African Journal of Biotechnology Vol. 11(86), pp. 15188-15199, 25 October, 2012
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.2471
ISSN 1684-5315 ©2012 Academic Journals

Review

Genomic and proteomic analysis with dynamically growing self organising tree (DGSOT) for measuring clinical outcomes of cancer

S. Nkhwa[1] and F. M. Fadlelmola[1,2,3]**

[1]Department of Biomedical Sciences, School of Medicine, University of Botswana,
Private Bag 00713, Gaborone, Botswana.
[2]Bioinformatics Department, Faculty of Computer Science, Future University, P.O. Box 10553, Khartoum, Sudan.
[3]Molecular Biology Department, Institute of Endemic Diseases, University of Khartoum, P.O. Box 11111,
Khartoum, Sudan.

Accepted 11 June, 2012

Genomics and proteomics microarray technologies are used for analysing molecular and cellular expressions of cancer. This creates a challenge for analysis and interpretation of the data generated as it is produced in large volumes. The current review describes a combined system for genetic, molecular interpretation and analysis of genomics and proteomics technologies that offers a wide range of interpreted results. Artificial neural network systems technology has the type of programmes to best deal with these large volumes of analytical data. The artificial system to be recommended here is to be determined from the analysis and selection of the best of different available technologies currently being used or reviewed for microarray data analysis. The system proposed here is a tree structure, a new hierarchical clustering algorithm called a dynamically growing self-organizing tree (DGSOT) algorithm, which overcomes drawbacks of traditional hierarchical clustering algorithms. The DGSOT algorithm combines horizontal and vertical growth to construct a mutlifurcating hierarchical tree from top to bottom to cluster the data. They are designed to combine the strengths of Neural Networks (NN), which have speed and robustness to noise, and hierarchical clustering tree structure which are minimum prior requirement for number of clusters specification and training in order to output results of interpretable biological context. The combined system will generate an output of biological interpretation of expression profiles associated with diagnosis of disease (including early detection, molecular classification and staging), metastasis (spread of the disease to non-adjacent organs and/or tissues), prognosis (predicting clinical outcome) and response to treatment; it also gives possible therapeutic options ranking them according to their benefits for the patient.

Key words: Genomics, proteomics, microarray, dynamically growing self-organizing tree (DGSOT).

INTRODUCTION

At a functional level, cancer is both a proteomic and a genomic disease (Hanahan and Weinberg, 2000). Cancer is a highly variable disease with multiple heterogeneous genetic and epigenetic changes. A cancer genetic defect is selected during cancer progression because the defect ultimately alters the protein network generating a survival advantage for the cancer cell (Hunter, 2000). The development and progression of cancer (Vogelstein and Kinzler, 1993; Weinberg, 1995; Levine, 1993) results from progressive alterations of

*Corresponding author. E-mail: faisal.mohamed@hotmail.com.

sequence of genetic and epigenetic changes which promotes the malignant transformation of the cell by disrupting key processes involved in normal growth control and tissue homeostasis. Three types of genes are responsible for tumorigenesis when undergoing alterations. Oncogenes are involved in promoting cell growth. Tumour suppressor genes are negative regulators of growth or other functions that may affect invasive and metastatic potential such as cell adhesion and regulation of protease activity. Stability genes control the rate of DNA mutation and contribute to the development of cancer when they incur alterations which result in mutations in the oncogenes or the tumour suppressor genes (Vogelstein, 2004; Bielas et al., 2006).

