DRUG DISCOVERY TODAY TARGETS

Determination of tumour marker genes from gene expression data

Miroslava Cuperlovic-Culf, Nabil Belacel and Rodney J. Ouellette

Cancer classification has traditionally been based on the morphological study of tumours. However, tumours with similar histological appearances can exhibit different responses to therapy, indicating differences in tumour characteristics on the molecular level. Thus, development of a novel, reliable and precise method for classification of tumours is essential for more successful diagnosis and treatment. The high-throughput gene expression data obtained using microarray technology are currently being investigated for diagnostic applications. However, these large datasets introduce a range of challenges, making data analysis a major part of every experiment for any application, including cancer classification and diagnosis. One of the major concerns in the application of microarrays to tumour diagnostics is the fact that the expression levels of many genes are not measurably affected by carcinogenic changes in the cells. Thus, a crucial step in the application of microarrays to cancer diagnostics is the selection of diagnostic marker genes from the gene expression profiles. These molecular markers give valuable additional information for tumour diagnosis, prognosis and therapy development.

Cancer is a genetic disease developed through the accumulation of abnormalities and aberrations in gene expression. Several different tumour-specific mutations, DNA amplifications and translocations lead to the development of different cancers with diverse clinical behaviour in terms of both therapy response and disease progression. Thus, it is of fundamental importance to precisely and accurately assign a given tissue sample to a diagnostic category. Currently, cancer diagnoses used to determine prognoses and guide therapy decisions are based on morphological features of the tumour, which are sometimes complemented by single-gene or single-protein assays. Examples include the prostate-specific antigen for prostate cancer diagnosis, CA 125 for ovarian cancer diagnosis and the carcinoembryonic antigen for colorectal cancer. However, tumours with similar histopathological appearances can follow significantly different clinical courses and possibly respond differently to therapy [1].

Molecular tumour classification on the basis of genomics experiments offers hope for a more individualized and more accurate diagnosis, prognosis and determination of treatment options [1,2]. The histological origin of primary tumours, their metastatic potential and optimal treatment might be discernable by gene expression analysis of tumour biopsies. As tumours are caused by genetic alterations, detailed gene expression data from genomic measurements are expected to be sufficient for the development of novel tumour classifications based on molecular characteristics. This novel classification could lead to a more complete understanding of molecular variations among tumours and, hence, better diagnoses and treatment strategies for the disease [3].

Miroslava Cuperlovic-Culf*
Rodney J. Ouellette
Beauséjour Medical Research Institute
Nabil Belacel
National Research Council, Canada,
Institute for Information, Technology-e-Health
*e-mail:
miroslavac@health.nb.ca

