doi:10.1017/S1431927616005882
Microsc. Microanal. 22 (Suppl 3), 2016

Imaging and Feature Selection Using GA-FDA Algorithm for the Classification of Mid-Infrared Biomedical Images

Rupali Mankar[1], Vishal Verma[2], Michael Walsh[2], Carlos Bueso-Ramos[3], David Mayerich[1]

[1.] Department of Electrical and Computer Engineering, University of Houston, Houston, USA
[2.] Department of Pathology, University of Illinois at Chicago, Chicago, USA
[3.] Division of Pathology/Lab Medicine, University of Texas MD Anderson Cancer Center, Houston, USA

Pathologists currently rely on chemical staining of tissue samples to perform disease diagnosis. However, these techniques are highly prone to variations due to the clinical environment and staining protocol. Mid-infrared spectroscopic imaging has the potential to overcome several of these problems by providing quantitative molecular information that can be used for highly specific tissue classification [1] [4]. However, there are bottlenecks that limit clinical applicability of these methods. For example, Fourier-transform infrared (FTIR) spectroscopic images are often time-consuming to acquire at a spectral resolution and SNR level that is viable for reliable classification. In addition, algorithms are computationally intensive and data sets can be terabytes in size. This research focuses on improving the clinical viability of mid-infrared spectroscopy by utilizing a QCL based imaging system, which allows feature selection from hyperspectral images (HSI) using a Genetic Algorithm and Fisher's Discriminant Analysis as a fitness function (GA-FDA) and classification of HSI images using Random Forest classifier with the optimized feature subsets selected by GA-FDA [3]. We demonstrate that GA-FDA is very promising feature selection algorithm with performance superior to unsupervised as well as supervised feature selection algorithms while being compatible with optimized QCL imaging methods.

Hyperspectral images are acquired using both FTIR spectroscopic imaging (Cary 620, Agilent) and a SPERO QCL imaging system (Daylight Solutions). MIR spectroscopy relies on the Beer-Lambert law and the absorbance is given by taking measurments of the light transmitted both with and without the sample present. These images are then pre-processed using baseline correction and normalization to the Amide I spectral band (1650 cm^{-1}). Imaging of one Tissue Micro Array (TMA) with an 8 cm^{-1} spectral resolution, 5.5μm spatial resolution, and 40 co-adds takes 15 hours on the FTIR spectroscope. Reducing the long imaging hours and processing time while achieving clinically usable classification results are main goals of this study.

The proposed GA-FDA algorithm is compared with other dimensionality reduction algorithms including PCA, mRMR, and Bhattacharya Distance (BD). GA evaluates each genome (i.e. feature subspace) from the population with FDA and ranks them according to their fitness scores. Fisher's discriminant analysis seeks to find a linear transformation $\hat{T}_{FDA}$ to a new reduced dimension feature subspace that maximizes fisher's ratio:

$$\hat{T}_{FDA} = \underset{T}{\operatorname{argmax}} \left\{ \operatorname{trace} \left[\frac{(T^T S^b T)}{T^T S^w T} \right] \right\}$$

Here, S^b and S^w define between-class and within-class scatter. The GA algorithm with FDA as a fitness function will look for features which maximize fisher's ratio (i.e. the fitness score of each gnome) when the $\hat{T}_{FDA}$ projection is applied on them.

Optimized features and neighboring baseline points from the feature selection algorithm are used for cell type classification using a Random Forest Classifier. The results are comparable with the use of whole spectra for the classification while significantly increasing computation efficiency. Validation of results is done on completely separate data sets. We found that GA-FDA performs better than PCA, mRMR and BA on complex data (Figure 2). The key point of GA is that it selects features subsets, whereas mRMR and BA select individual features. Hence, GA avoids selection of correlated features which is highly probable for algorithms which select features on an individual basis.

Conclusion

Classification based on spectral signature provides high accuracy for cancer-relevant tissue types. Dimension reduction using localized metrics (individual wavelengths) provides the potential for informed imaging using tunable QCLs, dramatically reducing image acquisition time.

References

[1] D. Fernandez *et al*, Nature Biotechnology, 23, 469-474 (2005).
[2] M. Baker *et al*, Nature Protocols, 1771-1791 (2014)
[3] M. Cui *et al*, IEEE: Applied earth observation and remote sensing Vol.6, No.3, June2013
[4] D. Mayerich *et al*, Technology, 03(1):27 (20

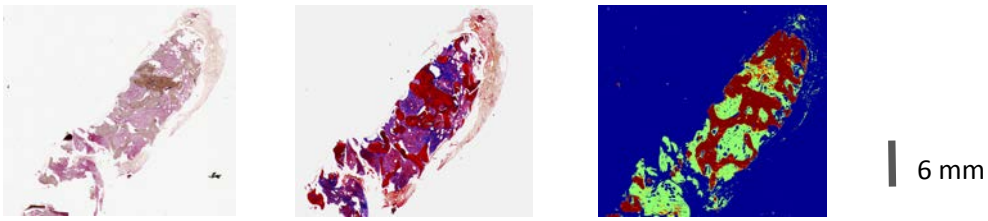

Figure 1. a) Reticulum stain of Bone Marrow. b) Mason's trichrome staining of bone marrow showing bone (red) and collage (blue). c) Classification results of bone marrow for bone with 6 features selected using GA-FDA selected showing bone (red) and stroma (green).

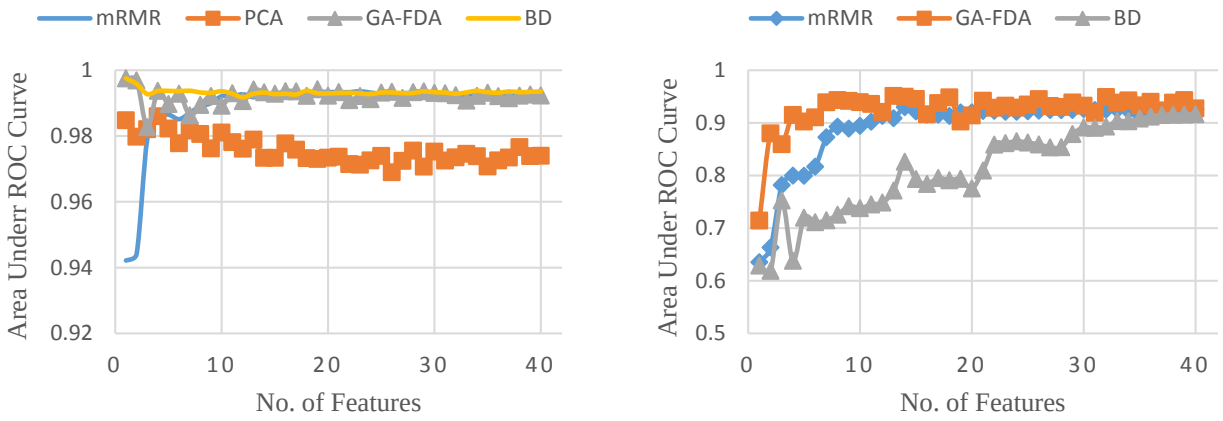

Figure 2. Plots of area under the ROC curve (AUC) vs no. of features used for the Classification of a) Bone and stroma in Bone Marrow biopsy (AUC = 0.9975 for all features) (left) b) Epithelium and fibrosis in the liver biopsy (AUC = 0.9556 for all features)(right)

https://doi.org/10.1017/S1431927616005882 Published online by Cambridge University Press