White Paper Presentation

General data analysis considerations in biomarker research

Dhammika Amaratunga

Senior Director & Janssen Fellow
Nonclinical Statistics & Computing
Janssen R&D of Johnson & Johnson

Introduction

- This is a "<u>white paper presentation</u>" regarding the derivation of a biomarker, signature or classifier when the data has been generated via a high throughput technology which is capable of measuring thousands of entities simultaneously.

- Examples of **high throughput technologies**:
  - *Gene expression microarrays*
  - *Deep sequencing*
  - *Multiplexed immunoassays*
  - *Genome-wide association*
  - *Copy number variation*
  - *Protein arrays*
  - *RNAi screens*

2

Data

- <u>Measurements</u>: *N* subjects; *G* features measured on each subject
- <u>Data</u>: Feature matrix $\{\boldsymbol{X}_{GxN}\}$ and response vector $\{\boldsymbol{y}_{1xN}\}$
- <u>Objective</u>: [combination of *G* features] ⇔ [***y***]

signal of interest

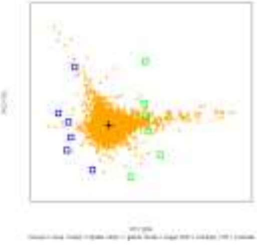
- <u>Notes</u>:
  - The <u>response</u> ***y*** could be binary, continuous, survival times, etc.
  - It is important that ***X*** (and ***y***) be <u>preprocessed</u> properly. This includes quality control, transformation, summarization, etc. It is assumed that any such preprocessing has already been done.
  - Any potential <u>secondary factors</u> (e.g., covariates, experimental artifacts) that could affect ***y*** should also be identified at this point.

Characteristics of the data

- Characteristics of the *N* subjects:
  - The number of subjects is relatively small (i.e., *N* is in the 10s or 100s).

- Characteristics of the *G* features:
  - The number of features is enormous (i.e., *G* is large, from a few thousands to the millions)
    → "megavariate" /"high-dimensional"
  - Data collection is automatic.
  - Some subsets of features are inter-correlated.
  - Since "all" features are being studied, it is likely that the features and feature combinations most relevant to ***y*** are present, perhaps in some low-dimensional subspace.
  - Many features carry no information regarding ***y***.

Initial look



- It is often useful to take an initial look at the data using simple data visualizations; they can provide information regarding …
  - data quality issues (e.g., outliers)
  - covariate effects
  - signal strength
- Biplots (or spectral maps) are particularly useful in this regard.

Questions of interest

- Which features are associated with ***y***?
  - Can be addressed by analyzing each feature individually.
    - *Methods*: Regular t-tests can be used, but modified t-tests that borrow strength across features (such as Conditional t or Limma) generally have much higher power.
- Which combinations of features can be used to predict ***y***?
  - Can be addressed using classification methodology.
    - *Methods*: Random forest, lasso, elastic net, svm.
    - <u>*Note*</u>: Methods developed for "large *N*" problems (e.g., data mining, machine learning) may not necessarily work well here. (1) Here "small *N*, large *p*". (2) Interpretability important.
- Which features are useful for predicting ***y***?
  - Can be addressed using some combination of the above.
    - *Methods*: Variable selection methodologies.

Questions of interest

♦ The objective of the signature should be articulated clearly and will usually determine what type of signature is being sought. http://www.fda.gov/downloads/Drugs/NewsEvents/UCM30073

Objective	Approach
Study MoA (Mechanism of Action) of Rx	▪ Study Rx vs PBO ▪ Adjust for Pre-Rx if available ▪ List of significant features
Develop a pharmacodynamic biomarker that can be used pre-Rx to predict response to Rx	▪ Study (R vs NR) for (Rx vs PBO) at pre-Rx ▪ Combination of a few features (classifier)
Develop a diagnostic biomarker that can be used to diagnose	Study (R vs NR) for (TRT vs PBO) at pre-Rx

General approach

♦ The objective of the analysis should be articulated clearly.

- Sometimes questions may sound the same but be actually quite different (e.g., prediction vs variable selection).
- Sometimes one study may have more than one objective.
- Common objectives are:
  - develop biomarker for diagnosis of a disease
  - develop predictor to predict response to treatment
  - identify features associated with phenotype
  - "derive signature"

♦ The approach will depend on the purpose of the exercise.

♦ A strategy as to how any findings from the analysis would be independently "confirmed" should be put in place.

Composition of signal

♦ It is likely that multiple signals will be present in the data (i.e., low dimensional projections that are substantially non-Gaussian).