1359-6446/04/$ – see front matter PII: S1359-6446(04)03393-3

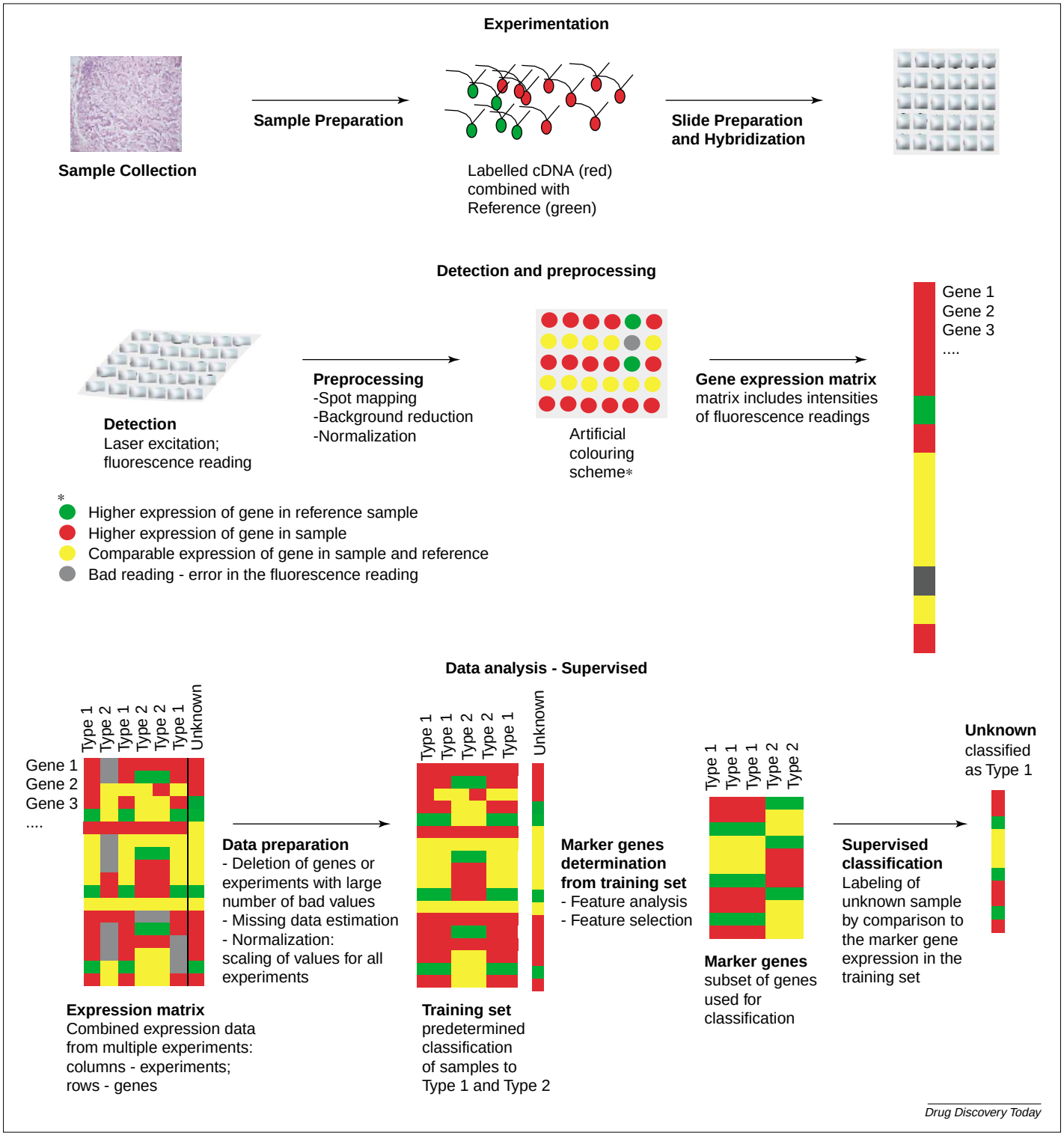

FIGURE 1

Schematic representation of microarray technology aimed at sample classification. The steps shown lead from sample collection (either tissues or cell culture samples), through RNA extraction, labelling and hybridization to DNA arrays, to detection and pre-processing (intensity determination). Data analysis includes the assembling of a gene expression matrix (database) from multiple experiments and the preparation of the data in the matrix for analysis. The missing values for some genes in some experiments (caused by errors in the array or hybridization) can be estimated from values of other genes. Once the gene expression matrix is prepared, a subset of classified samples is assembled into a training set that is used for the determination of marker genes and for classification. The gene expressions of marker genes from the training set are compared with their expressions in the unknown sample set. The unknown sample is then assigned to a class (type) with the most similar expression pattern. Although this figure represents spotted DNA microarray technology, the presented steps are comparable for any DNA microarray platform. Other array methods such as comparative genomic hybridization, cell arrays, protein arrays and glycol-chips follow the same logic.

Microarrays allow the simultaneous study of gene expression of all or a large portion of a genome of interest. In a typical microarray experiment, total RNA or mRNA is extracted from a sample (tissues or cells), labelled by