

Microarray Bio-informatics analysis

GIGPAD Overview

The GIGPAD platform tracks and manages microarray sample processing workflows. GIGPAD gathers sample annotations when microarray orders are placed. Case and control designations or other experimental group classifications can be captured through this order entry process. The investigator can also specify which experimental groups they would like compared through our bioinformatic pipeline. This information is passed to the Microarray LIMS system built into GIGPAD, which also captures data files through an integration with the Gene Chip Operating System (GCOS) supplied with Affymetrix Microarray instruments. Once raw data files are captured, they are sent to the Pipeline System, which is a generic framework that allows us to execute preconfigured bioinformatic analyses. Results generated through this pipeline process are then shared with the PIs/Researchers through a secure results return process.

The following section discusses the details of the bioinformatic pipeline analysis performed on Affymetrix microarray data.

The Bio-informatics Pipeline

The bioinformatic pipeline performs statistical analyses on raw expression data to identify the genes exhibiting statistically significant differences in expression the in different experimental groups. The pipeline currently supports only two class analysis. In the event that more than two classes are specified, an error message will be returned indicating that the pipeline does not support more than two class analysis.

The bioinformatic pipeline for microarray data analysis and reporting is accomplished through integration with an analysis/workflow tool called [Genepattern¹](#). Figure 1 illustrates the microarray pipeline process flow.

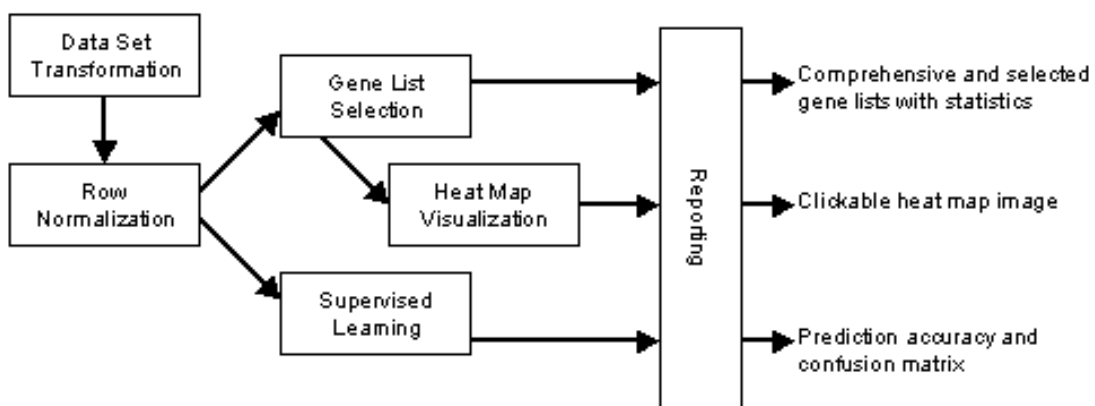


Figure 1 : Microarray Bio-informatics analysis pipeline components

This pipeline accomplishes the following three high level tasks:

1. Data Transformation
2. Data Analysis
3. Visualization and Reporting

Data Transformation

GCOS generates a table that contains the signal strength as well as detection values for each probe-set assayed for each sample. The format is known as the Affy MAS (Affymetrix Micro Array Suite) 5.0 format.

Using the sample classes specified during the order entry process, this table is transformed into two files:

1. A formatted expression data set in RES format
2. A class template in CLS format

The RES format is a tab-delimited text format that includes the signal and present/absent call for each probe across multiple chips. This format is one of two standard data set formats used by other GenePattern modules. The CLS format is also text and contains the experimental class names and defines the experimental class of each chip in the RES file.

This module was built in-house.

Data Analysis

In this phase, the data is analyzed using the following statistical algorithms:

Row Normalization

As a first step, the data is passed through an algorithm that performs row normalization. This removes differences between chips that are artifacts of the laboratory process that produced the data.

The PreprocessDataset module provided by GenePattern performs this step. Detailed documentation is available at:

<ftp://ftp.broad.mit.edu/pub/genepattern/modules/PreprocessDataset/broad.mit.edu:cancer.software.genepattern.module.analysis/00020/2/PreprocessDataset.pdf>

Gene List Selection

The normalized data set is run through an algorithm that ranks each gene according to the difference in expression level between the two experimental classes. The ranking is done using either the Signal-to-Noise Ratio or T-Test statistic.

