Disease Categories versus Disease States

Brian T. Luke (lukeb@ncifcrf.gov)

While all cancers have many factors in common, tissue differences and the body's response to different cancers make the test for ovarian cancer (CA125) very different from the test for prostate cancer (PSA). Even within the same tissue, all cancers are not necessarily the same. It is well known that there are two major types of lung cancer, small cell lung cancer (SCLG) and non-small cell lung cancer (NSCLC). It is also known that NSCLC has three major sub-types; adenocarcinoma (AC), squamous cell carcinoma (SCC), and large cell undifferentiated carcinoma (LCUC). Each of these has differences in the biochemical processes going on within the cancer cell and one should not expect that the detection, or necessarily the treatment, of these cancers will be the same.

Disease Categories and Disease States

Lung cancer is therefore composed of (at least) four disease categories; SCLG, AC, SCC and LCUC. Since the latter three are strictly differentiated by appearances of the cell under the microscope, it is possible that the underlying biochemical processes of an AC cell in one individual is significantly different than the biochemical processes in another individual with a cancer that appears similar. Therefore, each of these categories of lung cancer may be composed of one or more states. While the disease category represents the name of the disease based on some experimental observation, the disease state represents a grouping based on the underlying biochemical processes within the diseased cell. The detection of a disease and its treatment should be relative to specific disease states, not a disease category or individuals within that category.

A SELDI-based search for putative biomarkers begins with collecting samples from individuals in two or more categories (e.g. healthy versus having a particular disease). Since the underlying assumption is that the presence of a disease changes the blood concentration of one or more protein products, the objective is to identify peaks in the mass spectra whose intensities reflect this change. It is possible for different physiological changes to result in a person being placed in the same category, which means that a specific change in a given blood protein concentration may only hold for a subset of the individuals in a given category. Each state corresponds to a single category and any classifier attempts to correctly identify an individual's state.

The simple rule for any valid classifier is that it must reflect the assumed biological reality that particular changes in concentrations of specific blood proteins correspond to a specific state, which is then associated with a given category.

State-Specific versus Individual-Specific Markers

An example if a state-specific marker is shown in Figure 1. Each "+" represents, for example, the blood concentration of a particular biochemical. The individuals in the left column are in a specific disease state, while those in the right column are not and are therefore considered to be in a healthy state, at least with respect to this disease.

Individuals in each state have different blood concentrations of this biochemical due to genetic and environmental differences between individuals and any experimental uncertainty in the measurement. What is clear is that the range of concentrations for individuals in the disease state is significantly higher than for those not in this state. Such a marker can be used to classify the individuals into three groups; they are in the disease state if the blood concentration is above an upper threshold, they are not in this disease state if the concentration is below a lower threshold, and they are undetermined if the blood concentration is between these thresholds.

Since the values of many features are known for each individual, it is possible to construct classifiers using two or more features. An algorithm would search through sets of two or more features to find a set that optimally classified a given set of individuals, which is known as a training set. The goal is to maximize the number of correctly classified individuals, so if two features are used and one is that shown in Figure 1, the second feature would try to correctly resolve those in the undetermined region without upsetting the correct classification of the other individuals. Therefore, the action of this second feature in the classifier is to specifically act on those individuals in the undetermined region. The first feature (Figure 1) therefore is a state-specific marker since its intensity is largely controlled by the state of the individual, while the second feature is individual-specific since it only acts on those individuals who have an intensity of the first feature in the undetermined region.

This argument suggests that statistical methods which find features that are significantly different in magnitude depending upon an individual's state is all that is needed to find any state-specific markers. These independent markers can be found if both the healthy and diseased categories are represented by a single state. Figure 2a displays a situation where the diseased category (shown in red) is actually composed of two states (D1 and D2). This can only be seen through the action of a concerted pair of markers, Marker 1 and Marker 2. State D1 has a high intensity in Marker 1 while State D2 has a high intensity in Marker 2, while the healthy individuals have a low intensity in both features. Figure 2b shows intensity plots for these two markers under the assumption that there is a single diseased state and a single healthy state. It is questionable whether a given statistical method would find the difference in the intensities of these features significant. Only by correctly distinguishing the state if each individual can one see that Marker 1 is a good classifier for State D1 and Marker 2 for State D2 (Figure 2c).

