WACBE World Congress on Bioengineering 2007
Bangkok, THAILAND

Comparing 3 algorithms for Liver and bladder samples classification into Cancer or non-cancerous by the use of PCA

Last updated 4/20/07 by Nasser Abbasi

Accuracy result tables for the above 3 algorithms

These accuracy tables below where all generated using the following parameters:

- number of dominant components for 1 up to 5 (all where generated using the same random set per each trial run)
- Percentage of excluded number of samples from the set of samples to use to generate PHI vector (PCA) is 20% for all cases. In other words, 20% of the tumor samples where excluded from being used to generate tumor PHI, and similarly, 20% of the non-cancer samples where excluded from the set of non-cancer samples used to generate the normal PHI.
- Number of trials is 100.

Bladder data analysis

Algorithm 1: *Projection against tumor mode*. Accuracy of detection

Data set	Accuracy of detection of	One mode	Two modes	Three modes	Four modes	Five modes	Comments
Bladder Liver	Cancer	57.17 69.46	62.15 82.74	64.83 80.58	68.11 80.37	70.51 78.07	*Least accurate of all 3 algorithms. Accuracy improves with more modes added but remains least accurate. Tumor samples do not correlate positively as strongly with the tumor most dominant component.*
Bladder Liver	non-cancer	99.95 99.99	99.32 98.91	99.86 99.51	99.95 99.21	100 99.63	*Very good accuracy. Remains very good but become slightly less accurate as more modes are added. Normal samples correlate negatively very strongly with the tumor most dominant component.*

Algorithm 2: *Projection against normal mode*. Accuracy of detection.

Data set	Accuracy of detection of	One mode	Two modes	Three modes	Four modes	Five modes	Comments
Bladder Liver	cancer	80.35 81.47	77.35 78.44	73.20 81.30	69.23 82.61	70.26 80.96	*More accurate than using the first algorithm, but accuracy now decreases as more modes are added. But this change of accuracy is not consistent. Notice also that the accuracy decreased more with the bladder data than it did with the liver data.*
Bladder Liver	non-cancer	100 100	99.50 96.41	94.32 95.11	93.59 93.28	91.59 90.68	*This gives the most accurate result for detection of non-cancerous samples Normal sample correlate positively very strongly with the normal dominant component. But only one mode must be used. Adding more modes made the result less accurate*

Algorithm 3: *Projection against combination mode*. Accuracy of detection

Data set	Accuracy of detection of	One mode	Two modes	Three modes	Four modes	Five modes	Comments
Bladder Liver	Cancer	82.35 80.75	82.97 88.54	83.30 87.15	83.81 89.82	84.25 89.54	***Most accurate method for cancer detection****. In addition, accuracy Increases as more dominant modes are added. Consistent result from both the Liver and Bladder data*
Bladder Liver	non-cancer	99.86 99.99	99.41 98.72	99.71 98.54	99.81 98.44	100 98.94	*Very good accuracy also obtained for the detection of non-cancerous samples*

Effect of changing the samples working set size on the accuracy of cancer and non-cancer detection

Observation on the accuracy of cancer detection

In this analysis, the effect of changing the size of the samples set used to generate the dominant component on the accuracy of both cancer and non-cancer detection is investigated.

Referring to the size of the set of samples, which is selected at random from the pool of samples, as the *working set size*, we decreased the working set size from 95% to 5% of the population by decrements of 1%, and for each change, the detection accuracy was recomputed.

This was done for both liver and bladder data sets. In each run, the accuracy of each of the three algorithms for detection was examined. We generated the following plots to analyze the effect of changing the working set size on the accuracy.

As a result of the above analysis the following was observed:

The accuracy of cancer detection, in both liver and bladder data, was least affected by changing the working set size when using the method of projecting against the non-cancerous dominant component.

The accuracy remained at the same level, but started to show slight deterioration as working wet size went down to about 20% of the normal samples population.

The overall accuracy went down by only 3% as the working set size was decreased all the way from 95% to 5% of the normal samples population size. This shows that the method of projection against the non-cancer dominant component is better able to handle smaller working set as the basis for generating dominant component.

When using the projection against the tumor dominant component method (recall that in above, we used the projection against the non-cancer dominant component), the results were different. We observe that accuracy of cancer detection, even though it remained fairly steady, it did fluctuate much more as the working set size is decreased.

An interesting phenomenon is observed when using the combination mode for measuring accuracy (algorithm three). In this case, we observe that as the working set size is decreased, accuracy of cancer detection improves. The accuracy was largest when the working set size was smallest (5% of the overall population).

Now we look at how the accuracy of non-cancer detection changed as a function of the working set size.

Conclusion

Tumor samples do not correlate positively as strongly with the tumor dominant component when compared to how strongly the normal samples negatively correlate with the tumor dominant component.

Tumor samples correlate much strongly, but in the negative sense, with the non-cancerous dominant component. Hence, when attempting to decide if a sample is cancerous or not, it is not recommend to measure the strength of the positive correlation with the tumor dominant component, but instead one should measure the strength of how negatively the sample correlates with the non-cancerous dominant component.

An analogy might help. To detect if one end of a magnet is a positive pole (cancer), it is better to move this end closer to a known negative pole (this is the non-cancerous dominant component) and measure how strongly it is being pulled in (negatively correlated) than to move it closer to a known positive pole (this is the cancerous dominant component) and measure how strongly it is being pushed away (positively correlated).

The situation with non-cancerous samples is different. Non-cancerous samples do correlate very strongly in the positive sense the non-cancerous dominant principle component. They also correlate very strongly in the negative direction with the cancerous dominant component.

