High-throughput Genomic Paper Notes

Jonathan Yu

December 4, 2017

Efron and Tibhshirani, 2007

- Efron, B. and Tibshirani, R., 2007. “On testing the significance of sets of genes.” The annals of applied statistics (2007): 107-129

The authors propose two improvements to the Gene Set Enrichment Analysis (GSEA): (1) the maxmean statistic for summarizing gene-sets and (2) the restandardization for more accurate inferences. At the most basic level, the common approaches apply a two-sample t-statistic for each gene. Those genes with t-statistic above a cutoff (absolute value) are significant. GSEA suggests assessing the signficance for pre-defined sets instead of the individual genes to test whether gene-set is enriched. They would compute t-statistic for all genes in the data and then a score for each gene-set and then do permutations where you recalculate the score for each permuted dataset. A false discovery rate is computed. The maxmean proposed is comparitively higher power. The restandardization is then computed by scaling the maxmean statistic by the mean and standard deviation.

The choice of an efficient gene-set enrichment test statistic such as the maxmean is argued to give better power that is good for the location *shift* and *scale* of the z-values. The maxmean is defined as $S_{max} = max\{\bar{s}_S^{(+)}, \bar{s}_S^{(-)}\}$. This can detect unusually large z-values in both directions and robustlly does not let a few positive/negative gene scores dominate.

The restandardization can help improve inferences when score are selected randomly, theoretial null is distributed the same as that of the standardized scores and when the standardized scores are uncorrelated.

Hung et al., 2012

- Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, Charles DeLisi. “Gene set enrichment analysis: performance evaluation and usage guidelines.” Briefings in Bioinformatics, Volume 13, Issue 3, 2012, 281-291

The authors review a few methods to determine whether sets of genes are connected, i.e. those in the same gene pathway are overrepresented in differential expression, and suggest way to determine the best method when there is no comparison gold standard. The GSEA has been critically assessed by a few authors on the different variants and its choices of gene-set statistics. Compared to the others such as the median or Wilcoxon rank sum test, GSEA was proven to be least sensitive. However, in those papers, they have a gold standard, i.e. which gene sets were true positive and which gene sets were true negatives. In the case where there are no gold standard to see which method is appropriate, they introduce a concept called *mutual converage (MC)* which refelcts the extent to which the gene sets predicted by a particular method are reproduced by other methods. The results show that the GSEA most effect statstic MC while the mean and median tests had poor MC.

When performing GSEA, null hypothesis is defined by two situations: (Q1) the background distribution is obtained by shuffling genes and (Q2) the other is by shuffling the phenotypes/samples. Given uncorrelated properties for gene set-level and that a predictor has high fraction of mutually supported gene set, the MC of the predictor to be the ratio of the number of votes from the other predictors agreeing with its predictions divided by the maximum number of votes. It is the summation of the product of the prediction profiles of predictor k with the predicton profiles of the other predictors over the maximum number of votes from the predictors. This new metric has been shown to cover the location shift AND the shape changes of the obsreved differential expression distribution compared to the background distribution.

Benjamini and Hochberg, 1995

- Benjamini, Yoav, and Hochberg, Yosef. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B (Methodological), 1995, 289-300. -FDR paper

The authors present a different approach to problems of multiple significant testing by controlling the false discovery rate - expected proporotion of falsely rejected hypotheses. If you were control this, the power could be increased. As usual way to control for multiple testing problems, familywise error rate (FWER) are controleld by multiple comparison procedures (MCPs). The authors suggest a new point of view to multiplicity and thus, a posisble solution. They argue that the number of erronenous rejetions should be taken into account instead of errors that were made so by this logic, the false discovery rate (FDR) should be controlled. This will enable more power into the study at a cost of weaker control of FWER. They present some situations where FDR control is more important and than a simple Bonferroni-type FDR procedure. Through simulation, they show that the power of methods decreases when the number of hypotheses tested increases which proves the multiplicity problem and the power of FDR controlled is uniformly larger than other methods. This is increased with the number of non-null hypotheses and the number of features tested. In classical sense, control of FWER is better however, their argument is that by controlling the FDR, there is also weak control of the FWER.

storey and Tibshirani, 2003

- Storey, John D., and Tibshirani, Robert. "Statistical Significance for Genomewide Studies." Proceedings of the National Academy of Sciences of the United States of America 100, no. 16 (August 5, 2003): 9440-45. -q-value paper

