Multidimensional Shape and Color Distributions as a Computational Biomarker for Cancer Pathology – Predicting Lymph Node Status

Aladin Milutinovic1, Md. Alimoor Reza2, David E. Breen2, and Fernando U. Garcia1.

1Pathology Department, Drexel University College of Medicine, Philadelphia, PA, United States
2Computer Science Department, College of Engineering, Drexel University, Philadelphia, PA, United States.

Background

Imaging, both in 2D and 3D, is now a ubiquitous procedure in medicine, science and engineering. The resulting images contain a wealth of information that to date has only been partially utilized. We are developing a computational biomarker for the characterization of breast carcinomas to predict lymph node status.

Methods

Our method is based on a computational examination of a routinely applied prognostic panel that uses immunohistochemistry and hematoxylin & eosin (H&E) staining of 50 invasive breast cancer carcinomas with known lymph node status. This panel includes: estrogen and progesterone receptors, MIB-1 (proliferative activity), mutated p53 and HER2/neu. The highest staining marker from the panel was selected for further digital image analysis.

Using image processing techniques and geometric analysis, the architectural histologic pattern and the expression of these prognostic markers were analyzed to create feature vectors that numerically characterize each tumor’s features. Tumor characteristics including pathologic staging were added to the feature vectors. These feature vectors are fed into an algorithm called Support Vector Machines (SVM), which uses a training set of feature vectors to classify unknown samples. We applied the leave-one-out approach for the training and testing. One sample is removed for testing and remaining 49 samples are used for training. This process is iterated over all 50 samples and results for each classification combined to calculate sensitivity and specificity of the classification. Several kernel variations and their parameters were tested to optimize the sensitivity and the specificity of the classification. These kernels include linear, polynomial 2, 3, 4, 5, 6, 7, 8, 9, and 10, and also radial basis function kernels 5 though 150 incremented in interval steps of 5.

Additionally, we created feature vectors using tumor characteristics to test their predictability. We group them into three groups. One group includes age, tumor size and histologic grade. Another vector contained only the percentages of staining cells from each of the markers. Third, combined both of the groups.

Results

A total of 50 cases were processed to create the feature vectors. Of the 50 cases, 25 were lymph node positive (N1 and higher) and 25 negative (N0). In iterative testing of the algorithm, where one case is used as an unknown specimen and all others are used for training, the best results were obtained using a Radial Basis Function kernel with a sigma parameter of 35 using a feature vector of 13 distributions that included tumor characteristics. Of the 50 cases, 19 lymph positive and 16 lymph negative cases were correctly classified (0.76 sensitivity and 0.64 specificity).

Conclusions

1. SVM analysis can predict lymph node status in the majority of patients with invasive mammary carcinoma.
2. Age, Tumor Size, and Histologic Grade when used as a feature vector produce slightly predictive results. Percentage of staining cells for each IHC marker are not predictive. When both are combined with geometric and color feature vectors, the predictability improves.
3. Refinement of the feature vectors is currently in development to increase specificity and sensitivity. Each feature vector will be examined individually to test its predictive value. Additionally, statistics on the distributions will be used as feature vectors.
4. N0 lymph nodes will be further studied using NSABP protocol B-32 to improve specificity in around 11.4% (detection of occult metastases).
5. Future studies will explore other classification methods such as EMD, KNN, and MLP using additional study samples.
6. UPDATE: Our latest results using 100 total cases produced a Sensitivity of 0.80 and Specificity of 0.82 when using individual metrics to form a prediction with SVM, and then feeding the prediction into SVM again to perform stacked classification.

References