The term genomics refers to a comprehensive analysis of gene expression of large number of genes. This is achieved by assessing relative or semi-quantitative amounts of RNA in biological specimens. Genomic analysis uncovers mutations, deletions and epigenetic alterations that directly or indirectly alter gene expressions. Understanding proteins and their modifications may elucidate properties of cellular behaviour that may not be reflected in analysis of gene expressions is referred to as proteomics. Because of the multitude of potential post-translational modifications, compartment-talisation of proteins and the formation and regulation of multi-protein complexes, proteomic technologies and studies are rendered be more technically challenging (Chung et al., 2007). Genomics and proteomics technologies are a new and powerful tool for studying the molecular basis of interactions on a scale that is impossible using conventional analysis. These techniques make it possible to examine the expression of thousands of genes and proteins simultaneously. Most of the applications of genomics and proteomics technology come from the field of cancer research. Genomic and proteomic technologies facilitate the analysis of genetic and molecular alterations of thousands of tissue speci-mens in parallel, though not at the same time. examples include analysing the frequency of genetic and molecular alterations in large tumour materials to classify tumours according to their sites of origin (Su et al., 2001; Ramaswamy, 2001; Bloom et al., 2004), exploring the tumour's progression (van't Veer et al., 2002; Shipp et al., 2002; Leung et al., 2002), discovering previously unre-cognised subtypes of cancer (Adam et al., 2001; Carter et al., 2002; Rosty, 2002; Bittner et al., 2000; Perou et al., 2000) identifying predictive or prognostic factors and validating newly discovered genes as diagnostic and therapeutic targets (Kallioniemi et al., 2001; Scherf et al., 2000).

Analysis of these data requires the use of a system with algorithms that employ clustering technology. Clustering is a useful unsupervised method for identifying patterns from large data sets. Hierarchical clustering is of advantage when no prior knowledge of data sets is available and the clusters are not pre-defined. It can find different levels of patterns of data, identify trends in the data, and generalise the information. Currently a dynamically growing self-organising tree (DGSOT) proposed by Luo et al. (2003) uses a hierarchical cluster algorithm. To determine the true number of clusters it has a new cluster validation criterion called cluster separation, and to improve the cluster results it employs a new K-level up distribution (KLD) mechanism. The DGSOT grows in two directions, vertically and horizontally. In DGSOT each leaf represents a cluster that includes all data associated with it. The reference vector of a leaf is the centroid of all data associated with it. Each internal node represents a cluster that includes all data associated with its leaf descendants, and the reference vector of an internal node is the centroid of all data associated with its leaf descendents. In each vertical growth, the DGSOT adds two 'children' (sub clusters) to the leaf whose heterogeneity is greater than a threshold and turns it to a node. In each horizontal growth, the DGSOT dynamically finds the proper number of children of the lowest level nodes. The proper number of clusters is determined by cluster validation. Each vertical growth step is followed by a horizontal growth step after which a learning process is adopted. Each procedure is called a cycle which contains a series of epochs which consist of a presentation of all input data. Each presentation has two steps which are to find the best matching node and updating the reference vector. The input data is only compared to the leaf nodes determined by the KLD mechanism to be the best matching node which is known as the winner. The leaves whose heterogeneity is greater than a threshold will change itself to a node and create two descendent leaves. When the heterogeneity of all leaves less than a threshold T_R Vertical growth is stopped.

It is now possible to make the diagnosis of a particular cancer and cancer subtypes without examining histology. Genomic and proteomic technologies may not only eliminate diagnostic categorisation of the unknown primary cancer but may also improve the diagnostic accuracy of the current approaches. This technology also promises to lead to improvements in developing rational approaches to therapy as well as improvements in cancer diagnosis, prognosis and identification of gene and molecular sets associated with metastasis. One of the potential benefits of this technology within the next decade would be predicting who will develop cancer and how the disease will behave and respond to therapy after diagnosis (Staunton et al., 2001; Russo et al., 2003). In this study genomic and proteomics microarray technologies will be reviewed and the best technology available for analysing and interpreting the data generated will be recommended.

GENOMIC TECHNOLOGIES

The completion of the human genome sequence project has led to the study of gene expressions on a genomic scale (Brown and Botstein, 1999). Genomic technologies allow for the assessment of interactions between expressed genes to obtain a global view of cancerous tissue in a single unbiased experiment rather than focusing on one or a handful of genes at a time (Chung et al., 2007). Thirty years of molecular biology have provided numerous examples of genes that function under specific conditions and whose expression is tightly restricted to those conditions. At the level of transcript abundance, using DNA microarrays, the regulation of gene expression as well as the tight connection between the function of a gene product and its expression pattern makes it easy to measure the transcripts for every gene at once.

