Finding significant genes, enriched gene sets and meaningful signatures

Dhammika Amaratunga
Senior Director and Janssen Fellow
Nonclinical Statistics & Computing

Joint work with Javier Cabrera
Contributions by Nandini Raghavan & Volha Tryputsen

Weill Medical College of Cornell University - November 2012

Topics

1. Comparative microarray experiments
2. Finding significant genes
   - Conditional t
3. Finding enriched gene sets
   - MLP
4. Gene signatures
5. Concluding remarks

DNA microarrays

♦ DNA microarrays are widely used in genomics research to monitor the expression levels of many thousands of genes simultaneously.

♦ In a typical microarray experiment, the data is a matrix of the form:

$X = \{X_{gj} : g=1,\dots,G; j=1,\dots,N\}$

where

g indexes the genes (rows)

j indexes the samples (columns)

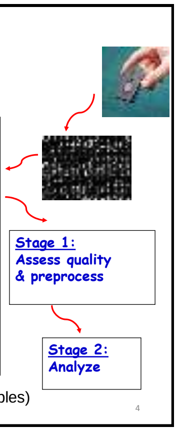
X_{gj} is a measure of gene expression for gene *g* in sample *j*

Example data set

♦ Expression measures for *G* genes in *N* samples:

	C1	C2	C3	T1	T2	T3 …
G1	83	94	82	111	130	122
G2	16	14	7	2	11	33
G3	490	879	193	604	1031	962
G4	46458	49268	74059	44849	42235	44611
G5	32	70	185	20	25	19
G6	1067	891	546	906	1038	1098
G7	118	111	95	896	536	695
G8	10	30	25	24	31	28
G9	166	132	162	27	109	213
G10	136	139	44	62	23	135
	. . . . . .	. . . . . .				

45101 rows (genes) x 12 columns (samples)



Stage 1:
Assess quality
& preprocess

Stage 2:
Analyze

Comparative microarray experiments

♦ Comparative microarray experiments are experiments in which the expression levels of *G* genes are simultaneously compared across two different groups of samples (e.g., normal vs diseased, non-cancerous vs cancerous, control vs treated).

♦ The *N* samples fall into two groups, *N*=2*k* (*k* in each of two groups).

♦ Usually *G*>>*N*. (e.g., *G*=40000 and *k*=10)

♦ It is of interest to determine which genes (and gene combinations) are differentially expressed across the two groups.

Ref: Amaratunga & Cabrera (*2004*)

A model for the data

♦ Let X_{gij} denote the preprocessed intensity measurement for gene *g* in array *i* of group *j*.

♦ Model: $X_{gij} = \mu_{gj} + \sigma_g e_{gij}$

♦ Effect of interest: $\Delta_g = \mu_{g2} - \mu_{g1}$

♦ Error model: $e_{gij} \sim F_g$ (location=0, scale=1)

♦ Can try to improve power by reducing parametrization and borrowing strength across genes:
- $F_g = F$
- gene mean-variance model: $(\mu_{g1}, \sigma_g) \sim F_{\mu,\sigma}$

Possible approaches (1)

- Parametric: Assume functional forms for F and $F_{\mu,\sigma}$ and apply either a Bayes or Empirical Bayes procedure
  → regularized test statistics:

$$\boxed{T_g = (\bar{X}_{g1} - \bar{X}_{g2})/s_g}$$ t test

$$T_g(c) = (\bar{X}_{g1} - \bar{X}_{g2})/(s_g + c)$$ SAM

or $$T_g(d) = (\bar{X}_{g1} - \bar{X}_{g2})/\sqrt{(d_g s_g^2 + d_0 s_0^2)}$$ LIMMA (+others)

Refs: Tusher, Tibshirani, and Chu (*Proc Natl Acad Sci USA*, 2001)
Smyth (*Stat Appl Genet Mol Biol.* 2004)
Wright and Simon (*Bioinformatics*, 2003)

Possible approaches (2)

- Semi-parametric: Use resampling to determine critical values ("Conditional t").

