Return to Contents PDF Version

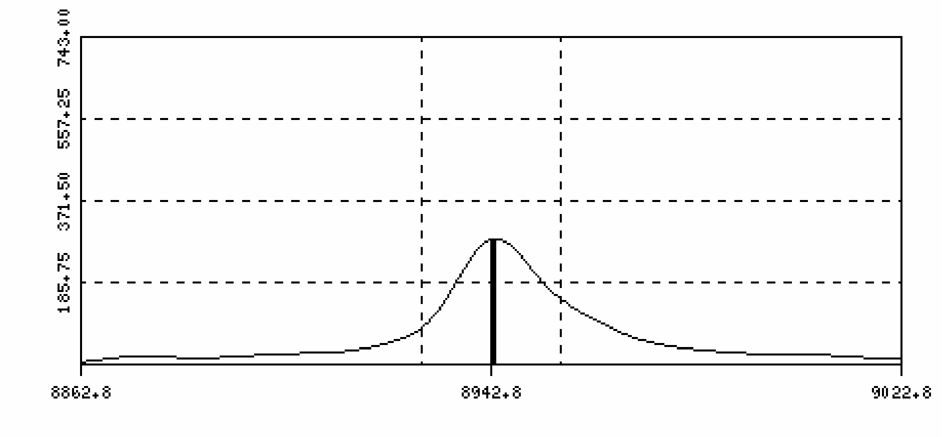
BioMarker Discovery Kit (BMDK)

Brian T. Luke (lukeb@ncifcrf.gov)

The BioMarker Discovery Kit (BMDK) represents a suite of programs with the eventual goal of constructing one or more biomarker-based classifiers. Each biomarker represents a particular feature that is associated with a particular State represented by a subset of the available individuals. In the following discussion it is assumed that the original data comes from a mass spectral analysis of a given biofluid from each individual. The overall process is as follows.

Spectral Pre-processing: The first step is to discard all spectra below an *m/z* value of 1200 to 1500; removing any spectral peaks produced by the energy absorbing matrix. Though generally done by software distributed with the mass spectrometer, it is also necessary to ensure that all baseline spectra have been removed. If this was not adequately done, BMDK contains a program that searches for baseline peaks by calculating the Pearson correlation coefficient between intensities at adjacently measured *m/z* values across all spectra. Baseline regions should show random fluctuations in intensity and a low correlation coefficient will identify baseline m/z values. A piece-wise linear fit to sections of baseline *m/z* values is then used to remove the baseline spectra for at all *m/z* values within this region. The next step is to consistently scale the spectra. In BMDK, this is done by setting the total ion current (sum of all remaining intensities) to a constant in each spectrum.

Peak Picking: The goal of a peak picking algorithm is to select regions of significant intensity and assigning a single intensity to each region, as shown in the figure below. Though several studies construct classifiers from specific *m/z* values [Bro-05, Con-04, Orn-04, Pet-05, Sri-06, Sto-05] tests examining duplicate spectra obtained from the same individual have shown that the *m/z* value representing the maximum intensity in one of the spectra is different from the maximum *m/z* value in the other. Combining several sequential m/z intensities into a single bin, known as binning, still results in a significant probability still exists for placing the maximum intensity for a given peak into different bins.



Instead of looking for regions of significant intensity in each spectrum, BMDK sums all spectra together so that the highest intensity resides at the average *m/z* value for each peak. The user supplies either an intensity threshold, given as a fraction of the average intensity across all *m/z* values in the summed spectrum, or a total number of regions to select. In addition, the user supplies the desired width of the region to scan, or read-window, as a fraction of *m/z*. This fractional width is denoted *w*. The program searches the summed spectrum and finds the m/z value with the highest intensity. This m/z value is placed in a peak-list and all of the intensities in the summed spectrum within *w*(*m/z*) of this *m/z* are set to zero. This process continues until either the maximum remaining intensity is below the threshold value or the desired number of *m/z* values has been added to the peak-list. The peak-list is then examined to determine if the selected region of sufficient intensity represents a peak shoulder instead of a peak. This is done by examining all *m/z* intensities within the read-window and determining if the maximum intensity occurs at the first or last few *m/z* values. If this occurs in a sufficient number of samples, this region represents a peak shoulder and the *m/z* value is removed from the peak-list.

