Poster D-11

From systems to symptoms: A combined genomics approach to expression-based breast cancer outcome prediction

Authors:
Victor J. Weigman (*Department of Biology, Program in Bioinformatics and Computational Biology, UNC-CH*)
Melissa Troester (*Department of Pathology and Laboratory Medicine, UNC-CH*)
Xiaping He (*Lineberger Comprehensive Cancer Center, UNC-CH*)
Chien-Hui Huang (*Lineberger Comprehensive Cancer Center,UNC-CH*)
Charles M. Perou (*Department of Pathology and Laboratory Medicine, Department of Genetics, Lineberger Comprehensive Ca*)

Short Abstract: Headway has been made in the microarray realm to further stratify breast cancer patients into affect groups, which have significantly improved outcome prediction and response to therapy from traditional pathology. Incorporation of genome variant information would greatly assist this approach by providing biological evidence concordant with transcriptome analysis.

Long Abstract:

It has been previously shown that microarrays can be used to gather information about cancer genes and rank their importance in patient outcomes. The predictive power of patient-subtype classification has had significant impact on outcome prediction. Given the strong biology of the members within each subtype (or differing cancer etiology), we can generate genes with strong prognostic value given fairly straightforward statistical metrics. It is known that these prognostic genes have explicit functional relevance, however, the exact manner in which they relate is can be speculative.

Currently, these gene sets undergo a various rounds of testing in order to craft a biological story around these areas. Most commonly, functional annotation is applied in order to utilize Gene Ontology memberships to characterize biological information. Hypogeometric mean and Z score metrics are typically used to quantify if the number of genes in a particular GO category is greater than expected. Typically, significant results from these analyses yield broad biological processes such as “DNA replication or Metabolism”, which give little more insight than observation of the gene names. Given the nature of microarrays to only represent the transcriptome of an organism, extrapolation of functional annotation based on microarray data alone may lead to spurious assumptions.

The need to expound upon the data from microarray studies forces us to utilize different information into the overall patient analysis. Incorporating multiple data types along with microarray data is not novel. Marcotte has stated earliest on the value of adding biological evidence to high throughput methods such as expression microarrays. Some laboratories have gone further to explain multidimensional frameworks in which this can be accomplished but with little explicit results. This leaves one to build on theoretical frameworks with more comprehensive data accompanying each patient.

Identification of critical genes involved with breast cancer and the effect mutations have on these genes have been around for a while. Generally, that has been due to meticulous work

done by laboratories who specialize in the quantification of small sets of genes. Twenty percent of breast cancer cases have been attributed to mutations in such high profile genes as BRCA1 and 2 and single percentages attributed to genes such as p53, RB, EGFR, KRAS, PTEN. High-throughput sequencing has allowed the resequencing process to be more efficient and more information regarding specific mutable regions within the chromosome. Arising out of this growing resource of sequence information were Single Nucleotide Polymorphisms (SNPs), insertions, deletions, or other nonsynonymous changes in codon structure. Some of these differences could be associated with belonging to a certain race/family, while others were shown to be predictive of cancer predisposition. This resource assists in the detection of novel genomic variants elements as well as further strengthens previous mutations and SNPs.

This study employs roughly 200 patients from which microarray data has already been prepared, analyzed, and resulting subtype predictions noted. We re-examine the sequence of 22 genes which are believed to participate in crucial breast cancer pathways by resequencing individual exons or other characterized regions. Sample DNA is derived from the same tumor tissue of which the expression values are drawn for consistency. Mutation statuses of these genes and particular loci will be documented to accompanying databases, from which the microarray data is stored. We hope to uncover new associations to gene modifications that can be attributed to certain breast cancer subtype development. These associations will accompany the prognostic genes and provide information on possible perturbations into crucial breast cancer pathways and provide biological insight that is lost from looking into expression data alone.

The affect adding these mutation data have to our predictive genes will be examined by looking into partitioning methods such as Random Forests and Bayesian Trees which have held success into methods that effectively stratify classification groups containing heterogeneous data. This will allow us to see if improvement can be made to existing outcome groups or if we can generate new, heterogeneous, prognosticators to generate new subtypes with improved outcome and response to therapy.

This work takes a preliminary look into what can be gained from a comprehensive systems study in which a patient’s transcriptome and sequence composition can be combined with clinical parameters to better quantify the mechanisms controlling their disease. Thus improving current outcome predictions as well as reclassify tentative patients or maybe provide basis for new patient subtypes.