Support Vector Machine Classification of Data Quality in Microarray Experiments

Timothy S. Davison, Sunil Mehta, Ingrid J. Burgetz, Ozgur Huner

University Health Network, Clinical Informatics

The use of microarray technology has greatly expanded our understanding of cellular function at the level of gene expression, and as such provides us with a powerful tool for gene-based disease diagnostics and drug efficacy experiments. Present applications of microarray experiments focus on the analysis of comparative gene expression levels in control, diseased (treated or untreated) and/or cell-cycle gene profiles. The information produced in microarray experiments is then derived from the log-ratio of fluorescent intensity values for each spot measured for two or more different hybridization experiments performed simultaneously on the surface of a slide. Although great care is taken in making each microarray slide, flaws in the spots inevitably arise due to processing errors including scratches, smudges, and asymmetric drying of the DNA. As a result, each of the spots in the array displays significant variability in the distribution of their fluorescence intensities. Consequently, the researcher must examine the images resulting from the microarray hybridization experiments to identify "bad" spots in order to remove them from further analysis. If performed at all, this process is typically done by eye which is both extremely time consuming and provides different results depending on the opinions and experience of each researcher.

Our research is focused on developing an unbiased, automated approach to the classification of data quality for each spot in a microarray experiment. This will both speed up the analysis and provide a standardized filter for tagging and removing "bad" data points from further analysis. To do this, we are using the QuantArray3 (Packard BioScience) software package that identifies individual spots in a tiff image while also providing eight quality metrics for each spot in the experiment. These metrics include spot diameter, area, footprint, circularity, spot uniformity, background uniformity, signal to noise, and confidence. Together with the spot and background intensities and standard deviations, these metrics provide a fingerprint of the data quality for each spot in the microarray experiment. In order to achieve a high-throughput application for the analysis of large quantities of data, we have implemented a supervised machine-learning technique known as a Support Vector Machine (SVM) with a Kernel-Adatron (KA) hard-margin optimization routine in the MATLAB (MathWorks) programming language. Using the set of fingerprint metrics for each spot to define the input space for the SVM, one of several kernels (including the radial basis function and polynomial functions) are used to map the training data into a high dimensional feature space. The KA optimization algorithm then rapidly finds the maximal-margin hyperplane, which separates the data into the binary classes of "good" or "bad" spots. Training of the classifier relies upon a supervised classification based on the statistical properties of the spot quality fingerprints in parallel and/or serial with an expert user's spot classification. Initial results indicate that our SVM classifier trained on as little as 100 spots can classify a 19K spot array with a generalization error of 2-3%. Once trained the SVM classifier produces deterministic results, thus providing a standardized benchmark for rapid spot quality classification across any microarray experiment.