By changing the level of transcription of specific genes, promoters function as transducers, responding to inputs of information about the identity, environment and internal state of a cell. Therefore, from the profile of transcripts obtained by DNA microarray, we can learn what information is transduced by the promoter of each gene. In a cell there are sets of genes which when expressed determine what the cell is made of, what biochemical and regulatory systems are operative, and how the cell is built (Brown and Botstein, 1999).

Genomic microarray platforms

It was first observed that single stranded DNA binds strongly to nitrocellulose membranes in a way that prevents the strands from re-associating with each other but permits hybridization of complementary RNA (Gillespie, 1965). In eukaryotes it was used to measure the number of copies of repeated genes, like those for ribosomal RNAs and transfer of ribonucleic acid (tRNAs) and to measure changes in the number of copies during processes such as amplification (Ritossa et al., 1971). Cloning technology provided a way for finding those clones which included specific sequences (Grunstein, 1975). This method was the direct antecedent of the blotting methods. Dot-blotting methods are more relevant to microarray (Kafatos, 1979). One could obtain information about the quantity of a particular message present in each RNA pool from immobilisation of RNA samples on the same matrix (Duggan et al., 1999).

This field was evolved from Edwin Southern’s key insight that labelled nucleic acid molecules used to interrogate nucleic acid molecules attached to a solid support (Southern, 1975). Northern blots (hybridization of RNA-DNA) and Southern blots (hybridisation of DNA-DNA) rely on hybridization between nucleic acids. Complementary gene sequences recognise each other and detect the presence or absence of DNA or RNA of interest in a sample using probes labelled in a variety of ways of detectors (fluorogenic, radiological or chemiluminescent detectors) (Chung et al., 2007). The two main types of microarray systems used today are complementary DNA (cDNA) and oligonucleotide microarrays (Schulze and Downward, 2001). cDNA array probes which are relatively long molecules are usually products of polymerase chain reaction (PCR). The PCR transcripts are generated from clone collections, using either vector-specific or gene-specific primers or are generated from cDNA libraries. They are printed on glass slides or nylon membranes as spots at defined locations. The spots are normally about 100 to 300 µm in size with the same distance spacing. Arrays consisting of more than 30,000 cDNAs can be fitted onto the surface of a conventional microscope slide using this technique. This type of microarray is mostly used for large scale screening and expression studies.

Oligonucleotide arrays consist of short 20 to 25 polymers synthesized *in situ*, either by photolithography onto silicon wafers (high-density-oligonucleotide arrays from Affymetrix), [http://www.affymetrix.com] or by ink-jet technology was developed by Rosetta Inpharmatics, (http://www.rii.com) and licensed to Agilent Technologies. They can also be directly synthesized onto glass slides. With oligonucleotides, probes can be designed to identify a unique part of a given transcript, making the detection of closely related genes or splice variants possible. The arraying of pre-synthesized longer oligonucleotides (50 to 100 polymers) has recently been developed to counteract the disadvantages of short oligonucleotides which may sometimes result in less specific hybridization and reduced sensitivity (Schulze and Downward, 2001; TJ, 2003; Relogio et al., 2002). The advantage of synthetic oligonucleotides is that the sequence information alone is sufficient to generate the DNA to be arrayed therefore no time consuming handling of cDNA resources is required (Kane, 2000). This microarray is used for the detection of mutations, gene mapping and expression studies and allows for the differential detection of gene family members or alternative transcripts that are not distinguishable by cDNA microarrays (Lipshutz, 1999).