Estimate F: $\hat{F} = \{(X_{gij} - \bar{X}_{gj})/s_g\}$

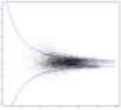
Estimate F_σ: $\hat{F}_\sigma = \{s_g\}$

Resample: $r_{ij}^* \sim \hat{F}$ and $s^* \sim \hat{F}_\sigma$ → $X_{ij}^* = s^* r_{ij}^*$

→ (t^{**}, s^{**}) → Repeat many many times

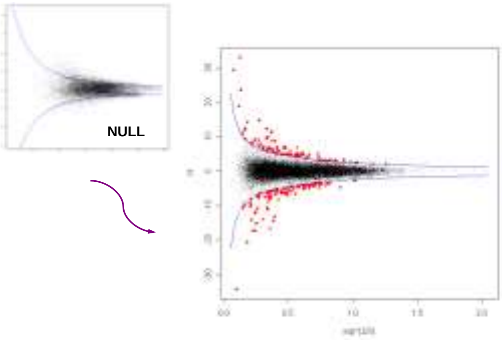
→ Form *critical envelope*, $t_\alpha(s_g)$, defined by

$$P(|T| > t_\alpha(s_g) \mid s_g \,;\, H_0) = \alpha$$



Ref: Amaratunga & Cabrera (*Statistics in Biopharmaceutical Research, 2009*)

Example



Some points worth noting

- Ct is semi-parametric: parametric in the use of a parametric model and t test statistic and nonparametric in the use of a resampling-based approach to determine critical values.
- In standard situations, Ct performs much better than the t-test and has similar performance to limma.
- In some non-standard situations, Ct performs much better than the t-test and slightly better than limma.
- Whichever method is used, a long list of *p*-values $\{p_g\}$ is generated.
- Typically no FWER-based multiplicity adjustments are applied since these studies are done considered exploratory; however, an FDR-based adjustment may be applied.

Interesting gene combinations

- In addition to genes of potential interest, gene combinations of potential interest can be identified via regression or classification techniques.
- **Regression:** minimize$_b$ Σ L(y; Σ $b_i x_i$)
  - All b's nonzero (with simple dimension reduction) or some b's nonzero (with filtering/selection)
- **Elastic net:** minimize$_b$ Σ L(y; Σ $b_i x_i$) + Σ $\lambda_1 |b_i|$ + Σ $\lambda_2 b_i^2$
  - The L1 component of the penalty (the "lasso" penalty) induces sparsity (i.e., very few b's non-zero).
  - The L2 component of the penalty (the "ridge" penalty) reduces sparsity and encourages retention of correlated genes.
  - Non-zero b's indicate genes of interest in combination.

Gene set analysis

- Now that we have identified genes (and gene combinations) of potential interest, how to interpret these findings?
- That's up to the biologists, but can we help? One way is to see whether we can identify "gene sets" that are "enriched".
- Many of the G genes (about 50%) can be categorized into "gene sets" based on their function or other characteristic.
- A gene set is said to be "enriched" if the *p*-values of the genes that comprise it tend to be smaller than a typical random gene set of the same size; enrichment could imply that the function associated with the gene set is operating differently in the two groups.

Example

- Example:
  Phagocytosis engulfment
  in KO vs WT experiment

Gene	p-value
11303	0.000651
14127	0.001703
14129	0.203787
14130	2.00E-05
14131	0.000292
16017	0.043791
17304	0.167931
19261	0.000415
56644	0.005529
70676	0.004842
380793	0.103618

This is a “significant” gene set

- **MLP statistic:**
  $\{p\} \rightarrow$ MLP = mean (-log p) = 2.34*

Assessing MLP’s significance

- Calculate the value of MLP for the gene set:
  MLP = mean(-log(p))
  - MLP should be large if the gene set is enriched.
  - n = number of genes in the gene set.
- To assess significance, draw a random sample of a size n from the set of all p-values and calculate the value of MLP for the pseudo gene set (call it MLP*); repeat many times.
- The p-value for the gene set is the proportion of times that MLP* equals or exceeds the observed value MLP:
  p=Pr[MLP*≥MLP]