Within genomics, familywise error rate and in this case, false positive discoveries, are common issues. With multiple testing of thousands of features, the goal is to identify as many significant features while controlling for false positives. They propose that the *q value*, an extension of the false discovery rate (FDR), is better than its originator. FDR is the rate of significant features that are truly null whereas the false positive rate is the proportion of the truly null features that were deemed significant. In the normal case, the simple p-value threhsold is too liberal of a false positive rate for the number of tests in 5 examples presented by the author. The usual correction is the familwise error rate which can be too conservative as well. A balance of these two would be the FDR which is the expected values of the proportion of false positives among the significant ones. This, as stated by the authors, is a good start to measure the overall accuracy of a set of significant features but it does not suffice to explain for the individual features - which brings them to the q vlaue measurement. The q-value is the expected proportion of false positives incurred when calling the specific feature significant or a more mathematical way, the minimum FDR that can be attained when calling the feature significant, $\min_{t \geq p_i} FDR$. To further expound on the appropriateness of the FDR and the q-value, two reasons were given: (1) When the null vaue is true, the FDR is equivalent to the FWER and (2) when the number of truly null features is less than the number of features, the FDR is smaller or equal to the FWER.

Smyth, 2004

- Smyth, Gordon K. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." Statistical Applications in Genetics and Molecular Biology 3 (2004)

Smyth proposed an extended hierarchical model for microarray experiments. Given the large number of gene-wise linear model fits, a moderated t, using the classical t-statistic as a basis, uses a posterior variance instead of the sample variance to detect differences in gene expressions. The innovation is that prior information is assumed for the variance among the genes and thus, the posterior can be derived to shrink the observed variances. The hypothesis would be that model coefficient that distinguishes differences between group sis zero. It was noted that when gene identifcation using either the current B-statistic model or the moderated t-statistic, they would lead to similar results albeit the t-statistic method would require less knowledge of the parameters. Simulation shows that a moderated t has a lower false discovery rate than other methods and that it can be used for single and two color microarray experiments.

Tusher, Tibshirani, and Chu, 2001

- Tusher, V. G., R. Tibshirani, and G. Chu. "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response." Proceedings of the National Academy of Sciences of the United States of America 98, no. 9 (April 24, 2001)

The method Significance Analysis of Microarrays (SAM) was developed to identify significant changes in the expression of thousand of genes during the many different biological states. A score is assigned to each gene for each change relative to the standard deviation for repeated measurements. The genes with scores above a specified threshold, permutations of the repeated measurements are used to estimate the false discovery rate (FDR).

First, the score comes from the signal-to-noise ratio where they ccounted for gene-specific fluctionations in the gene expressions d(i) $d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$ where the numerator is the difference between the average levels of gene i between two states I and U. The gene variation of the repeated measurement account is used in the denominator. Next, the differences are ranked from highest different to lowest relative differnece. Then, permutations of the repeated measurements were done in order to minimize potential confounding effects from differences between two cell lines. A statistic is calculated for each permutation which will result in a distribution of the relative differences where the mean reltaive difference can be derived. This can then compare and identify induced and repressed genes. Their method "for setting thresholds provides asymmetric cutoffs for induced and repressed genes.". The alternative method, t test, has a symmetric horizontal cutoff. SAM has neither strong or weak family wise error rate.

Bolstad et al., 2003

- Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias." Bioinformatics (Oxford, England) 19, no. 2 (January 22, 2003): 185-93. - Normalization methods description.

Three complete data methods: cyclic loes, contrasted based method, and quantile normalization are different norminalization techniques to reduce obscuring variation between oligonucleotide arrays. Current methods - scaling and non-linear method - utilized by Affymetrix do not reduce the variation as much as the other methods and performed poorly for spike-in regressions. Affymetrix methods do not deal with cases that are non-linear relationships between arrays. The cyclic loess and contrast based method are derived from the M vs A method and could be time consuming. All these methods depend on choice of baselines but the complete data methods performed matter on the matter of variance reduction and bias.

Cleveland & Devlin, 1988

- Cleveland, William S, and Susan J Devlin. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." Journal of the American Statistical Association 83, no. 403 (1988): 596-610. - Loess regression. Concept, statistics.

Locally weighted regression (loess) is introduced as a local fitting regression, similar to the least squares, that can be used for data exploration, diagnostic checking of parametric models and nonparametric regression surface. The dependent variable is smoothed as a function of the independent variable in a moving fashion. They show that this can be used to graph smooth surfaces to help choose a parametric model if needed during the exploratory phase. They then show that the adequacy of the parametric model chosen. Finally, it can be shown that the estimate can be used instead of parametric estimates. The multivariate smoother is an extension of the univariate loess smoother where they use the euclidian distance on each scaled independent variables. Loess is comprised of the weights and the neighborhood size. The M plot can help choose the fraction of points in the neighborhood. Assumption include that the loess estimate be a linear combinaton of the outcome observations, that the outcome is a normal distribution, and the loess estimate is not biased.

