Wolfl, Stefan [63]

Screening of lung cancer samples using gene expression profiling

Stefan Wolfl[1], Larissa Odyvanova[1], Torsten Kroll[1], Jorg Sanger[2] & Joachim Clement[1]

[1]*Klinik für Innere Medizin, Universität Jena, Jena, Germany*
[2]*Institut für Pathologie, Bad Berka, Germany*

Tumor initiation and progression is a complex process that differs not only between different tumors but also between distinct areas of a single tumor. Therefore it is important to analyze key players in cellular metabolism in tumors simultaneously. Lung cancers are one of the dominant causes of tumor-related death. The prospect for understanding the genetic background of these tumors and their complex pathological picture makes lung cancer a particularly interesting target for functional genomic analysis. We compared tissue samples from lung adenocarcinomas and squamous cell carcinomas with samples from adjacent tumor-free resection margins. Tissue samples were collected at surgery and immediately frozen in liquid nitrogen, and all sample material was histologically classified. For expression array analysis we purified total RNA using an optimized CsCl-cushion centrifugation protocol. We hybridized complementary DNA array membranes (Clontech Atlas Arrays and UniGene-based membranes from RZPD, Berlin) with ^{33}P cDNA and collected data by phosphoimaging. Comparison of the expression patterns led us to the following observations: (1) Tumor and tumor-free resection margins show a high degree of correlation in their expression patterns. (2) Comparisons between tumors or between tumor-free tissue samples show no obvious correlation. (3) The similarity between tumors and normal tissue from one patient is more pronounced in adenocarcinoma than in squamous cell carcinoma.(4) Filtering genes that represent plate epithelial carcinoma versus adenocarcinoma led to two potential discriminating candidate genes, those coding for interleukin-1α and granzyme A.

Xiong, Momiao [64]

Structural equation models for pathway identification

Momiao Xiong

University of Texas, Houston, Texas, USA

Genome-wide expression and protein profiles provide powerful tools for large-scale analyses of gene interaction and identification of pathways underlying cells' response to perturbations. Clustering algorithms, which identify distinct patterns by grouping genes with similar expression profiles, are the most widely used tools for gene expression data analysis. Although valuable, cluster analyses do not provide a complete picture of cellular processes, and more elaborate statistical and computational methods for determining these pathways (or genetic networks) must be developed. I propose a mathematical framework to describe the causal or logical relationships between gene expressions that exist in such pathways. Structural equation models are a powerful generalization of earlier statistical approaches, such as path analysis, and a widely used tool for causal inference. I employ structural equations to model relationships among genes using gene expression profiles. Solutions to the structural equations identify the pathway underlying a given causal structure or the logical relationship among the genes in the pathway. I use the method of generalized least squares to estimate the parameters in the structural equation models. Structural equation models can also assist in quantitative analysis of pathways. I have applied the proposed structural equation models to analyses of the yeast cell cycle and colon cancer apoptosis.

Yakhini, Zohar [65]

Statistical benchmarking and class discovery in gene expression data

Amir Ben-Dor[1], Nir Friedman[2] & Zohar Yakhini[1]

[1]*Agilent Laboratories, Haifa, Israel*
[2]*Hebrew University, Jerusalem, Israel*

Recent studies have elucidated putative disease subtypes from gene expression data[1–3]. In the data analysis phase of this process we seek a partition of the set of sample tissues into, say, two statistically meaningful classes. All current algorithmic approaches to this problem are clustering-driven, using similarity measures that account for all measured genes. Such methods fail to discover classes supported on small subsets of measured genes. Consider a candidate subtype. Label each sample in the data + if it is in the class or – otherwise. Some genes have dramatic + to – expression-level differences. Under a null model, in which a vector of labels of the appropriate composition is uniformly drawn, we can assign *P* values to all + to – expression-level differences. For actual biological classes we typically observe an overabundance of differentially expressed genes (compared with the null model). Efficient methods for calculating exact score distributions, under this null model, allow for a new approach to class discovery. For candidate partitions of the sample set we compute the abundance of differentially expressed genes. Statistical significance is assigned to the observed abundance using the aforementioned methods. Simulated annealing search heuristics (in the space of all possible classes) find the highest-scoring partitions. Thus grouping is based on subsets of the genes rather than on the entire set. The calculations are accurate and efficient, in contrast to sampling-based methods. We will discuss statistical and algorithmic approaches and use actual gene expression data to demonstrate the discovery process.

1. Alizadeh, A. *et al. Nature* **403**, 503–511 (2000).
2. Bittner, M. *et al. Nature* **406**, 536–540 (2000).
3. Golub, T. *et al. Science* **286**, 531–537 (1999).

Yamazaki, Victoria [66]

Efficient data mining of proteins involved in carcinogenesis by functional classification using Interpro

Dunrui Wang, Victoria Yamazaki & Tom Tang

Hyseq Inc., Sunnyvale, California, USA

Data mining in most cases relies on and is limited by a biologist's experience and knowledge. It is concentration- and time-intensive. We have developed an original hierarchical protein family classification, based primarily on protein sequence motifs, functional domains and cellular localization. This protein classification schema has been generated by classifying entries in Interpro 1.2, an integrated resource of protein families, domains and sites[1]. By using this new classification, one can effectively and rapidly mine data on proteins involved in oncogenesis. This classification is especially tailored toward the biopharmaceutical industry and drug discovery efforts, and it has been extensively used in Hyseq's internal data-mining processes. The hierarchical protein classification has 9 main classes, 56 subclasses, and 3,052 Interpro entries. These Interpro entries represent 574 domains, 2,418 families, 46 repeats and 14 post-translational modification sites from clustering PRINTS, PROSITE, ProDom, SWISS-PROT, TrEMBL and Pfam data. We generate Pfam models by multiple protein sequence alignments and express them mathematically in the form of hidden Markov models. To demonstrate the utility of our classification schema, we searched the SWISS-PROT and TrEMBL databases with Pfam models. Of the 31% of protein sequences that had