A separate program reads the *m/z* values from the reduced peak-list and processes the individual spectra. In particular, a read-window of width *w*(*m/z*), centered about the *m/z* value, is placed in each spectrum and the maximum intensity within this window is stored as that spectrum's intensity for that particular *m/z*. This continues until all *m/z* values in the peak-list have been applied to all spectra. This reduces each sample from intensities at tens of thousands of *m/z* values to a few hundred features.

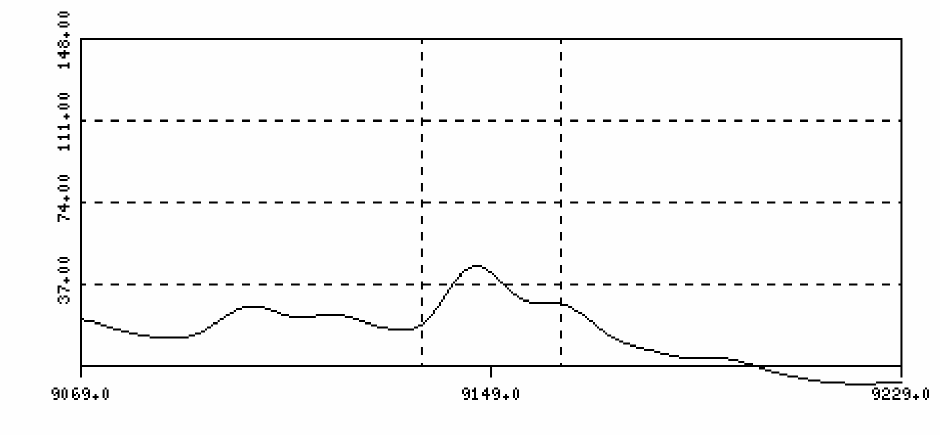
Processing Duplicate Spectra: Another area of concern is the handling of duplicate spectra. In many experimental investigations the number of samples that have a particular disease or meet specific requirements is very small. As the number of samples gets smaller, the confidence in the identification of a putative biomarker is reduced. Therefore, treating each spectrum taken from a given sample as an independent result increases the amount of data available. Unfortunately, if the duplicate spectra are extremely similar this could adversely affect the search for putative biomarkers by artificially increasing a particular features ability to group each spectrum with its duplicate. On the other hand the search for a putative biomarker should include any significant experimental variability. In addition, a slight defect in a particular surface or any problem in preparing the spectra may cause one member of the duplicate to be a bad result. If all duplicates are automatically averaged, the experimental variability would be reduced and a good spectrum may become contaminated by a bad one. BMDK includes the option of taking a middle ground and averaging duplicate spectra together only if they are highly similar. The Euclidean distance between samples taken over all of the features is used to measure the dissimilarity between two spectra. If the i^{th} sample produces spectra i_1 and i_2, the program determines the number of times a spectrum from another sample is closer to i_1 and the number of times another sample's spectrum is closer to i_2 than i_1 is to i_2. If the sum of these numbers is larger than a threshold, then these two spectra are considered to be sufficiently different and are not averaged. Otherwise they are averaged.

Outlier Detection: The final step in creating the dataset is the search for outliers. By definition, outliers are spectra that are quantitatively different over the several hundred features. BMDK identifies two different types of outliers. A Type-1 outlier generally has a very abnormal intensity in a small number of features, while a Type-2 outlier has smaller abnormal intensities in many features. A Type-1 outlier is identified by having a large distance to its nearest neighbor. Therefore, by calculating all nearest neighbor distances the average and standard deviation is determined. If a spectrum's nearest neighbor distance is more than two standard deviations above the mean, it is a Type-1 outlier. A Type-2 outlier has a large number of near-extreme intensities. For each of the several hundred features, the program determines the total range of intensities across all spectra. The number of times each spectrum has intensity in the upper or lower 5% of the range is determined, and that spectrum is a Type-2 outlier if this number is more than two standard deviations above the mean.