From the above, we conclude that it is best to always use a non-cancerous dominant component to correlate a given sample against since a non-cancerous sample will exhibit a strong positive correlation, while at the same time a cancerous sample would exhibit a strong correlation but in the negative sense. In other words, both types of samples have stronger correlations with the non-cancerous dominant component when looking at the absolute magnitude of the correlation than the case would be if we have used a cancerous dominant component to correlate samples against.

A medical explanation of the above phenomena can be as follows: Non-cancerous samples (from the same region of the body) have a uniform and consistent level of gene expressions. Therefore correlating a non-cancerous sample to the dominant non-cancerous component will show a very strong positive correlation. At the same time, the level of gene expressions in a tumor sample from the same part of the body (primary cancer) will exhibit a strong negative correlation with this non-cancerous dominant mode.

However, cancer gene expressions do not seem to be as consistent and of uniform level among the cancerous samples used to generate the dominant component, even though all the samples are from the same region of the body (liver or bladder in this case) and primary cancer, and they are all the same type of cancer. There are more variations and differences among gene expressions within tumor samples taken from the same region of the body than there are variations between normal samples also taken from the same part of the body.

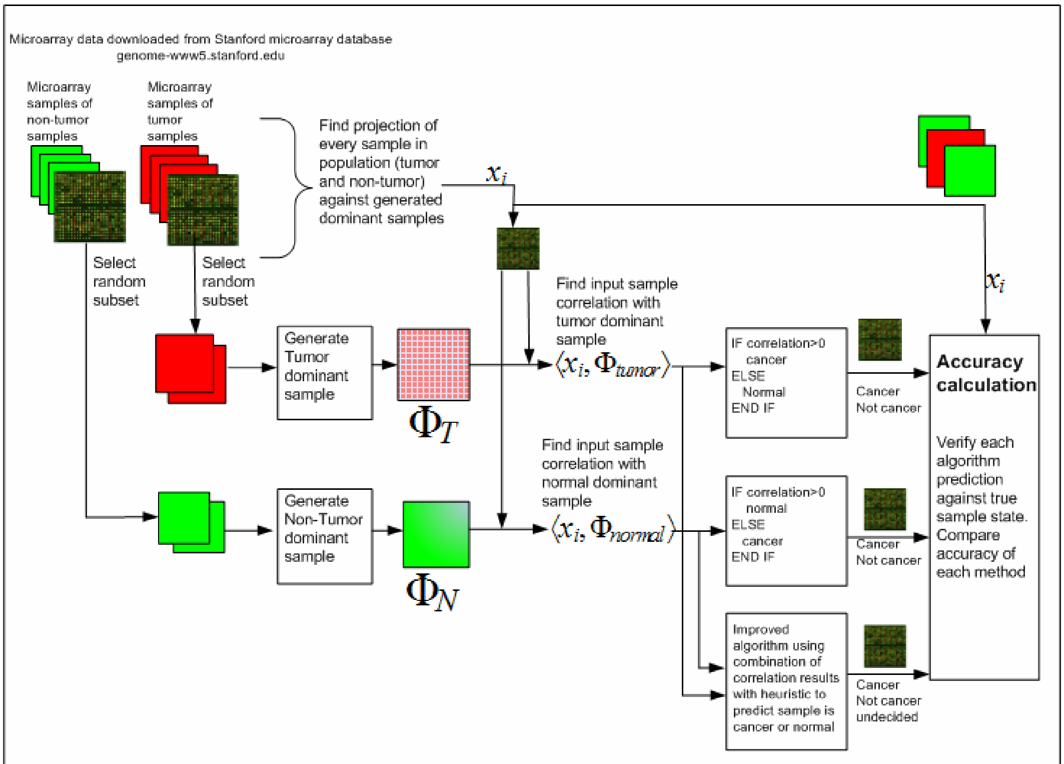
This is why correlating a cancerous sample against the cancer dominant component does not show as strong a positive correlation. One reason for these variations and differences among the gene expression of tumor samples taken from the same part of the body is that it seem to indicate that the tumor samples used where in different stages of growth, resulting if much more variation of gene expressions. Therefore, producing a dominant component that can exhibit all the main and prominent features of all the cancerous samples will be more difficult than the case is with non-cancerous samples. In other words, the more varied the samples, the harder it is to product a common sample which can exhibit the main distinguishing parts of all those samples. The task is easier to accomplish if there are less variations to start with among the samples.

The third algorithm introduces a heuristic algorithmic improvement in the detection of cancer. As a result of this improvement, we were able to improve cancer detection. However, since this improvement in detection is based on a heuristic improvement, more tests are needed against larger set of data.

Appendix

High level Workflow diagram

The following diagram illustrates the high level design of the software used in the analysis of this paper.



Microarray data downloaded from Stanford microarray database
genome-www5.stanford.edu
Microarray samples of non-tumor samples
Microarray samples of tumor samples
Find projection of every sample in population (tumor and non-tumor) against generated dominant samples
x_i
Select random subset
Select random subset
Generate Tumor dominant sample
Φ_T
Find input sample correlation with tumor dominant sample
$\langle x_i, \Phi_{tumor} \rangle$
Generate Non-Tumor dominant sample
Φ_N
Find input sample correlation with normal dominant sample
$\langle x_i, \Phi_{normal} \rangle$
IF correlation>0 cancer ELSE Normal END IF
Cancer
Not cancer
IF correlation>0 normal ELSE cancer END IF
Cancer
Not cancer
Improved algorithm using combination of correlation results with heuristic to predict sample is cancer or normal
Cancer
Not cancer
undecided
x_i
Accuracy calculation
Verify each algorithm prediction against true sample state. Compare accuracy of each method

High level algorithm used for determination of accuracy of cancer and non-cancer prediction based on the the use of Principle component analysis using 2 sets of data