

Emerging Science

abstr 0828

Statistical Aspects of Gene Signatures and Molecular Targets

Mithat Gönen, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

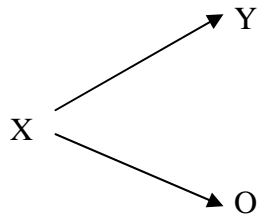
Evolution of high-throughput technologies has enabled the quantification of several thousands of gene expressions simultaneously. Many oncologic applications have emerged, two of which will be discussed: developing gene signatures and finding molecular targets.

A gene signature is a rule to predict patient outcome, usually survival or progression, from the expression of a relatively small number of genes. The scientific community has been frenzied by the proliferation of methods to develop gene signatures. These methods are usually derivatives of statistical regression or machine learning techniques. Despite the fanfare, there is little evidence that increased methodologic sophistication has resulted in substantial improvements in predictive accuracy. This can be explained by the possible abundance of “low-hanging fruit”: There are some, perhaps many, genes that are reasonably good predictors of outcome, and most sensible methods, including simpler ones, will capture some of these genes. The additional genes included in a rule will vary according to the methods used, but will make only small improvements in the overall predictive accuracy. This suggests that the return on investment on sophisticated methodology for making predictions is relatively low.

Application of findings that claim good predictive accuracy requires a good deal of care about validation. In any field, initial findings usually get the lion’s share of excitement and credit. Despite widespread agreement that independent validation of these findings is scientifically essential, credible validation can be overlooked. This is especially important in gene signatures: the number of genes is an order of

magnitude greater than the number of samples and it is easy to overfit. To ensure that the incremental gains are not sophistry, we should demand careful validation of gene signatures before they are adopted for routine use.

While genes signatures are specifically developed for predicting outcome, resourceful scientists have used them for identifying molecular targets. This looks like a free lunch at first: if overexpression of a gene increases the likelihood of progression or death, a molecular intervention suppressing the expression might be a reasonable treatment strategy. The fallacy in this thinking centers around statistical concepts of *correlation* and *causation*. Suppose in a simple world there are two genes, X and Y, and the expression of X is the only determinant of outcome O. It also happens that expression of X also derives the expression of Y but the expression of Y has no mechanistic connection on O. This can be depicted with the following diagram:



Note the absence of a direct link from Y to O. In this setting one can easily find Y to be a predictor of O because both of them follow from X. As long as the Y is an accurate predictor of O as established on a validation sample there is nothing wrong in using it in practice, despite the fact that mechanistically, Y and O are linked only through X. Nevertheless, it would be incorrect to decide that Y is an appropriate target because modifying Y will have no bearing on O.

Now consider a pathway with several genes in a map such as above, and it becomes obvious that there are several possible mechanistic configurations. A gene signature has no reason to uncover these links, but without some understanding of the links it is impossible to identify targets.

This reasoning explains another phenomenon that has baffled some oncologists. It is entirely possible to have two (in fact several) gene signatures that have no common genes. Either due to slight variations in methodology or sampling error, different genes may be used to represent the information contained in one set of genes. In other words, gene signatures are hardly unique.

This presentation will emphasize the differences in statistical methods used and required for developing gene signatures and identifying molecular targets. As for the former, validation methods should be considered routinely and carefully. Choice of methodology seems to be a less critical factor. Identifying molecular targets will require more careful and focused experimentation than large-sample microarray studies.