The explosion in DNA microarray technology in the last decade has given rise to extensive biological data in the form of expression profiles of tens of thousands of genes and proteins, often from only a handful of tissue samples. The principal objective of a high-throughput experiment can be generally characterized as one of class comparison, class prediction or molecular pattern discovery. Class comparison studies are designed to identify differentially expressed genes between different classes such as tissue types, patients or experimental conditions. In class prediction, the emphasis is on building a predictive gene set based on the class labels and expression profiles of known samples, and applying it to a new sample to predict its class. In molecular pattern discovery, however, the classes are not defined independently of the gene expression profiles and are unknown a priori.

The focus of my research is the development of novel statistical methodology that will enable analysis of large-scale biological data stemming from high-throughput experiments such as microarrays, comparative genomics hybridization, and proteomics. This includes methods for relating outcome variables (qualitative or quantitative) with large numbers of covariates, and molecular pattern discovery, based on supervised and unsupervised learning methods. Two methods we are currently investigating are non-negative matrix factorization and support vector machines. Class comparisons for identification of differential expression and class prediction fall within this framework. Our focus is on combining an information-theoretic approach with learning-theoretic methods for the discrimination of competing models and elucidation of clusters and hidden variables within such large-scale data.

We are exploring various matrix factorization methods in order to gain an understanding of their relative strengths and weaknesses, applicability and relationship to each other in this setting. One specific problem of interest is an attempt to associate large-scale molecular data and clinical data with survival time in the presence of censoring. This is an important issue in translational medicine; however, very little research has been done in this area.
unsupervised method for molecular pattern discovery using NMF, based on Renyi's measure of distance between two non-negative matrices A and B denoted by D (A||B), related to the Poisson likelihood. This measure is indexed by a parameter, \( \alpha \), and is given by

\[
R_\alpha(A, B) = \frac{1}{\alpha - 1} \sum_{i,j} [A^\alpha B^{1-\alpha} - \alpha A - (1-\alpha)B], \alpha \neq 1
\]

Renyi's measure includes various well-known distance measures as special cases. In the limiting case \( \alpha \rightarrow 1 \), D becomes the Kullback-Leibler divergence given by

\[
KL (A, B) = \sum_{i,j} (A \log(A/B) - A + B)
\]

In the special case that \( \alpha = 0.5 \), we have a symmetric distance measure (better known as the Bhattacharya distance). The case \( \alpha = 2 \) gives the Pearson chi-squared statistic. A similar measure is obtained for \( \alpha = -1 \). Renyi's measure is also invariant to any non-singular transformation of the original data. These powerful properties make our method widely applicable to a variety of data structures and distributions.

Our approach provides a unique and generalized framework for non-negative matrix factorization. We demonstrated the applicability of this method using cancer microarray data for elucidating cancer sub-types and identifying genetic networks.

**Gene expression analysis of time-course microarray data.** Devarajan, in collaboration with Wiest,§ Rhodes,§ Slifker§

Immature thymocytes were induced to differentiate in a synchronized wave and time points were taken (3, 6, 12, and 24 hours) that represent incremental advances from the undifferentiated (0 hours) to the fully differentiated (24 hours) state. Genes whose expression was modulated during differentiation were identified by performing expression profiling across the four time points listed. The list of differentially expressed genes was then interrogated using NMF and relevant biological correlations were extracted. The stochastic nature of the NMF algorithm provided an approach for evaluating the stability of the clustering and a quantitative evaluation of the performance of the method. A generalized approach to molecular pattern discovery using NMF, as described in Devarajan & Ebrahimi (submitted for publication, 2005), was adopted. NMF analysis identified six distinct clusters (1–6) of genes with high confidence. Cluster 6 was further sub-divided into two clusters (6a & 6b). Among the genes identified are many known to be functionally important in regulating thymocyte development. All of these genes exhibited significant changes in expression by the 6-hour time point and were enriched in clusters 3 and 6b. Based on this, we identified 30 candidate genes whose importance in thymocyte development and transformation will be assessed through functional screens. In addition, based on this analysis we selected the 6-hour time point, at which most genes known to regulate early thymocyte development are differentially expressed, as a source of mRNA from which to produce a cDNA library that will be employed in our forward genetic screen.

