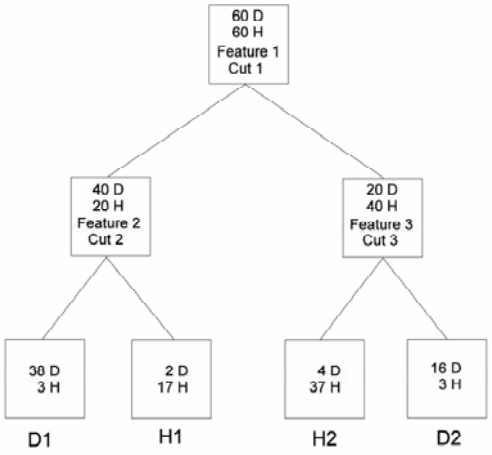
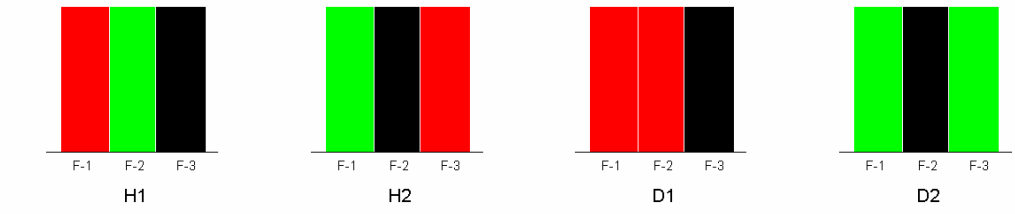
Fingerprint-based versus Biomarker-based Classifiers

Brian T. Luke (lukeb@ncifcrf.gov)

Informatic analysis of biofluids has led to a new paradigm for classification known as fingerprinting or pattern matching. In this paradigm, individuals are classified based upon a particular pattern of intensities obtained from spectroscopic or microarray investigations. If an untested individual has the same pattern as a known individual, then it is assumed that these two have the same classification. In other words, the classification is based upon an individual's pattern or feature values or fingerprint, and if an unknown individual has a sufficiently similar fingerprint to an individual that is in the diseased category, then this individual is classified as having the disease. This concept has been popularized through publications from the laboratories of Petricoin and Liotta [Bro-05, Con-04, Orn-04, Pet-05, Sri-06, Sto-05]. In addition to their medoid based classification algorithm (MCA), other classification schemes that use a fingerprint are a single decision tree (DT) and an artificial neural network (ANN). In contrast, we propose that many different procedures should be used to find state-specific markers, and only these markers should be used in the final classifier.

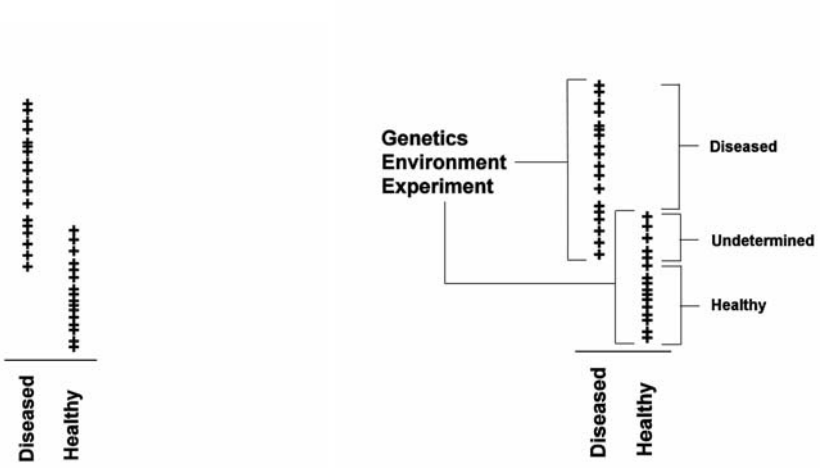
The simplest example of fingerprinting is a single decision tree, like the one shown at the right. Assuming that the entire dataset is composed of 60 diseased and 60 healthy individuals, the intensity of Feature 1 splits the dataset into two groups; 40 diseased and 20 healthy individuals if the intensity of this feature is below Cut-1 and 20 diseased and 40 healthy individuals if its intensity is above Cut-1. The left branch is further divided using Feature 2 into a diseased node (D1) that contains 38 diseased and 3 healthy individuals and a healthy node (H1) that contains 2 diseased and 17 healthy individuals. The right branch is divided using Feature 3 into a healthy (H2) and a diseased (D2) terminal node. If for each of the three features a red rectangle means that the intensity is below the cut point, a green rectangle means that the intensity is above the cut point, and a black rectangle means that the intensity can be any value, then the following fingerprints represent these four nodes.