Schulze and Downward (2001) provided a helpful schematic overview of probe array and target preparation for spotted cDNA microarrays and high density oligonucleotide microarrays. To distinguish between dot blots and DNA microarrays one can note that DNA microarrays use an impermeable rigid substrate such as glass which has a number of practical advantages over the porous membranes and gel pads used for dot blots (Khrapko et al., 1989). The development of methods for high-density spatial synthesis of oligonucleotides as well as the use of non-porous solids like glass sparked the explosion of interest in array technologies was facilitated by miniaturization and fluorescence-based detection (Lander, 1999). The first glass slide arrays were produced by Brown and colleagues at Stanford University (http://cmgm.stanford.edu/pbrown/index.html). They pioneered protocols for robotically spotting up to about 10,000 cDNAs onto a microscope slide and hybridizing with a double labelled probe. Fodor and colleagues adopted photolithographic masking techniques use in semiconductor manufacture to produce arrays with 400,000 distinct oligonucleotides (Brown and Botstein, 1999; Lander, 1999).

From Brown and colleagues the technology spread to few others who made important refinements to the micro-array technology as well as disseminating it and making available detailed protocols. They include Jeff Trent National Human Genome Research Institute (NHGRI, http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/), Vivian Cheung (University of Pennsylvania; http://w95vcl.neuro.chop.edu/vcheung) and Geoff Childs (Albert Einstein College of Medicine; http://sequence.aecom.yu.edu/bioinf/funcgenomic.html). The Pat Brown website also contains detailed specifications for building an arrayer and associated software. Additional protocols and hardware details can also be found at (http://chroma.mbt.washington.edu/mod_www/) (Bowtell, 1999). This technology further advanced to commercialisation by companies such as Affmetrix. The steps for performing microarray experiments are as follows:

1. mRNA from cells or tissue is extracted. Isolated RNA from samples of interest and a reference RNA are reverse transcribed into cDNA.
2. These are then labelled. In the case of spotted arrays the process of gridding is not accurate enough to allow comparison between different arrays therefore each is labelled with one of two spectrally distinct fluorescent dyes such as Cy3 or Cy5 to allow mRNAs from two different cell populations or tissues to be labelled in different colours, mixed and hybridized to the same array, which results in competitive binding of the target to the arrayed sequences.
3. Following labelling and purification, the labelled samples are pooled together to provide a target mixture containing cDNA representing a reference RNA as well as cDNA representing an experimental sample of interest.
4. This mixture is hybridized to the cDNA elements on the array surface of a single microarray. The labelled cDNA in the pooled samples hybridizes to probes with complementary sequence immobilized on the array.
5. Following hybridization, the microarray is washed to remove unbound and non-specific material.
6. After hybridization and washing the slide is scanned using two different wavelengths, corresponding to the dyes used, and the intensity of the same spot in both channels is compared. The wavelength radiation emissions are detected via a photomultiplier tube or a charged couple device camera.
7. The slide can then be visualised with a laser based device that measures the fluorescence of the two spectrally distinct fluorescent dyes at each of the probe spot positions. Most often a 16-bit TIFF image is generated representing the fluorescent signal intensities.
8. The fluorescent signal of the two hybridized samples at each probe position as well as determining background levels on the array can be compared using image analysis software. The intensity ratio between the two fluorescent dyes at each probe position is calculated using these signals and background values, thus providing an assessment of RNA abundance in the two samples as the ratio is indicative of relative amount of RNA for a particular probe in each of the two samples (Chung et al., 2007; Schulze and Downward, 2001).

PROTEOMIC TECHNOLOGIES

Recent advances in areas such as genome sequencing, robotics, bioinformatics and proteomics has led to an explosion of interest in the field of protein and antibody arrays. Compared to DNA arrays the generation of protein arrays is more costly and labour intensive. The medical rationale behind investigating changes in the structure or abundance of proteins is to improve our understanding of normal disease process as such changes can lead to disease. Advantages of analysing biological processes at protein level are that gene protein dynamics are non linear, and there is no reliable correlation between gene activity and protein abundance so it is difficult to predict protein dynamics, structure or interactions using genetics or DNA approaches (Anderson and Seilhamer, 1997; Cahill, 2001). Proteomics is the large scale study of expression, function and interactions of proteins (Geysen and Barteling, 1984). Proteomics aims to characterise the information flow within the cell and the organism through protein pathways and networks. The information flow here is mediated by protein-protein interactions that is, proteins deliver packets of information by modifying a protein binding partner for example, by phosphorylation/dephosphorylation, cleavage or alteration of its conformation (Liotta and Petricoin, 2000; Ideker et al., 2001; Schwikowski et al., 2000; Legrain et al., 2000; Blume-Jensen and Hunter, 2001; Pawson, 1995; Petricoin et al., 2002).

