

## **A Novel Computational Approach for Biomarker Detection for Gene Expression-Based Computer-Aided Diagnostic Systems for Breast Cancer**

Cancer produces complex cellular changes. Microarrays have become crucial to identifying genes involved in causing these changes; however, microarray data analysis is challenged by the high-dimensionality of data compared to the number of samples. This has contributed to inconsistent cancer biomarkers from various gene expression studies. Also, identification of crucial genes in cancer can be expedited through expression profiling of peripheral blood cells. We introduce a novel feature selection method for microarrays involving a two-step filtering process to select a minimum set of genes with greater consistency and relevance, and demonstrate that the selected gene set considerably enhances the diagnostic accuracy of cancer. The preliminary filtering (Bi-biological filter) involves building gene coexpression networks for cancer and healthy conditions using a topological overlap matrix (TOM) and finding cancer specific gene clusters using Spectral Clustering (SC). This is followed by a filtering step to extract a much-reduced set of crucial genes using best first search with support vector machine (BFS-SVM). Finally, artificial neural networks, SVM, and K-nearest neighbor classifiers are used to assess the predictive power of the selected genes as well as to select the most effective diagnostic system. The approach was applied to peripheral blood profiling for breast cancer where Bi-biological filter selected 415 biologically consistent genes, from which BFS-SVM extracted 13 highly cancer specific genes for breast cancer identification. ANN was the superior classifier with 93.2% classification accuracy, a 14% improvement over the study from which data were obtained for this study (Aaroe et al., Breast Cancer Res 12:R7, 2010).