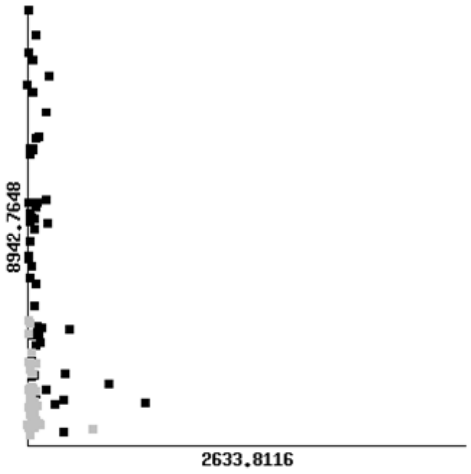


The philosophy behind a biomarker-based classifier is best described by the intensity plots shown below. The first plot (a) shows the intensities of a putative biomarker with the intensities of the diseased samples shown in the left column and those of the healthy samples in the right. While the intensities vary within each category due to genetic, environmental, and experimental effects (b), there is a shift in the range of intensities due to the presence or absence of the disease. This putative biomarker divides the samples into three groups; a diseased group, a healthy group, and an uncertain group (b).

(a) (b)



It should be stressed that a putative biomarker like the one shown above has a shift in the range for all samples within a given state, as opposed to a pattern of feature intensities associated with an individual. In other words, the shift in the intensity rage should be a consequence of the presence of the disease, and nothing else. The goal is to find these state-based markers and to build a classifier using only these markers. The situation shown in the intensity scatter plot to the right may complicate the search. This scatter plot represents the intensities for the best 2-feature distance-dependent 6-nearest neighbor classifier for individuals with benign prostate hyperplasia (BPH, black) and individuals with healthy prostates (grey) [Luk-07] using peak intensities from the study by Adam *et al*. [Ada-02]. The majority of the BPH individuals are characterized by high intensity in the peak centered at an m/z value of 8943, which has been previously identified as the blood form of complement C3a anaphylatoxin [Hab-06]. A small fraction of the BPH individuals do not have sufficient intensity in the 8943 peak, but a majority of them have a relatively higher intensity in the peak at *m/z* of 2634 (as does a single healthy individual). If the biomolecule associated with this 2634 peak is identified and found to be associated with



BPH, or a separate host response to BPH, it may be that there are actually two states associated with BPH and treatment of individuals in the two states may differ. If no connection between the biomolecule at *m/z* of 2634 and BPH is found, then this classifier is not valid and only C3a can be used in the classification.

In the examination of methods to search for putative biomarkers, the extent to which they can identify putative biomarkers associated with a small fraction of the Cases will be presented. It is possible that if a given disease state is rare enough; these methods may not be able to find the biomarker associated with this state. An alternative (untested) procedure may be to extract those Cases that are described by the identified marker(s) and then repeat the search to find a marker for the remaining Cases.

It is important to realize that a battery of tests used to identify different states within a given category (e.g. BPH) is fundamentally different from a panel of markers used to construct an individual-based fingerprint classifier. In addition the concept of "personalized medicine" should be replaced by "State-based diagnosis and treatment". It should also be noted that examinations of this dataset comparing BPH to healthy individuals [Luk-07], complete datasets that contain no information (i.e. completely random peaks), and a subset of peaks representing putative biomarkers shows that fingerprint-based classifiers have problems with coverage, uniqueness, and significance. Therefore, fingerprint-based classifiers are not generalizable to the underlying population and should not be used.

(Last updated 8/8/07)