA PFGE Strain Type

M.L. Bernauer, J. M. Lewis, J. J. Shroba, R. C. Mercier
University of New Mexico College of Pharmacy
Albuquerque NM, 87131

May 21, 2015

1 Background

Pulsed-field gel electrophoresis (PFGE) is a molecular typing method used in epidemiologic surveillance of MRSA. PFGE relies on electrophoretic migration of restricted DNA fragments, resulting in characteristic fingerprints which can be used to classify organisms. Current methods of analysis, such as Gel Compar II and BioNumerics rely on unsupervised hierarchial clustering algorithms to group organisms based on pairwise similarity. These methods are labor intensive, often requiring a significant amount of user intervention and oversight. This study presents an automated approach to PFGE typing to reduce the amount of user involvement and time required for analysis.

2 Methods

A total of 70 gel electrophoresis images were obtained from previous PFGE experiments conducted according to protocol. Lanes were extracted from the raw images using k-means clustering, resulting in 1067 fingerprints. Transformations applied to each of the fingerprints include; normalization, alignment and background subtraction. The training set was created manually by labeling each fingerprint according to PFGE type. Several fingerprints were removed due to the presence of artifacts, resulting in 843 (17% USA100, 72% USA300, 11% USA400) for the final analysis. Principal component analysis was used to reduce dimensionality by selecting the top 20 components. The support vector machine (SVM) was implemented using 10-fold cross validation. Accuracy of the other classification algorithms was assessed by splitting the data 70:30 training to test. Hyperparameters for k-nearest neighbors (kNN), random forest (RF) and the artificial neural network (aNN) were selected via grid search to yield the highest test set accuracy.

3 Results

The aNN and SVM were the highest performing algorithms with overall classificaiton accuracies of 0.9004, p=$1.15x10^{-12}$, 95% CI [0.8565, 0.9345] and 0.8845, p=$1.33x10^{-5}$, 95% CI [0.8383, 0.9212], respectively. Both kNN and RF algorithms performed similarly with classification accuracies of 0.8606, p=$5.38x10^{-8}$, 95% CI [0.8114, 0.9009] and 0.8685, p=$8.34x10^{-9}$, 95% CI [0.8203, 0.9077], respectively.

4 Conclusions

Supervised learning algorithms such as SVM and aNN provide feasible alternatives to traditional hierarchial clustering methods for determination of MRSA PFGE strain type.