Ref: Raghavan et al (*Journal of Computational Biology, 2006*)

Some points worth noting

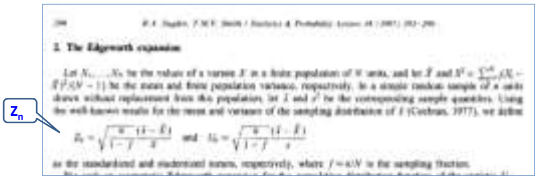
- The MLP statistic, **MLP = mean(-log(p))**, is essentially Fisher’s test statistic for pooling p-values.
- For gene set analysis, the MLP statistic generally offers higher efficiency than either the modified Kolmogorov-Smirnov statistic (which is the basis of GSEA) or Fisher’s exact test (which is basis of many software packages for gene set analysis).
- To determine significance, genes (rather than samples) are randomized, since we are interested in assessing enrichment in a specific gene set compared to a random gene set from the same system (rather than in assessing significance; i.e., the presence or absence of differential expression in that gene set).

Reducing computation time

- Attempt 1: **If p ~ unif[0,1], then MLP ~ gamma(n,1/n)**.
  However the baseline of interest for enrichment is the overall observed p-value distribution and this is generally not uniform.
- Attempt 2: **$\{p\} \rightarrow \mu=\text{mean}_X(-\log(p))$ and $\sigma^2=\text{var}_X(-\log(p))$**
  $\Rightarrow$ null distribution of MLP has mean μ and variance $f\sigma^2$ based on finite population sampling theory (f=FPC)
  $\Rightarrow$ approximately $Z_n=n^{1/2}(\text{MLP}-\mu)/(f^{1/2}\sigma)$ ~ N(0,1) based on the central limit theorem (for large n)
  However, since the distribution of {-log(p)} is very skewed and heavy tailed, convergence to normality will be slow.
- Attempt 3: Use an Edgeworth approximation to incorporate the skewness and kurtosis:
  $P(Z_n \le z_{obs}) = \Phi(z_{obs})$
  $\rightarrow P(Z_n \le z_{obs}) = \Phi(z_{obs})$ + adjustment

Ref: Amaratunga, Cabrera, De Bondt & Tryputsen (*2012*)

Edgeworth approximation



200 R.A. Hughes, T.M.F. Smith / Statistics & Probability Letters 44 (2001) 193–200

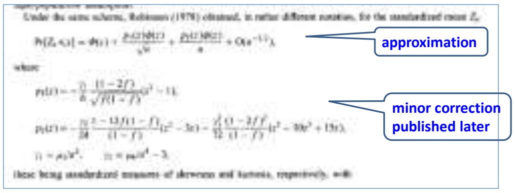
2. The Edgeworth expansion

Let $X_1,\dots,X_N$ be the values of a variate X in a finite population of N units, and let $\bar X$ and $S^2=\sum_{i=1}^{N}(X_i-\bar X)^2/(N-1)$ be the mean and finite population variance, respectively. In a simple random sample of n units drawn without replacement from this population, let $\bar x$ and s^2 be the corresponding sample quantities. Using the well-known results for the mean and variance of the sampling distribution of $\bar x$ (Cochran, 1977), we define

$$Z_n=\sqrt{\frac{n}{1-f}}\frac{(\bar x-\bar X)}{S}\quad\text{and}\quad T_n=\sqrt{\frac{n}{1-f}}\frac{(\bar x-\bar X)}{s}$$

as the standardized and studentized means, respectively, where $f=n/N$ is the sampling fraction.

