



# High-Performance Computing Examples in Bioinformatics

Xing Qiu

Department of Biostatistics and Computational Biology  
University of Rochester

Nov. 4, 2013



# Outline

- 1 Welcome to the world of \*-omics data!
- 2 Advanced Techniques



# Microarrays

- The most commonly used bioinformatics data are DNA/RNA microarrays (Affymetrix GeneChip platform, Illumina Beads Array, etc) and their variants.
- Microarrays measures the expression levels (concentration of RNA/DNA in the sample) for all genes through hybridization.
- Many synthesized **probes** are attached in the array to hybridize with specific target RNA/DNA fragments.
- Expression levels detected by these probes are then processed into a  $m \times n$  dimensional matrix for further analysis. Here  $m$  is the number of probesets (can be mapped to genes,  $m \approx 20,000 \sim 50,000$ );  $n$  is the number of samples (*a.k.a.* arrays/slides,  $n \approx 10 \sim 100$ ).
- This is a typical “**large  $p$ , small  $n$** ” scenario in statistical analysis.



## Other \*-omics data (I)

- Though we focus on microarray data analysis in this talk, we would like to point out that many other \*-omics data exist.
- PCR (Polymerase chain reaction) data. Low throughput (cheaper); high sensitivity; often used *after* microarray analysis to *confirm* differentiation of specific genes.
- Protein binding microarray (*a.k.a.* biochip, proteinchip). Records protein instead of DNA/RNA expressions. Good for identifying protein-protein interactions; transcription factor protein-activation; antibody measurements.



## Other \*-omics data (II)

- RNA-seq arrays. Uses the “next-generation sequencing technology”.
- Provides data at the nucleotide sequence level.
- Can be used as a substitute of microarray expression data.
- Can also be used to detect single nucleotide variation (SNP), *de novo* reconstruction of transcriptome, etc.
- **Cons:** 5% of high abundance transcripts (“house-keeping” genes) can exhaust 75% of reads, many important genes will be below detection threshold.
- Many, many other types of data exist for specialized studies.



# Microarray Data pre-processing

- Raw scanner level data on Affymetrix platform are images with “.CEL” suffixes. R/BioConductor provides a function `ReadAffy()` to read these CEL files into R.
- The default preprocessing method provided by R/BioConductor is RMA (Robust Multichip Average) which includes
  - 1 Background correction at the image level.
  - 2 Quantile normalization, which calibrates all arrays to the same scale (all quantiles must be the same).
  - 3 Summarization. Multiple (15 – 25) probes are used to detect one target so they need to be summarized to one number.
  - 4 Variance stabilization transformation. Most common VST:  $\log_2$  transformation<sup>1</sup>.

---

<sup>1</sup>Average in the log-scale is the *geometric mean*; multiplicative noise becomes additive in the log-scale, etc.



# Hypothesis testing (I)

- Now we have the data. Which genes are “interesting”?
- Definition of Differential expressed genes.
- Due to measurement error, the phenotypic difference for every gene is non-zero.
- Statistical hypothesis testing is a way to decide whether the observed differences are **significant** or not.

	Accept $H_0$	Reject $H_0$
$H_0$	True Negative	False Positive, type I error)
$H_1$	False Negative, type II error	True Positive)



## Hypothesis testing (II)

In principle, any parametric or nonparametric statistic designed for single hypothesis testing may be useful in microarray. In practice, the following test statistics are widely used in microarray data analysis:

- Two sample student  $t$ -statistic
- Wilcoxon rank sum statistic
- Kolmogorov-Smirnov statistic
- Cramér-von Mises statistic
- $N$ -statistic





# Permutation and Bootstrap

- In reality, gene expressions are not necessarily normally distributed, so we don't know the null distribution.
- Permutation/bootstrap are two popular tools in nonparametric inference.
- Use the poker game example to illustrate the permutation idea.
- Permutation is very time consuming. Convergence is of order  $O(\sqrt{p})$ .
- Due to extremely large number of genes (exons, SNPs, etc), we the number of  $p$  has to be very, very large.



# Machine Learning Techniques

- Principal component analysis (unsupervised learning). Functional version of PCA. sPCA.
- Cluster analysis (unsupervised learning). Cluster analysis on functional space; manifold learning.
- Discriminant analysis (supervised machine learning). Penalized method (SVM, LASSO, etc) are more suitable than traditional methods such as (un-penalized) linear discriminant analysis, logistic regression, and some nonlinear discriminant analysis techniques.



