Cross-study validation and combined analysis of Molecular Classification Data

Giovanni Parmigiani, Elizabeth S. Garrett, Anbazhagan Ramaswamy, Edward Gabrielson
Sidney Kimmel Cancer Center at Johns Hopkins University

Several recent studies have sought to refine the classification of cancer through gene expression profiling, using various gene microarray platforms. To initiate the process of cross-validating and integrating the results of these types of studies, we developed statistical approaches that allow overall assessments of profile similarities, comparisons of individual genes for association with outcomes, and combined supervised and unsupervised analyses. Our analysis focused on three projects that analyzed gene expression in a wide range of lung cancer histologic types, using cDNA and oligonucleotide array platforms. We first compared the data from these studies for consistency of co-expression relationships among pairs of genes. We then compared studies for associations of specific genes to outcomes, focusing on gene expression patterns associated with the differentiation of squamous cell carcinoma from adenocarcinoma. Two of the projects reported data on these two histologic classes of lung cancer. By using only the consistent genes identified by the analysis of coordinate gene expression, we reduced the overall variance between the studies by approximately 50%. Finally we developed an approach to the combined analysis based on latent categories signifying under-, over-, and baseline-expression. Following this approach we can generate results that are more easily interpretable, more easily translated into clinical tools, more robust to noise, and less platform-dependent.

Related materials: Software and preprints related to this presentation can be found at http://astor.som.jhu.edu/poe. Giovanni Parmigiani can be contacted at `gp@jhu.edu`. Related references include:

1. Parmigiani G, Garrett ES, Anbazhagan R, and Gabrielson E. A statistical framework for expression–based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B*, with discussion, 64:717–736, 2002. [Abstract and Full text]

2. Parmigiani G, Garrett ES, Irizarry R, and Zeger SL. (eds) *The analysis of gene expression data: methods and software*, New York: Springer, 2003. Chapter 16. [Website]

3. Parmigiani G, Garrett ES, Anbazhagan R, and Gabrielson E. *The Molecular Classification of Lung Cancer: A Cross-Study Comparison of Gene Expression Data Sets*, under review.

4. Scharpf R, Garrett ES, Hu J, Parmigiani G. Statistical Modeling and Visualization of Molecular Profiles in Cancer. *Biotechniques*, 34, S22–S29. [Full TExt, pdf]