Original Analysis of BPH and Healthy Datasets [Luk-07]

Brian T. Luke (lukeb@ncifcrf.gov)

Adam and coworkers [Ada-02] produced mass spectra for 78 individuals with benign prostate hyperplasia (BPH) and 81 individuals with healthy prostates. Each spectrum originally contained 48,538 intensities (ion counts) from m/z of -0.9991 to 198660. They were truncated to m/z of 1500 to 40,500 (17,686 intensities) and scaled to a constant total ion current. The sample from each individual was used to produce a duplicate pair of spectra.

52 BPH and 54 healthy individuals were randomly selected to form the training set; the remaining 26 BPH and 27 healthy individuals became the testing set. The 212 training spectra were examined using the peak picking procedure in the BioMarker Development Kit (BMDK) and 424 non-overlapping read-windows of sufficient intensity were identified. By requiring that the maximum intensity in each read-window not occur at the first or last two m/z values in this range for at least 60% of the BPH or healthy spectra reduced the number of read windows to 158. The maximum recorded intensity in each read-window was recorded, producing 158 feature intensities for each spectrum. Since conditional averaging and outlier detection are not commonly accepted procedures, the intensities for each feature in the duplicate spectra were averaged for all training and testing samples and no outlier detection was performed. The final dataset contained intensities for 158 features for a training set containing 52 BPH and 54 healthy individuals and a testing set containing 26 BPH and 27 healthy samples. This original dataset and a dataset using the same feature intensities with permuted sample labels (histologies) were examined using a preliminary version of BMDK [Hab-06], a decision tree (DT) classifier and medoid classification algorithm (MCA) classifier.

Correct Sample Labels (Histologies)

When the training data with correct sample labels was examined using the preliminary version of BMDK [Hab-06], 15 features were identified. Four features had highly correlated intensities producing 12 feature groups. When these features and all correlated features were visually examined, five feature groups had one or more well-defined peaks. If two or more peaks were present in a single group, the peak with the largest maximum intensity was selected. The intensity distributions for the five peaks selected as putative biomarkers are shown in Figure 1. For each peak, the intensities for the 52 BPH training samples are in the left column and the intensities for the 54 healthy training samples are in the right column. The number above the column of intensities is the maximum intensity observed in the training set, and the number immediately below is the minimum intensity. Below the intensity plots for each putative biomarker are the *m/z* values representing the center of the read-window selected during peak picking and are stored as the feature label in the datasets. It should be noted that the putative biomarker at m/z of 3321.7 only has a different intensity for a small number of BPH training samples. Either this is a poor putative biomarker or it identifies a small number of BPH samples that are in a different State from the rest of the samples.

When these putative biomarkers are examined to form the best distance-dependent 6-nearest neighbor classifier using one, or two peak intensities, the results listed in Table 1a are obtained. The peaks at *m/z* of 8942.8 and 4072.8 represent the two best 1-peak classifiers but only the third best 2-peak classifier when a standard Euclidean distance is used. The main reason for this is the large difference in maximum intensity between these peaks, and when the Euclidean distance uses a relative difference in peak intensities it becomes the best 2-peak classifier. The peak at 8942.8 has been previously identified as the protein fragment C3a-desArg and using the procedure developed earlier [Hab-06] an individual would be classified as healthy if the intensity of this peak was 89.0 or less, having BPH if the intensity was above 138.0 and "Undetermined" if the intensity was between these values. This produces an overall sensitivity of 88.6% and a specificity of 93.7%, but 15.7% of all individuals would be "Undetermined" and a different test would have to be used (Table 1b).

An exhaustive search of all symmetric 3-node decision trees on the training set found 12,344 unique trees with a sensitivity and specificity of 94% or higher. It is interesting to note that 1653 trees did not use any of the nine features associated with m/z of 4072.8 and 8942.8. The best 3-node decision tree had a sensitivity and specificity of 98.1% for the training set, but the sensitivity was only 81.5% for the testing set. This suggests that a necessary piece of the underlying fingerprint was missing from this decision tree (i.e. either an incorrect feature was used or the optimum cut point was not identified at one or more decision nodes). Using all data in the DT procedure produced the results in Table 2a. The first run converted a decision node into a terminal node only if it contained samples from a single category. It found 2000 6-node trees with a minimum sensitivity and specificity of 98.7 and 97.5%, respectively, and there sum was always 197.5 or higher. The second run converted a decision node into a terminal node if it contained at most a single BPH or healthy sample. It found 10 4-node and 1990 5-node trees with a minimum sensitivity and specificity of 96.2 and 96.3%, respectively, and their sum was always 195.0 or higher. Comparable results were found when a decision node was converted to a terminal node if it contained two or less BPH or healthy samples.