# DA Analyses

- Recently, we have developed gene differential association analysis (DA analysis) to select differentially associated genes (DAGs) which change their relationship (in terms of covariance/correlation structure) with the majority of other genes [2, 1, 3].
- Combining DA and DE genes in a pathway analysis leads to appearance of new connections between nodes, which represent molecular interactions already known from the literature [1]. This experiment demonstrates that indeed genes from DE and DA lists are *interacting partners* in biological processes.



# Summary Statistic for DA Analyses

- Pearson correlation coefficient is a natural way to measure DA.
- Not normally distributed. Fisher transformation,  $w_{ik} = \frac{1}{2} \log \frac{1+r_{ik}}{1-r_{ik}}$ , was used to make it “more normal”. This is the foundation of [2].
- Not the MPT when true correlation  $\rho_{ij} \gg 0$ . In [1], we define the *covariance distance* between two genes to be  $d_{ii'}^C = \hat{\sigma}(\delta_{ii'}^C) = \hat{\sigma}(x_i^C - x_{i'}^C)$
- Covariance distance is a measure of similarity between two random variables in terms of their second order moments. This measure is more sensitive to DA when  $\rho \gg 0$ .



## From Vectors to One $p$ -value

- Multiple testing statistics for one gene.  $N$ -statistic was used to quantify the difference of the **joint-distribution** of covariance vectors.
- Permutation (with or without replacement) is used for generating the null distribution of the  $N$ -statistic, from which  $p$ -values for each gene will be computed.
- Using permutation makes this method **nonparametric**. In addition, a trimming technique is employed so our method is also **robust** to outliers.



# Implementation

- Permutation is a computationally intensive method. An efficient implementation is critical.
- Our method was implemented in `Python/C++`. Computationally intensive functions are implemented in `C++` to ensure efficiency; `Python` bindings were provided for the ease of use and flexibility.
- This program is parallelized for running on high-performance Linux clusters.
- We have discovered a two-layer hierarchical parallelization design which can save up to 60% of computation time. Basically, the coarse-grain parallelization is implemented in `MPI`; fine-grain layer is implemented in `Pthreads` to take advantage of the memory-sharing feature of modern multi-core computers.



# ODE Parameter Estimation

- The equation:

$$M'_k(t) = \beta_{k0} + \sum_{i=1}^K \beta_{ki} M_i(t), \quad k = 1, \dots, K, \quad (1)$$

where  $M_k(t)$  is the mean expression curve of the  $k$ -th module (cluster);  $\beta_{k0}$  is the intercept and coefficients  $\beta = \{\beta_{ki}\}_{k,i=1,\dots,K}$  quantify the regulation effects of other modules, including self-regulation on the rate of expression change of the  $k$ -th module. The identification of network structure is equivalent to identify the nonzero coefficients  $\mathcal{S} = \{1 \leq k, i \leq K : \beta_{ki} \neq 0\}$ .

- Though the above equation is linear, its solutions are highly nonlinear temporal functions.



## A five step pipeline

- detection of temporally differentially expressed genes,
- clustering genes into co-expressed modules
- identification of network structure (LASSO/SCAD plus two-stage method)
- parameter estimate refinement (nonlinear optimization of RSS)
- functional enrichment analysis (for biological interpretation)





# LASSO and other regularized regression

- Least-square regression: minimize  $\|Y - X\beta\|_F^2$ .
- The sparsity principle. Find a subset of  $\beta$  which minimizes the RSS. N-P hard.
- A heuristic To obtain sparse  $\beta$ , add the following penalty

$$\min \|Y - X\beta\|_F^2 + \lambda \sum_j |\beta_j|.$$

- Interpretation: replace the  $l^0$  norm by the  $l^1$  norm.






# Parameter refinement

- Use a crude LASSO regression to find the model structure (say 30% of  $\beta_{ki}$  are nonzero). Then use ODE to find an optimum system to produce solutions that minimizes true RSS.
- The above differential equation is coupled, which means change one  $\beta_{ki}$  has influence of *all* genes. Thus such algorithm is very hard to parallelize.



# Bibliography I

-  Rui Hu, Xing Qiu, and Galina Glazko, *A new gene selection procedure based on the covariance distance*, *Bioinformatics* **26** (2010), no. 3, 348.
-  Rui Hu, Xing Qiu, Galina Glazko, Lev Klebanov, and Andrei Yakovlev, *Detecting intergene correlation changes in microarray analysis: a new approach to gene selection*, *BMC Bioinformatics* **10** (2009), no. 1, 20.
-  Mark Needham, Rui Hu, Sandhya Dwarkadas, and Xing Qiu, *Hierarchical parallelization of gene differential association analysis*, *BMC Bioinformatics* **12** (2011), no. 1, 374.