Figure 1: Values of a state-specific marker for individuals in a disease state (left) and in a healthy state (right).

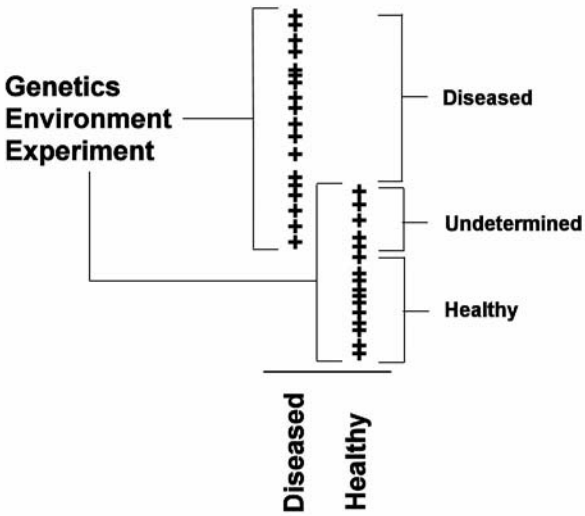
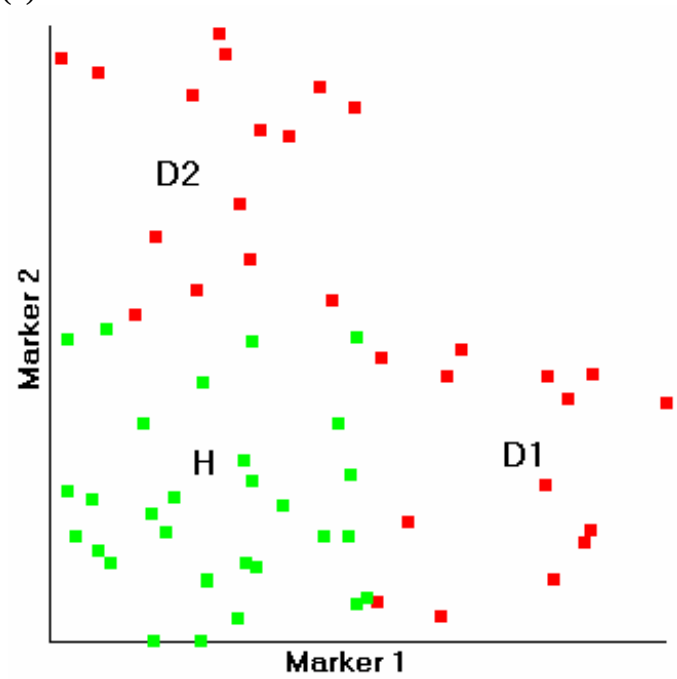
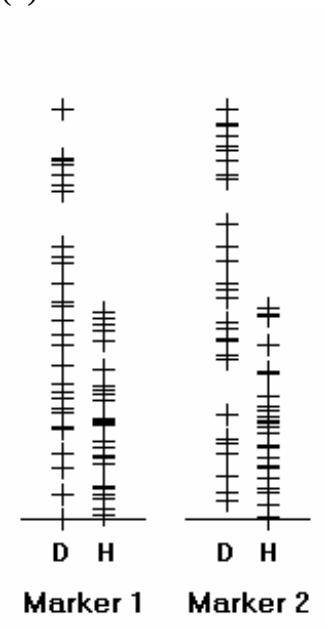


Figure 2: (a) Scatter plot for diseased (red) and healthy (green) individuals using the values of a pair of correlated markers; (b) Values for these markers assuming one state for each category; and (c) Values for these markers assuming two disease states and one healthy state.

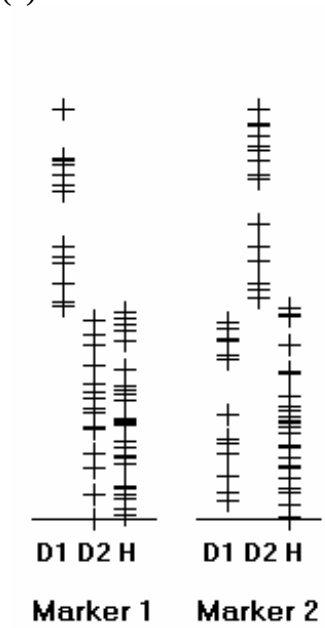
(a)



(b)



(c)



(Last updated 4/4/07)