Microarray Statistics Module 3: Clustering, comparison, prediction, and Go term analysis

Johanna Hardin and Laura Hoopes

Worksheet to be handed in the week after discussion

Name_____

Clustering algorithms: hierachical, K means, QT (hierarchical clustering is in BRB array tools; all three can be tried in MagicTool)
- Filtering before clustering

Looking at large numbers of genes: kinds of false conclusions and their expected frequency

Criteria for gene selection methods: sensitivity, specificity, PPV, NPV

Class comparisons (can be tried in BRB Array tools)
- ANOVA (t tests, F tests)
  - Multiple sample problems; corrections
- Permutation tests
- SAM
- Multiple comparisons

Class prediction (can be tried in BRB Array tools)
- Compound covariate predictor
- Cross validation
- Permuting p value

GO term analysis (can be tried in GenMapp II; you must register and get a password to download this free software)
- Categories of GO terms
- Relating pathways to mRNA trends by p values

Reading assignments:
Draghici chaps 6, 9

Radmacher, M, McShante, L, Simon, R (2002) A paradigm for Class Prediction Using Expression Profiles, J Computational Biol 9:505-511.

Hesterberg, T (2006) Bootstrap methods and permutation tests. Introduction to the practice of statistics, D. Moore and G. McCabe, 5th edition, WH Freeman& Co, New York, PP 1-3, 46-63, 69-70/

Dahlquist, KD, Salomonis, N, Vranizan, K Lawlor, SC, Conklin, BR. (2002) GenMapp, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31:19-20.

Doniger, SW, Salomonis, N, Dahlquist KD, Vranizan, K, Lawlor, SC, Conklin BR (2003) MAPPFinder: using Gene Ontology and GenMapp to create a global gene-expression profile from microarray data. Genome Biol 4 (1):R7.

Heyer, LJ, Kruglyak, S, Yooseph, S. (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9:1106-15.

Heyer, LH, Moskowitz, DZ, Abele JA, Kamik P, Choi D, Campbell AM, Oldham EE Akin Bk (2005) MAGIC Tool: integrated microarray data analysis. Bioinformatics 21:2114-5.

Vocabulary:

Average linkage – In order to decide whether two clusters should be linked, you want to determine the distance between those clusters. The average linkage distance is the average of all the pairwise correlations across the two clusters.

Class comparison – Using statistical methodology (i.e., hypothesis testing) to determine which genes are statistically significant (different) across the groups of interest.

Class prediction – Using a model building technique to predict to which group (class) a sample belongs.

Cross validation – A way of assessing your class prediction model. 1. Remove one sample (or 10% of the samples), and build the model on the remaining data. 2. Using the model, predict and record the group (class) membership of the removed sample. 3. Put the sample back into the full dataset, and remove a different sample. 4. Repeat steps 2 & 3 until you have removed & predicted the class membership for each sample. You should now be able to assess the model.

k - Nearest Neighbors – a way of model building (class prediction) where you find the k closest samples to the array in question (the "removed" array) and identify the most common group out of the k samples. Then you predict that the removed array is in the group which is most frequent in the k samples.

p-value – the probability of seeing data like you saw if in fact nothing interesting is going on.

Supervised – analysis using the class information (e.g., class comparison and class prediction)

Unsupervised – analysis without any class or grouping information (e.g., clustering)

Reporting prediction:

Let, for some class A,

n11 = number of class A samples predicted as A
n12 = number of class A samples predicted as non-A
n21 = number of non-A samples predicted as A
n22 = number of non-A samples predicted as non-A

Then the following parameters can characterize performance of classifiers:

Sensitivity = n11/(n11+n12)

Specificity = n22/(n21+n22)
Positive Predictive Value (PPV) = n11/(n11+n21)
Negative Predictive Value (NPV) = n22/(n12+n22)

Sensitivity is the probability for a class A sample to be correctly predicted as class A,
Specificity is the probability for a non class A sample to be correctly predicted as non-A,
PPV is the probability that a sample predicted as class A actually belongs to class A,
NPV is the probability that a sample predicted as non class A actually does not belong to class A.

Homework:

1. Begin the Clustering process. With your class data loaded into BRB Array tools, filter out 'empty' samples and flagged samples and filter out using the settings described in worksheet 2 so that you get down to between 100 and 1200 genes that are reliable and changed enough to be interesting. On the ArrayTools menu, choose clustering and select Launch cluster3.0 and Treeview. When you choose this option, you will encounter a screen that enables you to filter the samples, but don't because you already have done that. Click OK and continue with the clustering. Pick Hierarchical Clustering with the settings: centered correlation (or Spearman rank correlation) and average linkage – make sure to check that you want to cluster both genes and arrays. Later you can try k-means or other methods if you wish. (Note that clustering takes a while, so be patient.)
2. You will get a picture showing all the clustered genes with green down and red up, each column represents one of your arrays, and each row represents a gene. You can look at the top labels to find out what array each column represents. You can magnify a region of this picture using the zoom command. Choose a region to magnify by scrolling down until you see an area where your wild type and your mutants tend to have different colors. Fill in the zoom screen to have that magnified for about 15 genes where the program will also give you gene IDs. Print out and attach to this report.
3. Look up on SGD the IDs to find out if there is any common theme among the genes in the group you have selected. Comment:

4. Examine the hierarchy (looks like a branching tree) by clicking on it to select a part of the pattern. You can try this will both the array hierarchy and the gene hierarchy. See if you can click on the gene hierarchy and get the pattern you selected in part 2. Comments on ease of using a 'real' cluster versus a region you pick by eye:

5. Now, click on the ArrayTools menu and choose Class Comparison. Next choose 'between groups of arrays'. It will ask you what column in the Experimental Descriptors file to use. It will let you add a column, so make a new column in your Experimental Description File that will create groups to compare. For the 2006 data set, I would recommend three groups: snf1 and pde2 'mutant', wt stat/stat and asr1 stat/stat "wt", and wt stat/log "wt log". That's based on which ones look similar to me, but you can group it as you prefer. Run the test with the probability cutoff set at 0.005. You can view the data, which will show you genes that can discriminate between the three classes. Print out this output and attach to report; look up the functions of the genes. Comment on what kinds of genes tend to differ in expression:

6. In your Class Comparison, what is a false discovery of a gene? (explain in words, and give the results from the comparison you did in lab; look for the relevant tables near the bottom of your printout).

7. Now choose "Class prediction". PAM is the most powerful tool on this menu, but you must have 4 replicates to use it and we don't. We will choose the top option, Class Prediction. Use the groups as you defined them in part 5. Set the p value higher than the default (try p = 0.005 or 0.01 for example). Run the program and view the results. Print out and attach to this worksheet. Which classifier did the best job? Based on what criteria? When do you think you would use 1-nearest neighbor vs. 3-nearest neighbor?

8. Explain in words and numbers, using your print out for Class Comparison, the meaning of sensitivity, specificity, PPV and NPV for this set of predictions.

9. Compare and contrast unsupervised analysis (such as the clustering from steps 1-4) with supervised analysis (such as class comparison and class prediction to compare particular groups of arrays).

10. Explain why having many genes can sometimes be a problem.

11. Explain how cross validation can be used in these methods to examine the accuracy of your model. (For example, we model that stat/stat samples of wild type and *asr1* will be the same but different from wild type log/stat which will differ from the similar effects of stat/stat mutants of *pde2* and *snf1*.)

12. Load GenMappII onto your PC computer by going to this web site: http://www.genmapp.org/ where you need to register and receive a password by email. Get the program and the yeast gene list installed. Dr. Hoopes will give you a file of your data to use in this program. (Here is how I made it: the gene IDs, then a column called SystemCode with a capital D for every line, then averages of the three groups of paired arrays given in question 5, called mutant, WTlog, and WT and saved as a tab delimited text file.) Copy the file I have given you into the C drive under GenMapp and inside the Expression Datasets folder. Double click on the GenMappII symbol or name to open the program. Click on Data, then on Expression Dataset Manager. Under Expression Dataset, pick New Dataset. It will open the Expression Datasets file and you can click on the file you have put there. It will ask you to identify if any of the columns it thinks are data are actually text, usually they are not so you can skip that screen. GenMapp will undergo a process to remove doubtful genes (around 100 or so mismatched to the yeast list it has) and the rest will be added and saved into a file with the same name but with a different terminal extension (.gex), placed in the same folder. Then go to the DATA menu and choose that expression set with the gex extension.

13. Next you need to make 'color sets'. These will be ways you will color genes up or down in one of the types of arrays. Click on Color Sets and New. Then type Mutant into the Color set name box and choose the mutant average column from the gene value box. Now click NEW in the criteria builder box. Choose mutant average on the left, greater than or equal to in the middle and then add 3. The expression you are choosing will appear in the criterion box. Click on Color and pick a color for these genes (say Red if you want to be classical). Type in something like 'Up Mutant" in the Label in legend box. Then click Add. This criterion will appear as a line in the box at the bottom of the screen. Now click NEW again, choose mutant avg, less than, 0.33. Make that green and call it down mutant. Now go to the Color Sets button at the top again and choose NEW. Now you will put Wild Type in the name box, and choose WT avg. Make the same two types of criteria for that. Then go to the Expression Datasets menu at the top left and save. Now click on File and Open, pick the SC GO folder, and choose SC TCA cycle. It will open a chart showing all of the TCA cycle genes. Go on the scroll bar at the top and pick the mutant. You should see some red gene boxes and many grey ones; the red ones are induced three fold and the grey ones are not. [NOTE: If this program still has the bug (I was told this was fixed but maybe not) then you may be unable to see any color on any of the boxes even after you select the mutant color set. In that case, right click on the top gene. It will produce a box giving a weird number. Copy this number into Saccharomyces Genome Database and it will get the gene and yeast ID for you. Copy the yeast ID (YDL078C for example for the first gene under TCA cycle) from the SGD and paste it into the box where the weird number was; now click Search and then OK. Repeat for all of the gene boxes. Now go to the scroller and scroll to No Expression Data, then back to Mutant, and you should see the colors.]
    Print out the TCA cycle chart for mutant, for wild type, and for WT log. Comment on the differences you see.

14. Use the same data to look at ribosomal genes. Do they change in diauxie according to DeRisi? Comment on how general the change appears using the GenMapp method.