Altman and Martin, 2015

- Altman, N. and Krzywinski, M., 2015. Points of Significance: Association, correlation and causation. Nature Methods, 12(10)

Variability affects our internal validity (replication of experiments) and external validity (generalization of experiments to population). A well-designed experiment compromises between internal and external validity - if try to focus too much on one aspect, we lose the ability to have the other. In particular, two prinicples - precision to characterize a sample and variance from different sources together - can be combined into a nested design. Nuisance variation that occur from experimental design should be minimizaed to optimize power whereas those that occur from population should be sampled and quantified to make conclusions and determine uncertainty in the estimates of the conclusions.

Tumor Analysis Best Practices Working Group, 2004

- Tumor Analysis Best Practices Working Group. "Expression Profiling–Best Practices for Data Generation and Interpretation in Clinical Trials." Nature Reviews. Genetics 5, no. 3 (March 2004): 229-37. doi:10.1038/nrg1297.

Microarrays is a widely accepted technological way to analyze mRNA transcript levels for a genome. For this technique, there are a variety of methods methods for data generation and analysis for different experimental platforms: cDNA (sets of plasmids of specific cDNA in gridded liquid aliquots), spotted oligonucleotide (concentration of a known single-stranded sequence obtained from liquid handling on glass slides) and Affymetrix arrays (probes that are synthesized using light-activated chemistry and photolithography to find signals). The Tumor Analysis Best Practices Working Group deteremined the best practices for experimental design, probe-set analysis algorithms, signal/noise assessments and biostatistical methods. For human trials, longitudinal or corss-sectional design was determined best protocol as an experimental design for best power. On the note of technical variabliiltiy, reproducibility as well many other problems should be met with some thresholds mentioned such as 2 standard deviation from mean for scaling factor to normalize chips and percentage of present calls among samples should be within 10%. For signal/noise, they determined that each project will have its own signal/noise optimum and their own method of best analysis and compared the different algorithms (Table 1) on a number of criterions. The note that since 'feature or gene selection is vitally important when microarrays ar eused for differential diagnosis', they recommend users try different statistical methods such as standard parametric tests, non parametric methods, and global/local shrinkage methods. With aggregate gene expressions can help reduce dimension - a prevalent problem for analyzing gene related data - as well as gene selection, multiple testing and collinearity. Finally, they state back-end statistical methods such as data visualization and time-series studies can help with circumvent problems as well.

Lipshutz et al., 2005

- Liphsutz, Robert J., Stephen PA Fodor, Thomas R. Gingeras, and David J. Lockhart. "High density synthetic oligonucleotide arrays." Nature genetics 21, no. 1 (1999): 20-24

Lipshutz et al. have developped a tool to collect analyze vast amounts of genetic and cellular information simultaneously from nucleic acid strands. Using photolithography and light-activated chemistry, they are able to take advantage of the complementary properties of genes and the target design of the probes, they were able to design a DNA probe array to obtain complementary sequences and monitor a large amount of expression levels. For example, the array is coated with a chemistry protecting group to prevent DNA deposition and then a mask is placed on top to expose specified regions. A light is used to knocked off exposed protecting group and then add certain solution of nucleitide incubations that hybridize to the nucleitide at that location. This cycle is repeated until a synthesized polynucleotides of about 300,000 are created at specific locations on the 1.28 cm × 1.28 cm array. The array that can contain approximately 40,000 human genes are used for expression monitoring to understand a gene function. Fluorescence intensity image of the array show perfect match (PM) and mismatch (MM) probe pairs. The difference of PM and MM can help reduce background noise and cross-hybridization in order to detect variants in DNA sequence - to identify the difference and the position of nucleitides. For a single position on the DNA, there can a number of PM/MM probe pairs.

This technique is useful as there is a need for 'monitoring expression levels of a large number of genes repeatedly, routinely, and reproducibily' that do not need physcial intermediaries such as cDNA, PCR products or clones and the process to prepare, verify and cataloque them. Companies such as Affymetrix are developping software tools to manage, genotype, and sequence analyze these datasets.

Watson & Crick, 1953

- Watson, J. D., and F. H. Crick. "Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid." Nature 171, no. 4356 (April 25, 1953):737-38

After discovering some inconsistensies with current proposed structures of nucleic acid, Watson and Crick proposed a two helical chain structure where each are coiled around the same axis. Using the same chemical assumption of the 3', 5' linkages, the chains run in opposite directions and have bases on the insdie and the phosphates on the outside. In other words, the proposed DNA strucutre is a double-stranded helical model with two sugar-phosphate as backbones on the outside and hydrogen bonds between pairs of nitrogenous bases on the inside. The new feature includes having the two chains held by purine and pyrimidine bases - joined together in pairs by hydrogen-bond. Specifically, regarding bases, specific pairs bond together: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine). They thank Dr Jerry Donohue and Dr. M. H. F. Wilkins & Dr. R. E. Franklin for their criticisms and expirments for inspirations.