New biomarkers and therapeutic targets have been developed *via* different proteomic methods. These include 2-dimensional gel electrophoresis, 2-color and mass spectrometry (matrix-assisted laser desorption/ionization time-of-flight). These methods are time consuming, equipment is expensive and it requires experienced investigators. Tissue microarray is a more cost effective proteomic method which is available to most large pathology laboratories (Bubendorf and Koivisto, 1999; Popper and Kothmaier, 2008). Basically two general strategies have been pursued in proteomics. Firstly, function based microarrays, which are protein microarrays that assess protein interactions and biochemical activities by examining protein function in high throughput by printing a collection of target proteins on the array surface. The second type is abundance based microarrays. These are tissue microarrays which measure the abundance of specific bio-molecules using analyte-specific reagents (ASRs) such as antibodies (LaBaer and Ramachandran, 2005). Analysis of complex biological systems requires information that goes beyond protein expression level. Once again microarray technology has the great potential to provide us with powerful tools to identify and quantify proteins and to study their function in global perspectives (MacBeath, 2002; Templin et al., 2002, 2003).

Proteomic microarray platforms

The two main microarray platforms used in proteomics are; tissue microarrays (TMAs) and protein microarrays. TMAs were developed to address limitations of conventional techniques and to provide a major step forward in pathology research with subsequent potential utility in diagnostics and prognostic pathology (Kononen et al., 1998; Bubendorf et al., 1999). With TMA, molecular alterations in thousands of tissue specimens can be analysed in at the same time. TMA construction utilizes cylindrical core specimens from up to 1000 formalin fixed paraffin embedded (FFPE) tissue blocks containing hundreds of tissue cores on a single glass slide. These are then arrayed at high density into a recipient TMA block (Kallioniemi et al., 2001; Giltnane and Rimm, 2004). Up to 200 sections can be made and analysed by immunohistochemistry, *in situ* hybridization or immunoflorescence from a single array block. Thousands of replicate TMA slides can be constructed by sampling each donor block multiple times and positioning the tissues at identical coordinates in all TMAs (Ullmann et al., 2004). In a single experiment molecular characteristics of up to a 1000 specimens can be examined at once. The analysis carried out on TMAs extends the information available from gene expression microarrays by providing information on cellular origin of the molecular targets. TMAs are usually constructed from archival formalin fixed tissue materials which is a significant advantage as such specimens cannot be used in other high through technologies like cDNA microarrays (Kallioniemi et al., 2001).

Kallioniemi et al. (2001) have provided step by step instructions for technology for tissue microarray construction. In TMAs the pattern of protein expression can be studied in different cell compartments (nuclear, cytoplasmic, and membranous) and the distribution of proteins in tumour cells, stroma and adjacent normal parenchyma, including normal bronchial and alveolar epithelium can also be observed (Popper and Kothmaier, 2008). Typical tissue microarray technologies include multitumor, progression, prognosis and cryomicroarrays. Multitumor microarrays are composed of samples from multiple histological tumour types (Schraml et al., 1999). In progression microarrays stages of different tumour progressions within a given organ such as prostate, breast or kidney are contained in samples (Bubendorf and Koivisto, 1999; Kononen et al., 1998; Moch et al., 1999). Prognosis microarrays contain tumour samples from patients for whom clinical follow up data are available. Some studies have been published comparing molecular data with clinical end points (Kallioniemi et al., 2001; Barlund et al., 2000; Richter et al., 2000; Simon et al., 2001). Cryomicroarrays are superior to formalin fixed tissues in terms of RNA and protein integrity. The array uses frozen tissue embedded in an optimal cutting temperature compound (Russo et al., 2003; Fejzo and Slamon, 2001).

