

# **Analyzing the Peak Data: Some details for the manuscript “Serum Fingerprinting Coupled with a Pattern Matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men“**

by Bao-Ling Adam, Yinsheng Qu, John W. Davis, Michael D. Ward, Mary Ann Clements, Lisa H. Cazares, O. John Semmes, Paul F. Schellhammer, Enrique Dalmasa, Yutaka Yasui, Ziding Feng, and George L. Weight, Jr.

The analysis has three goals:

1. Finding potential markers for distinguishing the three groups,
2. Finding interactions between markers, and
3. Constructing an accurate classifier for diagnosis of prostate cancer.

1. Choosing the relevant features using the area under the receiver operating characteristic curve(ROC).

For the first goal, we evaluate the discriminatory power of each peak for distinguishing between normal and PCA, BPH and PCA, and normal and BPH by estimating the area under the ROC curve (AUC), which goes from 0.5 (no discriminating power) up to 1.0. The AUC statistic can be interpreted as the probability that the test result from a diseased individual is more indicative of disease than that from a nondiseased individual (Pepe, M.S., 2000). In Table 1 we listed 10 peaks with the highest AUC for separating normal from PCA, 10 peaks with the highest AUC for separating normal from BPH, and 10 peaks with the highest AUC for separating BPH from PCA. To be able to separate two groups completely

with a single peak, its AUC should be 1. Such kind of peaks has not been found in this analysis. Therefore using multiple peaks is necessary.

The peaks with AUC close to 0.5 are irrelevant for classification. In our next step analysis, we ignore the peaks with AUC statistics below 0.62. Only 124 of the 779 peaks have AUC above 0.62.

## 2. Decision tree

In the next step, we try to separate 326 training samples, including 82 normal, 167 cancer and 77 BPH with a decision tree, using the 124 peaks as predictors.

### 2.1 Constructing a decision tree

The algorithm for constructing a decision tree consists of (1) selecting a peak and a cutpoint such that the training samples can be splitted into two subgroups (nodes), each of them is as homogeneous as possible, and (2) do the same for each of new born nodes. For example, the decision tree in Figure 1 begins with the whole training set (the root). The first peak selected located at the mass 7819.75Da, and the cutpoint is 0. The answer to the question “whether the peak at 7819.75 has intensity less than or equal to 0” splits the root into two nodes, “yes” goes to the left, and “no” goes to the right. Note that most cancers go to the right, and most BPH go to left.

For each node we defined a cost function reflecting the heterogeneity of the node and seek for the peaks and cutpoints to reduce the total costs in the two splits. We use the negative log likelihood of the multinomial distribution as the cost function. Let  $c$  be the number of classes, for a given node  $p_j$  is the probability of class  $j$ , and  $n_j$  is the number of samples in class  $j$  ( $j = 1, \dots, c$ ), the cost function for this node is

$$-\log L = -\sum_j n_j \log(p_j),$$

where  $p_j$  is evaluated by  $n_j/n$ , where  $n = \sum_j n_j$ . In our case, the number of classes  $c$  is 3. The cost for the root node is 336.00. Among the 124 candidates, we selected the peak at mass 7819.75 with cutpoint 0, because this selection reached the maximum reduction of cost in the two descendant nodes. The costs for the left and the right descendants are 73.33 and 128.00, respectively. The total is 246.33. The cost reduction, or, the information gain, for this split is  $336.00 - 246.33 = 89.67$ . The left and the right descendants are splitted further. This recursive partitioning procedure stops when each node containing only one class samples, or further splitting has no gain. The nodes without further splitting are called the terminal nodes, or, leaves, denoted by rectangles in the figure. This tree contains 10 leaves from L1 to L10. The classification rule is simple. For example, if an unknown sample has no peak at mass 7819.75 but has a peak at mass 7024.02, then this sample falls in leaf L1. In the training set there are 1 normal, 27 PCA, and 1 BPH in this category. The majority is PCA, so we will classify this new sample to PCA. Likewise, if a sample falls to the category of L2, it will be assigned to BPH. For detailed description of the tree method see Breiman, Friedman, Olshen, and Stone(1984).

## 2.2 A probabilistic interpretation of the decision tree

Deterministic classification rules ignore the stochastic nature of reality. Even for L2, which contains only BPH samples in the training set, the possibility for normal or PCA can not be ruled out for a new sample. A better way is to calculate the expected probability for each of the three classes for each leaf.