When the training data was examined using the MCA procedure, 2000 classifiers using five features produced sensitivity=specificity=100% for the training data. 100 of them also produced a perfect result for the testing set, but %undetermined varied from 20.8% to 60.4% (average = 38.0%). Only 19 classifiers gave %undetermined=0% for the testing data, but the sensitivity was 50% or less in all 19. On average, 33.9% of the testing samples received an "Undetermined" classification across the 2000 classifiers, so coverage is a problem. This means that if there is an *a priori* division of the available samples into a training set and a testing set, thereby producing a truly "blinded" test, it is extremely likely that a large fraction of the testing set's classifications will be "Undetermined".

When all data was placed into the training set but the calculation was limited to not allow more than 52 BPH-cells and 54 healthy-cells, the results in Table 3a are obtained. For classifiers using five or six features the processing of the samples occurred in two

different orders. The first (Original Order) processed the BPH training set, the healthy training set, the BPH testing set and then the healthy testing set, while the second (Inverted Order) inverted the order of the datasets. Each ordering was run twice using different seeds to the random number generator. All four runs found at least 2000 6-feature classifiers with a sensitivity and specificity of 100% and no "Undetermined" classifications.

Permuted Sample Labels (Histologies)

To examine the significance of the results obtained for these datasets, the sample labels (histologies) were permuted independently among all training samples and then all testing samples. Re-running the search for putative biomarkers in the training set using the preliminary version of BMDK identified 30 features. Correlation of intensities among these features causes 22 feature groups to be identified. All other features that correlate with these feature groups were collected, and instead of visually inspecting all features to find well defined peaks, the feature with the highest maximum intensity in each group was selected. This resulted in a set of 21 features since one feature correlated with features from two different groups and it had the highest intensity among all features in both groups. An intensity plot of these 21 features is shown in Figure 2, and it is clear that none of them represent as strong of a putative biomarker like as the 8942.8 peak shown in Figure 1. Instead of finding the best distance-dependent 6-nearest neighbor classifiers for the training set and then examining the testing set only the training set was examined. The order of the training samples was scrambled and a 10-fold cross-validation was run. This process was repeated a total of five times and the resulting sensitivities, specificities, and %undetermined are shown in Table 1c. For the training and testing sets from each cross-validation run a distance-dependent 6-nearest neighbor classifier was used and the best results using one, two and three features are listed in this table. A comparison of Tables 1a and 1c clearly shows that permuting the sample labels (i.e. removing any biological information) strongly degrades the quality of the final classifier. The best 1-peak classifier had an average sensitivity and specificity below 64% and no dramatic improvement was found using any combination of two or three peaks.

When the DT procedure was used, a run using only the training set found 200 trees with an average sensitivity and specificity above 88%, but the testing set results were generally not better than a random guess. Though this was expected, one could argue that this was due to an incomplete coverage in the training set. To accommodate for this potential lack of coverage, the search for good DT classifiers was repeated using two different procedures. The first combined all samples into a single dataset and the other determined the quality of the classifier by averaging the sensitivities and specificities of the training and testing sets and then adding these averages together. Both procedures were run eight times using different seeds to the random number generator and the results are listed in Table 1b. This table lists the ranges of the average sensitivity and specificity across the top 200 decision trees. Combining all data consistently produced sets of 200 decision trees with an average sensitivity and specificity of 81% or better. Averaging the sensitivity and specificity of the training set with the sensitivity and specificity of the

testing set produced one run where the average sensitivity and specificity varied from 86.3 to 85.4% across the top 200 decision trees.

The results in Table 1b only converted a decision node into a terminal node if it did not contain any cases or controls. To study the effect of pruning, all samples were run and a decision node was converted to a terminal node when it contained at most zero, one or two BPH or healthy individuals. These results for all 2000 unique decision trees in the final population are shown in Table 1b. Converting to a decision node to a terminal node when a decision node had one or two BPH or healthy individuals caused the average sensitivity and specificity to be 82.3% or higher across all 2000 decision trees.