TMAs have several limitations: for example the cylindrical tissue cores do not permit a complete pathological evaluation. Immunohistochemistry (IHC) assay is limited to known candidate proteins for which specific and high quality antibodies are available. Another limitation is the ability to validate the antibody using IHC for example, staining cross reactive proteins rather than the analytes of interest. Although there are many automated tools available for scoring and standardisation, it can be subject to inter examiner inconsistency. Recovery of intact and good quality genomic and proteomic material is difficult with formalin fixed specimens because of the intense cross-linking induced between the bio-molecules by formalin fixation (Chung et al., 2007; Wang et al., 2001; Bauer et al., 2000). Protein microarrays are composed of immobilised protein spots which contain a set of 'bait' molecules (Liotta et al., 2001; MacBeath, 2002; Zhu and Snyder, 2003). Antibodies, cells or phage lysates, recombinant protein or peptide, a drug or nucleic acid may be displayed on the array spot. The array is queried with a labelled antibody or ligand (probe) or analytes of interest contained in an unknown biological sample (for example, cell lysate or serum sample) (Zhu and Snyder, 2003; Lal et al., 2002; Templin et al., 2002; Wilson and Nock, 2003; Paweletz et al., 2001; MacBeath and Schreiber, 2000; Humphery-Smith et al., 2002; Petach and Gold, 2002). Molecules are then tagged with a signal generating moiety which generates a pattern of positive and negative spots. The intensity of the signal is proportional to the quantity of the applied query molecules bound to the 'bait' for each spot on the array. Images can then be captured and analysed.

Protein microarrays can be used in the analysis of the interactions between proteins and other proteins, low molecular weight compounds, peptides, oligosaccharides or DNA (for example, tumour proteins are compared with those of normal adjacent tissues or to standard protein lysates). The microarrays allow for the identification of a large number of target proteins from a minute amount of sample within a single experiment (MacBeath, 2002;

Templin et al., 2002, 2003).

Protein microarray platforms can be categorised into two groups. These are called forward phase arrays (FPA) and reverse phase arrays (RPA) depending on whether the analyte is captured from solution phase or bound to solid phase (Liotta et al., 2003; Sheehan et al., 2005). In forward phase arrays antibodies for a target protein usually referred to as capture molecules, are immobilised onto a substratum like such as a glass slide similar to DNA microarrays. One type of immobilised antibody is contained in each spot. The target protein is contained in the cellular lysate (example protein lysate from tumour or normal tissue). The bound protein can then be detected using a fluorochrome labelled secondary antibody. This method multiple analytes can be measured at once. In *reverse phase arrays* complex protein lysates/mixtures are immobilised onto glass slides. The slides are then incubated with specific antibodies against a protein of interest. A single analyte end point is measured and directly compared against a large number of samples on a single glass slide. Fluorescent, isotopic and chemiluminescent horseradish peroxidase/luminol systems can be used for detection (Liotta et al., 2003). This method is limited to the availability of candidate proteins of interest and availability of specific and high quality antibody against the protein of interest. Self assembling microarrays are the new emerging protein microarray platforms. They promise a much wider and easier use of the technology to probe protein interaction and function (LaBaer and Ramachandran, 2005; Ramachandran et al., 2004).

ANALYSIS OF MICROARRAYS AND THEIR CLINICAL OUTCOMES

Genomics data analysis

Thousands of data points are generated in a typical genomic microarray and this creates serious challenges for storing and processing data. In order to manage the information on the genes represented on the array, construction of databases is required. After completion of data acquisition, the appropriate data filtering normalisation and background correction approach is most appropriate for the given data set decided on. Various methods are available for detecting and quantitating gene expression levels. These include sequencing of complementary deoxyribonucleic acid (cDNA) libraries and serial analysis of gene expressions (SAGE) (Adams et al., 1991; Okubo et al., 1992; Velculescu et al., 1995; McAdams and Shapiro, 1995). Ermolaeva et al. (1998) developed software that is capable of both analysing microarray data and linking to databases such as Entrez and UniGene. This software can be found and downloaded at (www.nhgri.nih.gov/DIR/LCG/15K/HTML/). A sophisticated program for analysing microarray data (Gem Tools) was developed by Synteni. Silicon Genetics provides the Gene Spring package for analysing data from Affymetrix GeneChip and other microarray experiments (http://www.sigenetics.com). Other commercial readers and arrayers provide software for data analysis and mining (Duggan et al., 1999).