We use a Bayesian approach for calculating the expected probabilities and their credible intervals. The numbers of observations in the 3 classes  $(n_1, \dots, n_c)$  follow a multinomial distribution with parameters  $(p_1, \dots, p_c)$ . Naturally, we use a Dirichlet distribution with hyper parameters  $Dirichlet(\alpha_1, \dots, \alpha_c)$ . The posterior distribution for the probabilities  $(p_1, \dots, p_c)$  given the data  $(n_1, \dots, n_c)$  is a Dirichlet distribution with parameters  $(n_1 +$

$\alpha_1, \dots, n_c + \text{alpha}_c$ ). In our case  $c = 3$ , and the probabilities for normal, PCA, and BPH, are denoted by  $p_1$ ,  $p_2$ , and  $p_3$ , respectively. Thus, for each node, the expected probabilities for the  $j$ th class is simply  $(n_j + \alpha_j)/(n + \alpha)$ , where  $n = n_1 + n_2 + n_3$  and  $\alpha = \alpha_1 + \alpha_2 + \alpha_3$ . We choose  $\alpha_1 = \alpha_2 = \alpha_3 = 1$ . Thus, for the node L2, the expected probabilities for normal, PCA, and BPH, are 0.0167, 0.0167, and 0.9667, respectively. Although, there is no observation of normal and PCA in the data, the expected probabilities  $p_1$  and  $p_2$  are not zero. The prior let us more conservative (The maximum likelihood estimates for  $p_1$  and  $p_2$  are  $n_1/n$  and  $n_2/n$ , both are zero).

To calculate the 95% credible interval we simply generate 4,000 samples for the posterior Dirichlet distribution, and sort them and take the 100th and the 3900th samples as the lower and the upper bounds of the 95% credible interval. Table 2 shows the expected probabilities and their 95% credible intervals for 8 terminal nodes in the decision tree.

### 2.3. Interpretation and classification

A decision tree provides information on the interaction of the potential markers. For instance, our decision tree shows, among other things, that if no peaks at mass 7819.75 and mass 7024.02 and mass 5382.13 (L2), then most likely, this is a BPH with expected probability equal 96.67% (the 95% credible interval is between 90.72% and 99.52%). This fact may lead to some meaningful discovery.

The decision tree can also serves as a classifier. In an independent testing set with 15 normal, 15 BPH and 30 PCA, the number of misclassified samples are 0 for normal, 1 for BPH and 5 for PCA. The overall misclassification rate in the maximum likelihood sense is 10%. The expected misclassification probabilities and their 95% credible intervals calculated from posterior distributions for the training set and the total samples (training plus testing) are shown in Table 3.

## REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees*, Wadsworth International Group, Belmont, California.

Pepe, M.S. 2000. Receiver operating characteristic methodology, *American Statistical Association*, 95, 308-311.

Table 1: Top ten list of the potential markers sorted by the magnitude of the areas under the ROC curves (AUC)

	normal vs PCA		normal vs BPH		BPH vs PCA	
	mass	AUC	mass	AUC	mass	AUC
1	9149.12	0.8394	7819.75	0.9562	7024.02	0.8981
2	3896.64	0.8377	9149.12	0.8781	7819.75	0.8800
3	6949.22	0.8144	4071.18	0.8532	9719.99	0.8469
4	5074.16	0.7952	6949.22	0.8487	7844.00	0.8400
5	9655.75	0.7851	8943.08	0.8366	7933.52	0.8083
6	8295.64	0.7758	3896.64	0.8159	9655.75	0.8018
7	4071.18	0.7704	7983.76	0.8145	6099.08	0.7964
8	8355.56	0.7677	7775.62	0.7841	7054.17	0.7961
9	7775.62	0.7434	5074.16	0.7796	6889.72	0.7752
10	4102.07	0.7428	7480.79	0.7738	9192.45	0.7692

Table 2: Expected probabilities and their 95% credible intervals estimated using 4,000 simulations from posterior distributions for 8 leaves in the decision tree

Leaf	Class	No. of Obs.	Expected Prob.	95% credible interval
L1	normal	1	0.0625	0.0081,0.1693
	PCA	27	0.8750	0.7423,0.9630
	BPH	1	0.0625	0.0087,0.1698
L2	normal	0	0.0167	0.0005,0.0584
	PCA	0	0.0167	0.0004,0.0628
	BPH	57	0.9667	0.9072,0.9952
L3	normal	1	0.2000	0.0247,0.4793
	PCA	5	0.6000	0.3027,0.8592
	BPH	1	0.2000	0.0248,0.4753
L4	normal	0	0.0714	0.0018,0.2509
	PCA	0	0.0714	0.0019,0.2579
	BPH	11	0.8571	0.6311,0.9823
L5	normal	0	0.1429	0.0040,0.4725
	PCA	4	0.7143	0.3557,0.9567
	BPH	0	0.1429	0.0040,0.4504
L6	normal	74	0.9494	0.8950,0.9858
	PCA	2	0.0380	0.0082,0.0879
	BPH	0	0.0127	0.0003,0.0459
L7	normal	0	0.0204	0.0005,0.0738
	PCA	46	0.9592	0.8893,0.9951
	BPH	0	0.0204	0.0005,0.0726
L10	normal	2	0.0337	0.0068,0.0784
	PCA	81	0.9213	0.8595,0.9674
	BPH	3	0.0449	0.0123,0.0964

Table 3: Misclassification rates and their 95% credible intervals estimated using 4,000 simulations from posterior distributions in training set and updated by testing set

	In training set		Updated by testing samples	
	error rate	95% Credible interval	error rate	95% Credible interval
normal	0.0586	0.0161, 0.1090	0.0502	0.0138, 0.0929
PCA	0.0295	0.0074, 0.0553	0.0499	0.0209, 0.0803
BPH	0.0773	0.0257, 0.1356	0.0747	0.0272, 0.1270