When the MCA procedure was used to examine the datasets with permuted sample labels, the results in Table 3b were obtained. Between five and eight features were used in the MCA classifier with the requirement that there are at most 52 BPH cells and 54 healthy cells. For each number of features the training set was examined before the testing set (Original Order) and then the testing set was examined before the training set (Inverted Order). For each ordering the search for good classifiers was run twice with different seeds to the random number generator. The results show the sensitivity and specificity for the best and 100th best classifier in the final population of unique classifiers. These results show that a sensitivity and specificity of 94.9 and 95.1%, respectively, was found for a 6-feature classifier. An improvement was found when the number of features was increased to seven (sensitivity=98.7, specificity=93.8%) or eight (sensitivity=96.2, specificity=97.5%) features.

Conclusions

These results show that coverage is a real problem for fingerprint-based classifiers. The DT analysis of the original datasets showed that better results were obtained when all samples were analyzed together and this was most likely due to improved detection of the correct features and cut points to produce the optimum fingerprints. Coverage was particularly important for the MCA method. If an *a priori* division of the samples into a training set and a testing set is performed, it is extremely unlikely that good results will be obtained for the testing set without finding any samples with an "Undetermined" classification.

Uniqueness is also a problem for fingerprint-based classifiers. When all samples were examined both the DT and MCA classifiers found a large number of unique classifiers that produced a good result. In contrast, a biomarker-based classifier only produced a small number of accurate classifiers (Table 1a).

Finally, significance is a problem for fingerprint-based classifiers but not a biomarker-based classifier. Permuting the sample labels (histologies) caused a significant degradation of the classifier's accuracy for the training data when a biomarker-based classifier is used (Table 1c). In contrast, both the DT (Table 2c) and MCA (Table 3b) classifiers produced fairly accurate results. Suggesting that accurate results must be due to some underlying biological factor is clearly incorrect for a fingerprint-based classifier.

Figure 1: Intensity plots of the five putative biomarkers identified [Luk-07] by the preliminary version of BMDK [Hab-06].

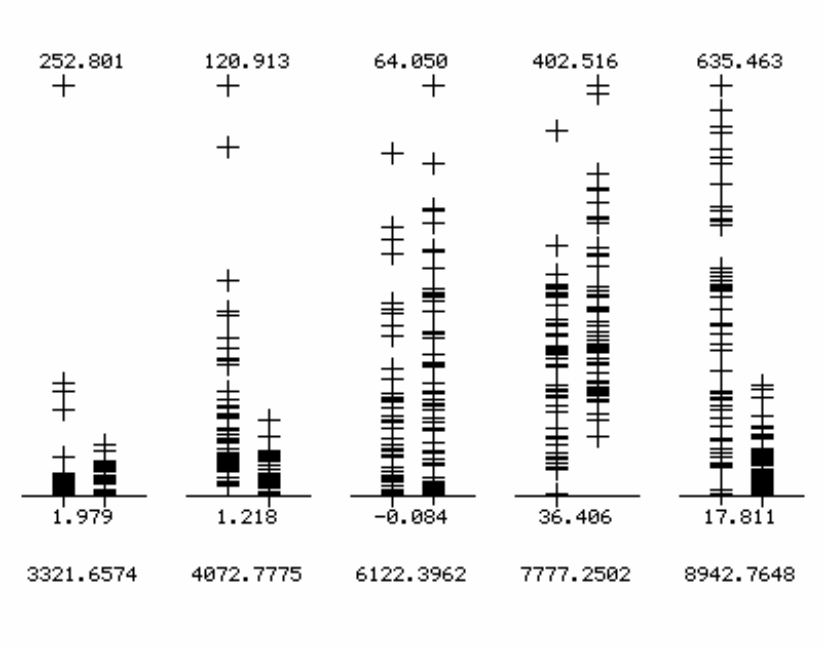

Figure 2: Intensity plots for the 21 features selected from the training data after the sample labels (phenotypes) were permuted.