A Type-2 outlier can occur if an abnormal peak is present in the spectrum. The initial scaling of the spectrum would cause all other intensities to be reduced relative to the intensities of the other spectra. Since the search for Type-1 outliers uses all spectra independent of the individual's histology, it is possible to be a Type-2 outlier without also being a Type-1 outlier. By combining conditional averaging with outlier detection, a single aberrant spectrum will not be combined with its duplicate and only this single spectrum would be removed from consideration.

Identifying Putative Biomarkers: Once the original spectra are converted to a smaller set of feature intensities, BMDK uses 10 different methods of analysis to identify putative biomarkers. These methods determine how well each feature distinguishes some or all of the individuals in a given histology. The union of all features that have one of the top five scores for each of the 10 methods produces the set of putative biomarkers.

Examining the Putative Biomarkers: A single biocompound may produce more than one putative biomarker by having separate peaks for the +1 and +2 ion or the biocompound alone and complexed with the compound used as the energy absorbing matrix. Therefore, the Pearson's correlation coefficient between all pairs of putative biomarkers across all samples is used to combine the putative biomarkers into groups. All other features in the dataset are then compared to the putative biomarkers within each group and are selected for examination if the correlation coefficient is 0.70 or higher. The original spectra for each putative biomarker and all correlated features are then visually examined to determine if they represent well defined peaks in the spectrum. For example, the peak shown above represents a well defined peak while the spectrum below represents a compound peak.



The relative size and separation of the two component peaks varies from one spectrum to the next. The magnitude of the highest recorded intensity within this read-window could be due to more than one biocompound and the experimental variability in their separation. Though these two peaks can be deconvoluted, the current implementation of BMDK only allows the intensity from a well defined peak to be used in the final classifier. If more than one well defined peak is present in a group of correlated peaks, only the peak with the highest maximum intensity is retained. Therefore, the final set of putative biomarkers contains an uncorrelated list of maximum intensities from well defined peaks.

Searching for Bias: Before these putative biomarkers are used to construct a final classifier, they should be examined to determine if their intensity is due to some factor other than the State of the individual. This is done by determining if there is any correlation between the peak's intensity and the individuals age [Hab-06], whether or not they are under specific medication, the order in which the spectra were generated [Con-04], or some other external factor. If any correlation is observed, this peak should be removed from the list of putative biomarkers.

Constructing a Classifier: The final classifier is based on a distance-dependent K-nearest neighbor (DD-KNN) algorithm. All of the putative biomarkers are individually used to fine the best 1-peak DD-KNN algorithm, and this is followed by an exhaustive search over all sets of two and three putative biomarkers. In practice, six nearest neighbors are generally used but this number can be increased if there are a large number of samples; the number of neighbors should not decrease below six. The quality of the classifier is determined using a leave-one-out procedure since this method preserves the coverage (range of intensities) for the samples to the greatest extent. Each time an optimum classifier is found, the distribution of samples in feature-space is plotted to determine the number of States present for each category. If a given category divides into two or more States, the specifics of the individuals in each State need to be examined with respect to the putative biomarker that distinguishes it from the other States. If this division does not make biological sense, then this classifier should not be used. While a DD-KNN algorithm is not the only classifier that can be constructed using the putative biomarkers, tests have shown that a decision tree (DT) and medoid classification algorithm (MCA) classifier greatly overstates the accuracy (sensitivity and specificity).

Testing the Classifier: The leave-one-out procedure produces a fairly accurate estimation of the accuracy of the classifier. The only time the accuracy was found to be inflated is when one or more putative biomarkers divide the individuals in a given category into multiple States. This is seen in a visual inspection of the samples in feature-space. A Bootstrap or n-fold cross-validation can also be used, but for these the number of samples placed in the testing set should be very small. If a large number of samples are excluded from the training set, there may be an incomplete coverage of the range of intensities for each putative biomarker leading to an artificially poor result for the testing set. Other procedures are being developed to estimate the quality of the classifier if it is applied to the underlying population of individuals.

(Last updated 4/28/07)