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 significant Pfam hits, 4,520 sequences were in the cancer subclass. At present 59 Interpro entries exist in this subclass. These entries include breast cancer susceptibility proteins, retinoblastoma protein domains, p53 tumor antigen, Xeroderma pigmentosum proteins and the Burkitt's lymphoma receptor.

1. Apweiler R. *et al. Nucl. Acids Res.* **29**, 37–40 (2001).

Yang, Ping [67]

Neutrophil elastase gene in lung cancer development: evidence from molecular genetics and clinical epidemiology

Ping Yang, Ken Taniguchi, Claude Deschampes, Eric Bass, Rebecca Meyer & Wanguo Liu

Mayo Foundation, Rochester, Minnesota 55901, USA

Neutrophil elastase (NE) is a powerful protease capable of degrading various protein components of the extracellular matrix, coagulation and component cascades. Local production of NE is involved in invasion and associated with a poor lung cancer prognosis. We propose that NE is important in lung cancer development, especially in the invasive phase. We tested this hypothesis using 344 cases and 299 controls. We detected single-nucleotide polymorphisms at the NE locus by denaturing high-performance liquid chromatography analysis, DNA sequencing and cloning of the NE gene promoter region. We performed transfection analysis, using relative luciferase activity on a human lung cancer cell line, to determine the regulatory role of the detected single-nucleotide polymorphisms in setting transcription levels of NE promoter. Two new single-nucleotide polymorphism markers, REP_A (T and G) and REP_B2 (G and A), were detected. There was no GG at REP_A and very little AA at REP_B2. TT and TG were contrasting allele types at REP_A, and GG and AA/AG were contrasting allele types at REP_B2. Measured by odds ratio, the TT type at REP_A or GG type at REP_B2 was associated with a 2.3 or 1.4 times higher risk of developing lung cancer than the TG and AA/AG types, respectively. When assessing the combined effects of the high-risk alleles, the odds ratio was 24.8. We demonstrated a 1.9-fold increase of relative luciferase activity in the T-G construct compared with the G-A construct, providing evidence that the TT-GG type correlates with a high NE level. Our findings indicate an important role for NE in lung cancer development.

Yang, Xiang-Jiao [68]

Monocytic leukemia zinc-finger protein MOZ and its related factor MORF are new histone acetyltransferases

Nathalie Champagne, Nadine Pelletier & Xiang-Jiao Yang

Molecular Oncology Group, McGill University Health Centre, Montreal, Quebec H3A 1G4, Canada

The monocytic leukemia zinc-finger protein (MOZ) gene is rearranged in t(8; 16)(p11; p13) and inv(8)(p11q13) associated with acute myeloid leukemia. The other fusion partners involved are CBP and TIF2, both of which are known transcriptional coactivators with intrinsic histone acetyltransferase activity. We have cloned MORF, a new human protein related to MOZ, and demonstrated that MOZ and MORF have intrinsic histone acetyltransferase activity. Moreover, like MORF, MOZ possesses a weak transcriptional repression domain at its N-terminal part and a strong activation domain at its C-terminal part. These results indicate that MOZ and MORF are acetyltransferases involved in regulating transcription and thereby shed light on how aberrant MOZ proteins lead to leukemogenesis.

Yeung, Ka Yee [69]

Transcriptional analysis of Barrett's epithelium and normal gastrointestinal tissues

Ka Yee Yeung[1], Michael Barrett[2], Jeff Delrow[2], Patricia Blount[2], Brian Reid[2] & Peter Rabinovitch[3]

[1]*Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA*
[2]*Fred Hutchinson Cancer Research Center, Seattle, Washington, USA*
[3]*Department of Pathology, University of Washington, Seattle, Washington, USA*

Barrett's esophagus is a premalignant condition caused by chronic acid reflux in which the normal squamous epithelium of the esophagus is replaced by a metaplastic columnar epithelium. Of interest is the distinction between neoplastic Barrett's epithelium (BE) and surrounding normal tissues of the upper gastrointestinal tract. For example, although it arises in the esophagus, BE more closely resembles the epithelium of the duodenum at the histological level. We compared the transcriptional profile of BE to the profiles of normal upper gastrointestinal tissues, including gastric epithelium, squamous epithelium of the esophagus and duodenal epithelium. We collected endoscopic biopsies from each tissue from a series of patients during routine surveillance. Poly(A) + RNA was prepared from pooled samples (2–4 patients per pool) of BE (four pools), esophageal squamous epithelium (four pools), gastric epithelium (three pools) and duodenal epithelium (three pools) and used to interrogate Affymetrix HU6800 and FL6800 chips. We found no difference in the correlation coefficients between neoplastic BE and each of the normal tissues. In addition, we searched for tissue-specific patterns of gene expression in the normal tissues and the neoplastic BE. We compared the performance of several clustering algorithms. Our results suggest that the data set consists of approximately eight clusters, and the best performance was obtained using CAST and k-means. From this analysis, we applied CAST to identify eight clusters, among which some were specific for squamous epithelium (203 genes), duodenal epithelium (211 genes), gastric epithelium (105 genes) and BE (36 genes).

Yuan, Bo [70]

Physical mapping and functional annotation of 60,000 human genes

Degen Zhuo[1], Wei Zhao[1], Hee-Yung Yang[2], Jian-Ping Wang[1], Russell Sears[1], Do-Hun Kwon[1], David Gordon[1], Solomon Gibbs[1], Dai Dean[2], Troy Baer[3], Don Stredney[3], Al Stutz[3], Ralf Krahe[1], Fred Wright[1] & Bo Yuan[1]

[1]*Bioinformatics Group, Human Cancer Genetics Program, Ohio State University, Columbus, Ohio, USA*
[2]*Labbook.com, Columbus, Ohio, USA*
[3]*Ohio Supercomputer Center, Columbus, Ohio, USA*

The recent release of the first draft of the human genome provides an unprecedented opportunity to integrate all human genes and their functions within a complete positional context. However, at least four significant technical hurdles remain: to create a complete and nonredundant human transcript index, to assemble the still-fragmented human genome draft, to place the individual transcript indices accurately on the human genome and to annotate all human genes functionally. We report the extension of the UniGene database through the assembly of its sequence clusters into nonredundant sequence contigs. The resulting consensus was aligned to the draft genome. We determined a unique location for each transcript within the human genome by the integration of the restriction fin-