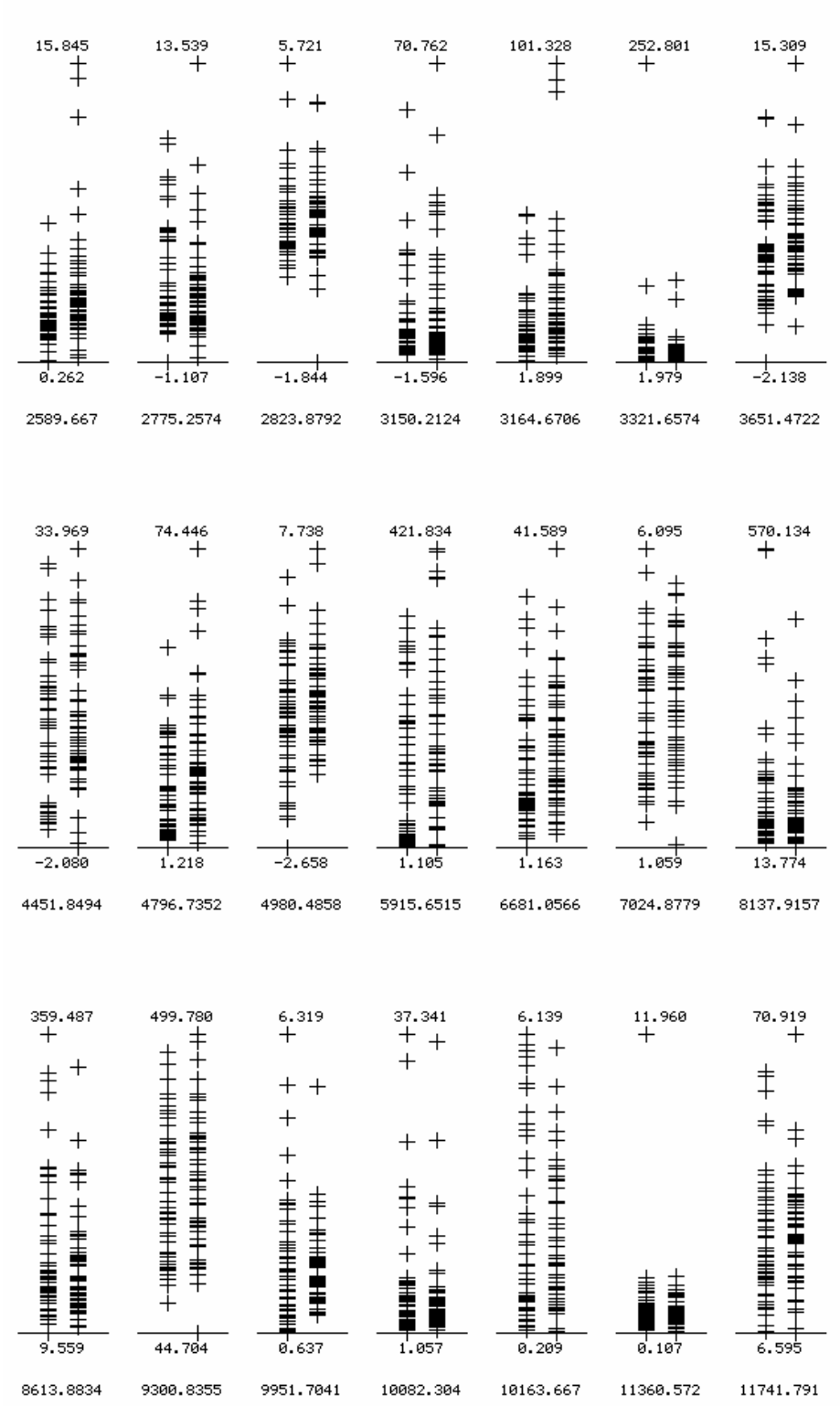

Table 1: Results of the classification of BPH versus healthy samples: (a) Best 1- and 2-feature, 6-neighbor DDKNN results using only the five putative biomarkers shown in Figure 1 for datasets with true sample labels (phenotypes); (b) Predicted results using the decision rules BPH if I(8942.8)>138.0, healthy if I(8942.8)<89.0, otherwise undetermined; and (c) The best 1-3 feature results for the training data with permuted sample labels (phenotypes), using 5 runs of 10-fold cross-validation.