**Evaluation of in vivo and in vitro pharmacology and toxicology of preventive agents using human mutant cells from dominantly heritable cancers.** Devarajan, in collaboration with Knudson,§ Bellacosa,§ Clapper,§ Ross,§ Caretti,§ Godwin,§ Yeung§

The purpose of this study is to identify potential molecular targets of cancer chemopreventive agents. The greatest opportunity to identify such early biomarkers is provided by dominantly inherited cancer syndromes whose responsible germinally mutant genes have been characterized. We examined the mRNA expression profiles of primary cultures from selected tissues (breast, ovary and kidney epithelial cells, and colon fibroblasts) obtained from individuals with six representative heritable cancer syndromes using Affymetrix technology. This was done in the presence and absence of a panel of putative chemopreventive agents. We preprocessed gene expression data using the Robust Multi-chip Average method (Irizarry et al., *Biostatistics* 4:249, 2003). We compared expression profiles between genotypes and treatments using the local pooled error method, analysis of variance and the Wilcoxon test. Statistical significance was measured in terms of \( q \)-values adjusted to control the false discovery rate (FDR), while biological significance was measured in terms of mean fold-changes between the classes being compared. The \( q \)-value for a given gene is the minimum FDR incurred when calling the gene significant. A volcano plot of \( q \)-values versus fold-change on the logarithmic scale enabled us to visualize the relationship
between statistical and biological significance. We identified differentially expressed genes between treatments within genotypes, genotypes within treatments, and identified genes for which the effect of treatment was different between genotypes. We applied unsupervised clustering methods such as NMF, principal component analysis and hierarchical clustering to identify potential sub-groups of interest within the genes and samples.

A mouse model recapitulating molecular features of human mesothelioma. Devarajan, in collaboration with Altomare,§ Vaslet, b Skele,§ De Rienzo,§ McClatchey, d Kane, b Jhanwar, d Testa§

Malignant mesothelioma (MM) has been linked to asbestos exposure and generally has a poor prognosis because it is often diagnosed in advanced stages and is refractory to conventional therapy. We reported a mouse model of asbestos-induced carcinogenesis that develops MM. These contained molecular alterations considered to be hallmarks of human mesothelial cell tumorigenesis, including inactivation of key tumor suppressor genes and activation of Akt. Overall, asbestos-treated Nf2 (+/−) mice exhibited decreased MM tumor latency compared to asbestos-treated wild-type littermates. Moreover, Nf2 (+/−) mice were found to have significantly shorter survival times than wild-type mice (using the log-rank test), and a statistically significant difference was found between the two groups (p-value = 1.54 × 10−3). Using Fisher’s exact test, the association between genotype, Nf2 (+/−) or Nf2 (+/+) and incidence of MM was tested. At the 5% significance level, there is a statistically significant association between genotype and incidence of MM (two-sided p-value = 0.039). The odds of MM incidence are 4 times as high in the Nf2 (+/−) mice as in the wild-type mice.

Age-dependent morphological alterations of human ovaries from populations with and without BRCA mutations. Devarajan, in collaboration with Xu,§ Godwin,§ Ozols,§ Daly,§ Klein-Szanto,§ Lynch, e Edelson, § Cai §

Ovarian cancer often develops in women of perimenopausal age, when ovulation ceases but gonadotropin levels are increased. We investigated a recent collection of 52 ovaries from prophylactic oophorectomies of a high-risk (HR) population and 66 control ovaries from non-neoplastic diseases referred to as normal-risk (NR) group, to determine if ovarian morphological changes relate to BRCA1/2 genotypes or to reproductive history. All statistical tests for association were carried out using the two-sided Fisher’s exact test at the 5% significance level. We found no statistically significant difference in frequency of these histopathologic features between HR and NR groups. However, inclusion cysts and deep invaginations were found much more commonly in women age 45–54 of either HR or NR groups. When age was categorized into two groups based on peri-menopausal status (45–55 years and other), a statistically significant difference was found between age group and frequency of occurrence of morphological features. The odds of occurrence of inclusion cysts were 5.43 times as high in women aged 45–54 relative to women in other age groups (p-value = 0.009). Likewise, the odds of occurrence of deep invagination were 6.42 times as high in women aged 45–54 relative to other women (p-value = 0.008), and the odds of occurrence of pseudo-stratification were 3.77 times as high in this group of women as in other women (p-value = 0.039). Thus, ovarian morphological features appear to be related to perimenopausal status. The study suggested, however, that no significant increase in the presence of non-neoplastic ovarian morphological changes is associated with BRCA1/BRCA2 mutations.

Publications


§ Fox Chase researcher

a N. Ebrahim: Division of Statistics, Northern Illinois University, DeKalb, IL
b C. Vaslet, A. Kane: Department of Pathology and Laboratory Medicine, Brown University, Providence, RI
c A. McClatchey: Department of Pathology, MGH Cancer Center and Harvard Medical School, Charlestown, MA
d S. Jhanwar: Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY
e H. Lynch: Department of Preventive Medicine and Public Health, Creighton University, Omaha, NE

Fox Chase Cancer Center 2005 Scientific Report