♦ Some of these are the signals we are seeking …

- direct
- downstream
- interactions with secondary effects

♦ … while others are non-specific and confounding

- signal in certain features due to secondary effects
- overall signal due to secondary effects
- signals due to downstream effects

♦ In addition, the large number of features could induce spurious signals.

Enriched methods

♦ To find a meaningful signal, it helps to "enrich" the analysis by reducing the influence of genes that are less likely to be carrying the signal of interest.

♦ An analysis may be enriched in various different ways …

- Assign greater weight to more interesting genes
- Filter out less interesting genes
- Penalize less interesting genes

♦ "Unsupervised" enrichment is possible.

Example: $X_g \to VAR_g \to w_g$

♦ "Supervised" enrichment is likely to be more helpful here.

Example: $X_g \to T_g \to p_g \to q_g \to w_g$

♦ Preferred methods: Enriched random forest, enriched elastic net.

Signatures

♦ The term "signature" refers to a characteristic or combination of characteristics that is able to differentiate two (or more) predefined classes of samples.

- e.g., list of significant genes
- e.g., gene combination (such as a classifier)

♦ Considerations (beyond usual sensitivity and specificity):

- Contextual relevance
- Broad specificity to condition (direct / downstream)
- Correlation / Redundancy / Representation of processes
- Combinations (formula / up-down)
- Portability (generalizability, lab, platform, assay, scale-up)

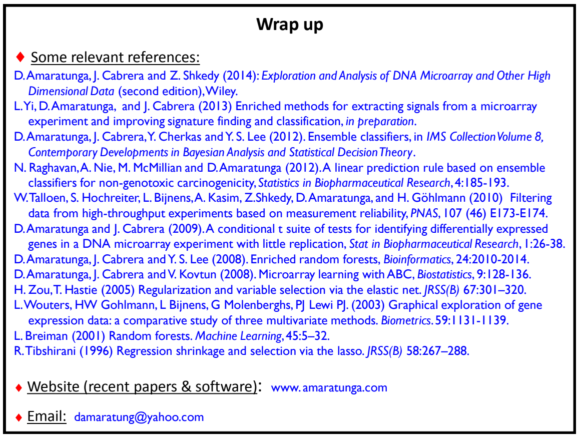
Confirmation

♦ Within study qualification:

- Performance assessment: Cross validation or bootstrap should be used to assess the performance (specificity and sensitivity) of the procedure.
- Signature: These yield multiple signatures; either some consensus combination or the top-level signature can be used as the signature.
- Problem: The "final" signature may have to be modified due to specificity or contextual considerations, possibly invalidating the initial assessment.

♦ Independent qualification is crucial:

- Confirmation: Independent or follow-up study; leave-out set.



Wrap up

♦ Some relevant references:

D. Amaratunga, J. Cabrera and Z. Shkedy (2014): *Exploration and Analysis of DNA Microarray and Other High Dimensional Data* (second edition), Wiley.

L. Yi, D. Amaratunga, and J. Cabrera (2013) Enriched methods for extracting signals from a microarray experiment and improving signature finding and classification, *in preparation*.

D. Amaratunga, J. Cabrera, Y. Cherkas and Y. S. Lee (2012). Ensemble classifiers, in *IMS Collection Volume 8, Contemporary Developments in Bayesian Analysis and Statistical Decision Theory*.

N. Raghavan, A. Nie, M. McMillian and D. Amaratunga (2012). A linear prediction rule based on ensemble classifiers for non-genotoxic carcinogenicity, *Statistics in Biopharmaceutical Research*, 4:185-193.

W. Talloen, S. Hochreiter, L. Bijnens, A. Kasim, Z. Shkedy, D. Amaratunga, and H. Göhlmann (2010) Filtering data from high-throughput experiments based on measurement reliability, *PNAS*, 107 (46) E173-E174.

D. Amaratunga and J. Cabrera (2009). A conditional t suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication, *Stat in Biopharmaceutical Research*, 1:26-38.

D. Amaratunga, J. Cabrera and Y. S. Lee (2008). Enriched random forests, *Bioinformatics*, 24:2010-2014.

D. Amaratunga, J. Cabrera and V. Kovtun (2008). Microarray learning with ABC, *Biostatistics*, 9:128-136.

H. Zou, T. Hastie (2005) Regularization and variable selection via the elastic net. *JRSS(B)* 67:301–320.

L. Wouters, HW Gohlmann, L Bijnens, G Molenberghs, PJ Lewi PJ. (2003) Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics*. 59:1131-1139.

L. Breiman (2001) Random forests. *Machine Learning*, 45:5–32.

R. Tibshirani (1996) Regression shrinkage and selection via the lasso. *JRSS(B)* 58:267–288.

♦ Website (recent papers & software): www.amaratunga.com

♦ Email: damaratung@yahoo.com