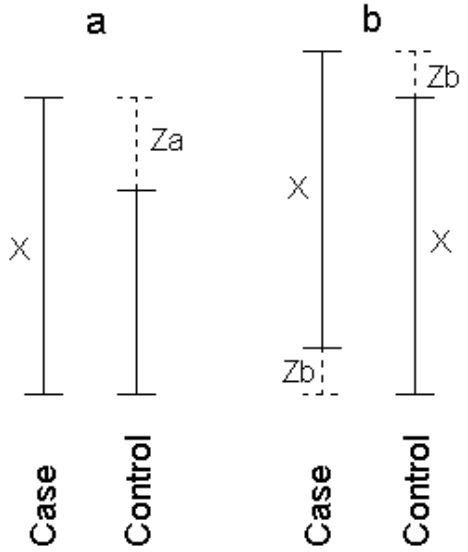
Identifying Putative Biomarkers

Brian T. Luke (*lukeb@ncifcrf.gov*)

This section describes the 10 methods currently employed to identify putative biomarkers, where a putative biomarker is defined as a feature whose intensity can distinguish some or all of the subjects in one State from those in another. In other words, the intensity ranges for the samples in each State should be different; the larger this difference the stronger the marker. Two examples of a marker are shown in the figure at the right.



In Feature-a, the Cases have an intensity that span a total range of X while the Controls have a range that is smaller by an amount Za. In Feature-b, both the Cases and Controls have a range of X, but the range for the Cases is shifted higher than the range for the Controls by an amount Zb. If there are equal numbers of Cases and Controls, the maximum range (X) is the same for both markers, and there is a uniform distribution of intensities in each State, the same number of samples will be distinguishable as long as Za=2Zb. For example, if X=100 and Za=20 (Zb=10), then 10% of the samples should be distinguishable. In Feature-a, 20% of the Cases should have an intensity that is larger than any of the controls. In Feature-b, 10% of the Cases will have an intensity above all Controls and 10% of the Controls will have an intensity below all Cases. Since in this example only 10% of all samples should be correctly distinguishable, this would be considered a very weak marker. As Za and Zb increase, the marker becomes stronger since it is able to correctly distinguish more of the samples.

For each of the methods described below there will be a table that examines the ability to detect for weak markers as a function of Za (2Zb) and the number of Cases and Controls. For this examination, X is set to 100 and Za is varied from 10 to 40 (Zb from 5 to 20). The first step is to estimate the maximum possible value that a given method can achieve for a feature that contains no information (Za=Zb=0) for a given number of samples. This is done by examining 10,000 randomly generated features with intensities between 0.0 and 100.0. Then for each value of Za or Zb, 10,000 new features will be randomly generated and there will be a count of the number of times a feature has a score that is better than that obtained from the features with no information. Part of this table for the *dtgini* procedure is shown below.

Each	Score	10a	10b	15a	15b
30	0.333	1	1	8	1
45	0.390	12	5	38	9
60	0.410	8	2	81	6
90	0.431	13	0	297	10
150	0.462	952	34	6778	318
300	0.483	9892	3746	10000	9634

The first column lists the number of Cases and number of Controls in these artificial features. The second column lists the minimum score ($GINI_{split}$) obtained from 10,000 features with no information, and is an estimate of the minimum value one may expect. It should be noted that this value increases as the number of Cases and Controls increases, meaning that as the number of samples increases it becomes harder to get a meaningful separation of the samples using a one-node decision tree. The third column builds feature intensities Feature-a in the figure above with Za=10. Since this is a weak marker, only one of the 10,000 features produced a $GINI_{split}$ that was less that the best result obtained from features with no information for 30 Cases and 30 Controls. If there are 300 Cases and 300 Controls, almost all of the features have a score that is better than was obtained from features with no information. Conversely, the fourth column shows that if 2Zb=10, the number of features that are better than random is significantly reduced; the 10,000 features with 300 Cases and 300 Controls only found 3746 with $GINI_{split}$ values that are less than random.

A terse description and links to the 10 methods that examine the features for putative biomarkers are as follows.

1. *catboot* (formerly known as *fqual* [Hab-05]) This method performs a Bootstrap analysis and determines the centroids for the remaining samples in each State. A distance-dependent K-Nearest Centroid algorithm is used for the classification of the removed samples, where K is the number of States in the dataset.
2. *student* (formerly known as *impf* [Hab-05]) This method performs a Student t-test to test for independence distributions for Cases and Controls.
3. *dtgini* (formerly known as *fqual* [Hab-05]) This method examines each feature by using it in a single node decision tree using the GINI Index to determine the optimum cut point.
4. *dtinfg* (formerly known as *infgl* [Hab-05]) This method examines each feature by using it in a single node decision tree using the Information Gain to determine the optimum cut point.
5. *nnfeat* This method uses each feature to construct a Feed-Forward Back-Propagation Artificial Neural Network. Each network has a single input node (the feature's intensities), two processing nodes in the hidden layer and a single output node.

6. *chisq* This method determines an approximate chi-square value by dividing the total intensity range into regions by requiring that there are at least five expected Cases and Controls in each region.
7. *kruswal* This method performs a Kruskal-Wallis one-way analysis of variance using the ranks of the intensities for samples in each State.
8. *kolsmir* This method performs a Kolmogorov-Smirnov test (K-S test) to measure the maximum difference in cumulative fraction plots for the Cases and Controls.
9. *extreme* This method measures the maximum number of samples from a given State at either extreme of the intensity distribution.
10. *vartest* This method is derived from the relevance index of Yip and coworkers [Yip-03]. It finds features with a minimum intra-State variance relative to the total variance.

Conclusions contain overall points relating to the results from all 10 methods of examining features.

(Last updated 4/4/07)