(a)

	Training Data				Testing Data		
Features	Sens	Spec	Undet		Sens	Spec	Undet
8942.8	75.0	79.6	0.0		84.6	74.1	0.0
4072.8	71.2	77.8	0.0		73.1	81.5	0.0
3321.7	56.9	63.0	0.9		46.2	63.0	0.0
7777.3, 8942.8	82.4	87.0	0.9		96.2	77.8	0.0
3321.7, 8942.8	74.0	90.7	1.9		88.5	88.9	0.0
4072.8, 8942.8	75.0	87.0	0.0		88.0	81.5	1.9

(b)

Group	Sensitivity	Specificity	Undetermined
Training	84.8%	93.5%	13.2%
Testing	95.8%	94.4%	20.8%
Overall	88.6%	93.7%	15.7%

(c)

	Training Data				Testing Data		
Features	Sens	Spec	Undet		Sens	Spec	Undet
9951.7	60.1	62.1	2.0		61.6	65.9	1.9
3164.7, 4451.8	66.6	57.4	3.9		66.0	59.4	4.3
2823.9, 3164.7, 4451.8	66.3	61.5	4.1		66.0	63.2	4.5

Table 2: (a) Minimum (Sensitivity+Specificity), Sensitivity, and Specificity for the top 2000 decision trees using all samples, along with the size of the trees and the features that are regularly used when the tree is terminated if the number of BPH or healthy samples is at most 0, 1 or 2; (b) Results for seven-node decision trees using permuted phenotypes showing the average quality (%) of the classification for the best and 200th best solution for eight different runs when either all samples are used in the training (All) or when the score is the average of the sensitivity and specificity for both the training and testing sets (Sum); and (c) Minimum (Sensitivity+Specificity), Sensitivity, and Specificity for the top 2000 decision trees using all samples with permuted sample labels, along with the size of the trees when the tree is terminated if the number of BPH or healthy samples is at most 0, 1 or 2.

(a)

NSTOP	Minimum (Sens+Spec)	Minimum Sensitivity	Minimum Specificity	Nodes (# Trees)
0	197.5	98.7	97.5	6(2000)
1	195.0	96.2	96.3	4(10), 5(1990)
2	195.0	96.2	97.5	4(4), 5(1996)

(b)

Run	All	Avg
1	81.51 to 81.03	83.44 to 82.55
2	83.55 to 81.36	84.35 to 82.98
3	83.55 to 82.93	86.31 to 85.36
4	83.64 to 81.16	85.42 to 83.97
5	82.93 to 82.31	84.85 to 83.01
6	83.52 to 82.88	83.35 to 82.87
7	84.26 to 83.07	86.25 to 85.29
8	83.48 to 82.29	84.28 to 83.35

(c)

NSTOP	Minimum (Sens+Spec)	Minimum Sensitivity	Minimum Specificity	Nodes (# Trees)
0	162.2	73.1	79.0	7(2000)
1	164.7	70.5	84.0	7(2000)
2	164.6	73.1	86.4	6(1), 7(1999)

Table 3: (a) Number of 5- and 6-feature medoid-based classifiers and that produce 100% correct classifications when all data is used with the caveat that there cannot be more than 52 BPH cells and 54 healthy cells for the dataset with correct sample labels; and (b) Sensitivity and specificity of the best and 100th best medoid classifier for each of two runs where the number of features (#F) was varied from five to eight and the classifier was required to have at most 52 BPH cells and 54 healthy cells after the sample labels are permuted.

(a)

Original Order			Inverted Order	
#Features	#Classifiers		#Features	#Classifiers
5	45		5	55
5	43		5	55
6	2000		6	2000
6	2000		6	2000

(b)

	Original Order				Inverted Order			
#F	**1st Classifier**		**100th Classifier**		**1st Classifier**		**100th Classifier**	
	Sens	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**
5	98.7	87.7	94.9	85.2	85.9	98.8	87.2	92.6
5	98.7	87.7	93.6	86.4	85.9	98.8	87.2	92.6
6	97.4	92.6	96.2	88.9	91.0	98.8	88.5	96.3
6	100.	87.7	97.4	87.7	94.9	95.1	89.7	95.1
7	100.	91.4	97.4	90.1	94.9	97.5	89.7	98.8
7	98.7	93.8	97.4	90.1	92.3	100.	93.6	93.8
8	100.	93.8	94.9	96.3	96.2	97.5	93.5	96.3
8	97.4	96.3	98.7	91.4	96.2	96.3	92.3	97.5

(Last updated 5/2/07)