General data analysis considerations in biomarker research

Dhammika Amaratunga
Senior Director & Janssen Fellow - Nonclinical Statistics & Computing

Version date: December 21 2013

♦ Preface

This is a “white paper” regarding the derivation of a biomarker, signature or classifier when the data has been generated via a high-throughput technology capable of measuring thousands of features simultaneously.

♦ Introduction

Biomarkers are measurable indicators of a biological condition. They can be exploited in pharmaceutical R&D in multiple ways, including diagnosis and prognosis (e.g., assessing likelihood of response to treatment).

Biomarkers can be derived in many ways. More and more, modern high-throughput biological technologies, such as microarrays, deep sequencing systems and multiplexed immunoassays, are playing a pivotal role in biomarker derivation. These technologies allow researchers to monitor the activities of potential markers (e.g., genes, proteins) at the rate of thousands of markers at a time. Developing a biomarker out of the data generated by these experiments can be quite challenging; overfitting is a concern; overall, many issues need to be considered and specialized analysis techniques may be needed.

♦ Data

The data in a typical biomarker study is of the following form:

- *G* features have been measured on each of *N* subjects.
- A “feature” refers to a gene, a protein, etc.
- The feature measurements could be continuous, count, etc.
- The data for the features have been arranged into a *G*x*N* feature matrix *X*.
- The responses have been arranged into a 1x*N* response vector *y*.
- The response *y* could indicate groups or be binary, continuous, onset times, etc.

♦ Objectives

Roughly, the objective of the study is to determine “signatures”, features or combinations of features associated with the response. These features can be used to elucidate which features have been affected or to develop a prediction rule for response.

The exact objectives of the study should be defined right from the start. Sometimes questions may sound the same but be actually different (e.g., prediction vs feature selection). Therefore the questions must be articulated carefully. Another point to keep in mind is that sometimes one study might have multiple objectives and each objective may need to be addressed differently.

Common objectives are:

- develop a biomarker for early diagnosis of a disease
- develop a predictor to predict response to treatment
- identify molecular features associated with a phenotype
- derive a “signature” for a phenotype (a term that is used generically)

How the data are analyzed will depend on the purpose of the exercise. For example, in a clinical trial, if the objective is to derive a pharmacodynamic biomarker that could be used pre-treatment to predict response to treatment, the analysis should compare responders versus non-responders for the treated patients versus the placebo patients using pre-treatment data. For practical purposes, a classifier that involves only a few features is likely to be sought. On the other hand, if the objective is to study the mechanism of action of a treatment, the patients in the treatment group could be compared to the patients in the placebo group post-treatment (with perhaps an adjustment for baseline) and the result would be a list (maybe a long list) of significant features, which can be assessed for biological relevance (e.g., via pathway analysis).

♦ Preprocessing

It is important that X (and y) be preprocessed properly. This includes quality control, normalization, transformation, summarization, etc. The particular type of preprocessing will depend on the technology. It is assumed that any such preprocessing has already been done.

Methodology: For microarray data, quantile normalization and log transformation are standard.

♦ Design considerations

Any potential secondary factors (e.g., covariates, experimental artifacts) that could affect the response should be identified.

It is best if the study is designed on the basis of standard Design of Experiments principles (such as randomization, replication, balancing of covariates across treatment groups) as much as possible. However, sometimes this is not possible, such as if the data is opportunistic (e.g., observational or a study of treatment responders). In this case, the analysis and the interpretation of the analysis findings should take into account any covariate imbalances. In general, these types of studies cannot be regarded as standard confirmatory studies.

♦ Characteristics of the data

Usually the number of subjects is relatively small (i.e., N is in the 10s or 100s) compared to the number of features (i.e., G is large, from a few thousands to the millions). Such data is called “megavariate” or “high-dimensional”.

Some characteristics of the feature data:

- Data collection is often automatic using a standard platform (such as an Affymetrix chip for gene expression, an Illumina system for deep sequencing, a RBM panel for proteins).

- ➢ Some subsets of features are inter-correlated, since molecules involved in biological processes function by interacting with each other as well with external stimuli.
- ➢ Many features carry no information regarding *y*.
- ➢ Since a very large number of features are being studied, it is likely that features and feature combinations most relevant to *y* are present in some low-dimensional subspace. This is the fundamental premise underlying this type of study.

♦ Initial look

It is often useful to take an initial look at the data using simple data visualizations as they can provide information regarding …

- ➢ data quality issues (e.g., outliers)
- ➢ covariate effects (e.g., gender, investigator site, array/plate)
- ➢ signal strength (see below)

Methodology: Biplots (or spectral maps or principal components analysis plots or factor analysis plots) are particularly useful. Other EDA (Exploratory Data Analysis) techniques such as boxplots or correlation plots are also often useful.

♦ General questions of interest

Broadly speaking, biomarker derivation studies have two types of objectives.

(1) *Comprehension*: Here the objective is to identify features associated with *y* in order to study, at a cellular level, the effect of a disease or a treatment. For example, this is useful if the purpose of the experiment is to understand which biological processes are affected by treatment.