The ComparativeMarkerSelection module²³⁴⁵⁶⁷ provided by GenePattern performs this step. Detailed documentation is available at:

<ftp://ftp.broad.mit.edu/pub/genepattern/modules/ComparativeMarkerSelection/broad.mit.edu:cancer.software.genepattern.module.analysis/00044/3/ComparativeMarkerSelection.pdf>

Supervised Learning

Supervised learning or class prediction methods represent an important paradigm in molecular classification and pattern recognition. This analysis involves selecting the features (genes) most correlated with a phenotypic distinction of interest. These features or "marker genes" are biologically interesting in themselves but they can also be used as the input of a classification algorithm that uses existing "labeled" samples to build a model to predict the labels for future

samples. Genes correlated with a binary class distinction, for example a morphological or clinical phenotype, is directly identified and selected by using a "distance" metric, for example,

$$\begin{aligned} \text{t-test statistic} &= (\mu_1 - \mu_2) \div (\sigma_1^2 + \sigma_2^2)^{1/2} \\ \text{or Signal to noise ratio} &= (\mu_1 - \mu_2) \div (\sigma_1 + \sigma_2) \end{aligned}$$

[μ and σ are the means and std. dev. per class]

The model is tested in the leave-one-out cross validation method in which one sample is held, a predictor is trained on the remaining samples. The left out sample is classified by this predictor and the process is repeated iteratively.

The model that minimizes the total error in cross-validation is chosen. This model can then be validated on an independent test set.

The WeightedVotingXValidation module⁸⁹ provided by GenePattern performs this step. Detailed documentation is available at:
ftp://ftp.broad.mit.edu/pub/genepattern/modules/WeightedVotingXValidation/broad.mit.edu:cancer_software.genepattern.module.analysis/00028/2/WeightedVotingXValidation.pdf

Note: The prediction is performed only if there are at least 5 chips in each class.

Visualization and Reporting

The visualization and reporting module converts the analyzed data into user-readable formats (.xls, .html, .png etc.)

Visualization

The top ranked 400 are extracted from the data set and represented as a heat map image based on signal values. A clickable (HTML) version of the image is also generated. Each horizontal row in the image represents the value for a single marker and the samples are listed vertically. Against each gene's heat-map image is the description of the gene that can be clicked to view the Affymetrix description of the accession.

The HeatMapImage module provided by GenePattern creates the image visualization. Detailed documentation is available at:

ftp://ftp.broad.mit.edu/pub/genepattern/modules/HeatMapImage/broad.mit.edu:cancer_software.genepattern.module.analysis/00032/5/HeatMapImage.pdf

The module generating the HTML version was built in-house.

Other Reports

Excel and HTML versions of the selected genes including the fold value, mean and standard deviation for each condition as well as the signal strengths for each sample are generated.

An excel sheet containing the above statistics for all the genes is also generated for the user.

In addition, if the prediction algorithm is run, then the accuracy of the prediction algorithm is presented to the user along with the prediction results in an excel format.

A confusion matrix is also generated that gives the accuracy of prediction based on a given confidence threshold.

A consolidated report sheet giving a description of the analysis performed and information about the different visualizations and results is also presented.

The module generating these reports was built in-house.

-
- ¹ Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, and Mesirov JP. GenePattern 2.0. *Nature Genetics* 38 no. 5 (2006): pp500-501 doi:10.1038/ng0506-500.
- ² Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. **57**(1): p. 289-300.
- ³ Golub, T., Slonim, D. et al. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. *Science* **286**, 531-537.
- ⁴ Good, P. (1994) *Permutation Tests: A Practical Guide for Testing Hypotheses*, New York: Springer-Verlag
- ⁵ Lu, J., Getz, G., Miska, E., et al. (2005) MicroRNA Expression Profiles Classify Human Cancers. *Nature* **435**, 834-838
- ⁶ Storey, J.D. and R. Tibshirani (2003) Statistical significance for genomewide studies. *PNAS*, **100**(16): p. 9440-9445.
- ⁷ Westfall, P.H. and S. S. Young (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. *Wiley Series in Probability and Statistics*. New York: Wiley.
- ⁸ Golub TR, Slonim DK, et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 531-537 (1999).
- ⁹ Slonim DK, Tamayo P, Mesirov JP, Golub TR, Lander ES. (2000) Class prediction and discovery using gene expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB) 2000*. ACM Press, New York, pp. 263–272.