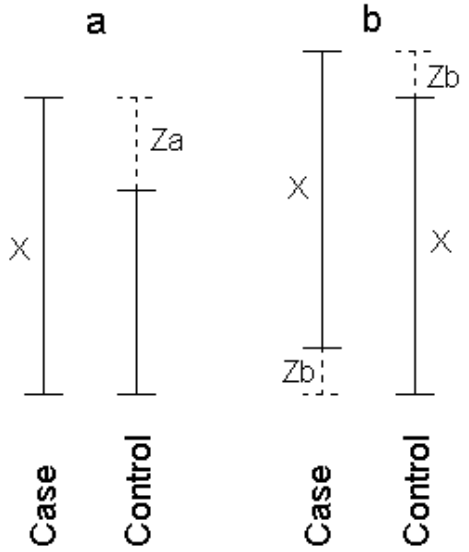
Comparison of Analysis Method

Brian T. Luke (lukeb@ncifcrf.gov)

As stated in Identifying Putative Biomarkers, a total of 10 methods are available in the BioMarker Development Kit (BMDK) to search a dataset of feature intensities for putative biomarkers. Included in the description of each method is an examination of the minimum strength a feature must have to have a 50% probability or better of obtaining a score that is better than a feature with no information. Two forms of a putative biomarker are examined, as shown in the figure to the right. For a feature of type Feature-a, the relative strength of the putative biomarker is determined by Za, which represents the extent to which the maximum intensity for samples in one category exceeds that of the other. For example, if Za=30, the range of intensities for one category is only 70% that of the other, or approximately 85% of all samples have another sample of a different category with the same approximate intensity.



The following table lists the minimum value of Za needed for a peak of type Feature-a to have a 50% probability or higher of obtaining a better score than a non-informative peak as a function of the number of Cases and Controls (Each).

Each	catboot	student	dtgini	dtinfg	nnfeat	chisq	kruswal	kolsmir	extreme	vartest
30	60	55	50	40	55	65	65	60	45	50
45	50	40	35	35	50	45	55	45	40	45
60	50	35	30	30	45	40	45	45	30	35
90	45	30	25	25	35	35	35	35	20	40
150	35	30	15	15	30	30	25	25	10	30
300	25	20	10	10	20	20	20	20	10	20

For the largest dataset (300 Cases and 300 Controls), dtgini, dtinfg and extreme only require Za to be 10. This means that if one category has a maximum intensity that is 90% of the other so that 95% of all samples have intensity in the overlapped region. For the smallest dataset (30 Cases and 30 Controls) dtinfg requires that one category have a maximum intensity that is 60% of the other while catboot and kolsmir require that it be 40% or less.

For putative biomarkers of type Feature-b, the following table lists the minimum value of 2Zb needed to find at least 50% of Feature-b peaks with scores better than non-informative peaks as a function of the number of Cases and Controls (Each).

Each	catboot	student	dtgini	dtinfg	nnfeat	chisq	kruswal	kolsmir	extreme	vartest
30	95	55	85	70	70	100	80	95	75	60
45	80	50	60	60	60	65	65	70	55	60
60	75	45	55	50	55	55	50	70	55	45
90	60	35	45	40	40	40	40	50	35	45
150	50	30	25	25	35	35	30	40	15	30
300	30	20	15	10	25	20	25	30	10	20

For the largest dataset dtinfg and extreme have at least 50% probability of identifying a putative biomarker if there is a 95% overlap in the ranges of intensity for the two categories; again with 95% of all samples having intensity in the overlapping region. For the smallest dataset, student has at least a 50% probability of identifying the feature if there is at most a 72.5% overlap in the intensity ranges while for chisq this overlap can be at most 50%.

The major goal of this exercise is to demonstrate that there is a limit to the level of detection of a putative biomarker using any of these methods. For the largest dataset examined, at least 5% of the samples (30 samples) must have an intensity value in a range not covered by samples in the other category. As the sample size decreases, the fraction of samples that must have intensity values in a range not covered by the other category increases. This means that if a single category is composed of multiple States, and at least one of the States contains a small fraction of the samples, identifying a marker for this State may only be likely if the total number of samples in this category, and therefore the number of samples in this State, is reasonably large.

It should be stressed that these limits of detection are only approximate. In some cases one value of Za or 2Zb finds slightly below 50% of the features. Since the tests are only performed at 5% increments, the minimum level shown in these tables identifies significantly more than 50% of the features. In addition, these artificial features are produced using a random number generator with a uniform distribution. If the distribution of intensity values is different from uniform, it may be harder or easier to identify a putative biomarker. For example, if the distributions of intensity values have a Gaussian or normal distribution about the mean, finding a putative biomarker may be considerably harder. Conversely, a putative biomarker would be easier to identify if it has an excess density of intensity values in the range not covered by samples in the other category.

(Last updated 4/30/07)