Methodology: This type of question can be addressed by analyzing each feature individually, i.e., via *uni-feature analysis*. 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. More complex models may be needed depending on the situation. FDR assessments are useful to control the proportion of false positives. In addition, a follow-up gene set analysis or pathway analysis may be performed to interpret the findings.

(2) *Prediction*: Here the objective is to predict *y* for a “new” subject.

Methodology: This type of question can be addressed by identifying a combination of features that can predict *y*, i.e., via *multi-feature analysis*. Regression and classification methodology can be used. Penalized regression methods (such as lasso and elastic net) are useful to force parsimony. Classification trees are useful as they generate easily understandable classifiers. Other possible methods include partial least squares, random forest and support vector machines.

♦ A note regarding “Big Data” techniques

Methods developed for “large *N*” problems (e.g., data mining, machine learning) may not necessarily work well in this situation, since here:

- ➢ *N* is small and *G* is large (rather than the other way around).
- ➢ May not generalize well (i.e., may overfit).
- ➢ Interpretability of findings is important.

♦ A note regarding multiplicity

Some degree of overfitting is inevitable in this kind of situation. If done carefully, a classification rule can generally be developed with a reasonable amount of protection against gross overfitting. Selecting the most useful features, however, is more prone to overfitting. There will almost certainly be a fair number of false positives. FWER adjustments are not possible as they decrease power too much. On the other hand, FDR assessments are useful and provide a certain amount of control over overfitting.

Due to the high likelihood of overfitting, independent confirmation of results is crucial.

♦ Signal

The premise underlying these studies is that the primary signal (low-dimensional non-Gaussian projection) of interest lies in a low-dimensional subspace; i.e., that there is a low-dimensional non-Gaussian projection of *X* that is associated with *y*.

There are different types of signals:

- ➢ Strong signal carried by many features
- ➢ Strong signal but very few features carry signal
- ➢ Weak signal

It is useful to categorize the type and strength of signal early on in the analysis. In some cases, there may not even be a signal or it might be too weak to detect.

A common difficulty is that often there are multiple signals in the data.
Some of these are the signals we are seeking:

- ➢ Direct effects
- ➢ Downstream effects
- ➢ Interactions with secondary effects

Others are non-specific and confounding:

- ➢ Signal in certain features due to secondary effects
- ➢ Overall signal due to secondary effects
- ➢ Non-specific signals due to downstream effects

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

♦ Enriched analysis

To find a meaningful signal, it helps to “enrich” the analysis by reducing the influence of features that are less likely to be carrying the signal of interest. An analysis could be enriched in various different ways …

- ➢ Assign greater weight to more interesting features

- Filter out less interesting features
- Penalize less interesting features

Enrichment may be supervised (i.e., take into account the response information to do the enrichment) or unsupervised.

Methodology:

- For unsupervised enrichment, calculate the variance V_g for each feature and assign higher weights to features with higher variance; e.g., $W_g = \log(V_g)$ or $W_g = V_g$.
- For supervised enrichment, calculate the t statistic T_g for each feature and assign higher weights to features with greater separation; e.g., $W_g = -\log(q_g)$ or $W_g = 1/q_g$, where q_g is the FDR-adjusted p-value or q-value associated with $T_{g.}$
- Supervised enrichment is likely to be more helpful here but could lead to overfitting.

♦ General approach

For classification or regression, methods such as cart or lasso or elastic net may be run on enriched data. These could be run in a loop, using a resampling technique such as cross validation or bootstrapping, with the model fitting done on the “in-bag” samples and performance assessments done on the “out-of-bag” samples. This provides some protection against overfitting, provided that as much of the procedure as possible, including enrichment, is done within the resampling loop. This is the principal behind methods such as enriched random forest and enriched ensemble lasso.

These methods provide a series of models. Either the topmost model or a consensus or average model can be taken as the final model.

♦ Characteristics of signatures

The term “signature” refers to a detected signal (here, a set or combination of features) that is able to differentiate two (or more) predefined classes of samples.

- e.g., a list of significant features
- e.g., a combination of a subset of features (such as a classifier)

Signatures can be derived using the methods listed above and can be used to develop biomarkers.

Considerations (beyond usual sensitivity and specificity):

- Whether the components of the signal are contextually relevant.
- Whether the signal is broadly specific to the response, either directly or downstream.
- Whether there are inter-correlated (perhaps redundant) features in the signal (maybe ok).
- Whether multiple processes are represented in the signal.
- Whether the combination is of a form (formula / up-down) that is usable.
- Whether the signal is portable (generalizable, lab, platform, assay, scale-up)

Based on these considerations, the detected signal may have to be modified to produce a “final” signature. Note, however, that this invalidates prior performance assessments that might have been done via cross-validation or bootstrapping.

♦ Confirmation

Independent qualification, using an independent or follow-up study or a leave-out set, is crucial.

♦ Some relevant (mostly in-house) 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, *Statistics 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 and T. Hastie (2005) Regularization and variable selection via the elastic net. *JRSS(B)* 67:301–320.

L. Wouters, HW Gohlmann, L Bijnens, G Molenberghs, and 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.

♦ Contact information

Websites:

- teamsna.jnj.com/prd/panonclinicalstatistics
- www.amaratunga.com

Email:

- damaratu@its.jnj.com
- damaratung@yahoo.com