Z_n



Under the same scheme, Robinson (1978) obtained, in rather different notation, for the standardized mean Z_n

$$\Pr[Z_n\le x]=\Phi(x)+\frac{p_1(x)\phi(x)}{\sqrt{n}}+\frac{p_2(x)\phi(x)}{n}+O(n^{-3/2}),$$

where

$$p_1(x)=-\frac{\gamma_1}{6}\frac{(1-2f)}{\sqrt{f(1-f)}}(x^2-1),$$

$$p_2(x)=-\frac{\gamma_2}{24}\frac{1-6f(1-f)}{(1-f)}(x^3-3x)-\frac{\gamma_1^2}{72}\frac{(1-2f)^2}{(1-f)}(x^5-10x^3+15x),$$

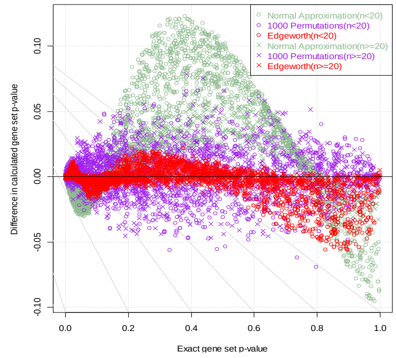
$$\gamma_1=\mu_3/S^3,\qquad \gamma_2=\mu_4/S^4-3,$$

these being standardized measures of skewness and kurtosis, respectively, with

approximation

minor correction published later

Example



Some points worth noting

- The Edgeworth approximation can be used to reduce computational time.
- The saddlepoint approximation gives better accuracy in some circumstances but has some computational and stability issues.
- It is worth considering multivariate approaches as means of finding enriched gene sets; however, the interplay between correlations, the number of genes that exhibit separation, the relative effect sizes and power is actually quite complex.

Signatures

- The term "signature" refers to a characteristic or combination of characteristics that is able to differentiate two (or more) predefined classes of samples.
- It tends to be used rather loosely and could have different meanings to different people in different contexts.
- In transcription profiling, it could mean e.g. a list of significant genes or some sort of gene combination (e.g., a classifier).
- The objective of the signature should be articulated clearly and will usually determine what type of signature is being sought.

Example

- Scientist: "We recently concluded a Phase II clinical trial involving our wonder drug TRT. There were 100 patients enrolled in the study; 50 received placebo (PBO), 50 received TRT. The response rates in the two groups were 20% and 50% respectively, a tad lower than we expected. We were able to collect gene expression data from many of these patients and would like to see whether we can leverage this to improve the response rates. We have pre-Rx gene expression data for 30 PBO patients, including 6 responders, and 40 TRT patients, including 23 responders. There is also some Rx-phase gene expression data available. We want you to take a look at the pre-Rx gene expression data for the TRT-treated patients and derive a gene expression signature for response."

Example: objectives and signatures

- What is the objective of the analysis?
  - **prediction of response to TRT using pre-Rx data**
    - ***study (R vs NR)(TRT vs PBO) at pre-Rx***
  - **prediction of response using pre-Rx data**
    - ***study (R vs NR) for all (TRT + PBO) at pre-Rx***
  - **mechanism of action of TRT (using pre-Rx and Rx data)**
    - ***study (TRT vs PBO) for (Rx vs pre-Rx)***
  - **gaining an understanding of the disease at the cellular level**
    - ***study PBO for (Rx vs pre-Rx)***
- Different signatures would be needed for different objectives.
- Whatever the objective, the identification of significant genes and gene sets is an important step in the development of a signature.

Practical considerations

- Sensitivity:
  - Evidence of differential expression (e.g., the gene or gene combinations comprising the signature should be significant).
  - Should have contextual relevance.
- Specificity:
  - Should not be picking up random variations.
  - Should be broadly specific to condition (direct / downstream)
  - Yet should be generalizable (e.g., beyond population subgroup)
- Other considerations:
  - Correlation / Redundancy / Representation of processes
  - Combination for prediction (classification)
  - Portability (for future use; lab, platform, assayability, scale-up, …)

Concluding remarks

- The development of signatures is (arguably) the single most important problem in the analysis of gene expression data.
- Identifying significant genes (and gene combinations) is important for developing signatures.
- Identifying significant gene sets is important for assessing the contextual relevance of the significant-looking genes.
- Ct and MLP work well for these purposes.
- Room for more research …

Wrap up

Contact: damaratung@yahoo.com

Web: www.amaratunga.com