These software analysis methods can be broken down into two categories *viz.* supervised and unsupervised analysis. An *unsupervised analysis* is mostly used for molecular classification of tumours or class discovery based on gene expression patterns. Other unsupervised analytical tools are self organising maps (SOM), K means clustering and principle component analysis (Tamayo et al., 1999; Tavazoie et al., 1999; Pomeroy et al., 2002). *Supervised analyses* are usually applied for gene selection and class prediction to determine sets of genes that distinguish one group from another (Wu, 2001).

Application of clinical outcomes of genomics

Classification of diagnosis

Gene expression profiling has been evaluated to augment the accuracy of diagnosis, especially the 3 to 5% of new cancer cases with unknown primary origin. Ramaswamy et al. (2003) showed the feasibility of comprehensive molecular cancer diagnosis by analysing 218 tumour samples from 14 common tumour types and 90 normal samples for multi-class cancer diagnosis based on tumour gene expression analysis. Their results gave them an overall accuracy of 78% as poorly differentiated cancers could not be accurately classified with their corresponding organs. Tothill et al. (2005) evaluated the identification of cancer origin from metastatic tumours of unknown primary origin based on the gene expression profile. The tumour type could be predicted with 89% accuracy using a support vector machine (Su et al., 2001; Ramaswamy, 2001; Giordano et al., 2001; Tothill et al., 2005; Briasoulis and Pavlidis, 1997). Golub et al. (1999) introduced a concept of identifying previously unknown tumour subtypes and predicting a tumour to be within an already defined class based on gene expression profile using acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL) as a model. Other large studies have been published to show that microarray can diagnose and identify subcategories of hematologic malignancies (Hofmann et al., 2001; Yeoh et al., 2002; Armstrong et al., 2002; Ferrando et al., 2002; Shimada et al., 2002; Valk et al., 2004; Bullinger et al., 2004).

Prediction of metastasis

The understanding of metastasis is an important area in clinical cancer research; mechanisms of metastasis have been scientifically researched for years. Some insight into the complex processes undergone during metastasis has been provided by genomic technology, and it has proven its ability to predict metastatic behaviour from the analysis of primary tumours (van de Vijver et al., 2002). Dhanasekaran et al. (2001) used 9984 element spotted microarrays and were able to distinguish normal prostate, benign prostatic hyperplasia (BPH), localised prostate cancer and metastatic cancer samples. Zajchowski et al. (2001) in a recent study identified 24 genes differentially expressed between weakly and highly invasive breast cancer cell lines and showed that their RNA expression profiles were sufficient to predict the aggressiveness of previously uncharacterized cell lines.

Using a sub-megabase resolution tiling array (SMRT), (Fadlelmola et al., 2008) demonstrated how microarray-based comparative genomic hybridization (Array CGH) revealed gains and losses of 9 novel regions not previously reported in the literature in Hodgkin Lymphoma (HL) cell lines L428 and KMH2. Gene mapping to these regions include cell cycle-associated genes, signaling pathway genes, genes encoding tight junction proteins *CLDN4* (claudin4) and Jak/Stat signaling pathway and tumour suppressor gene *ING3* (Fadlelmola et al., 2008).

Classification for prediction of recurrence and survival

Clinicians routinely observe variable clinical outcomes within patients with comparable histopathology, staging and treatment. This can be followed, as cancer is a heterogeneous disease. Tumour subclasses defined by expression profiling can predict disease-free and overall survival of patient (Sorlie et al., 2001). Many candidate markers have emerged from other studies and are being further investigated as potential diagnostic markers which could highlight risk of recurrence after medical interventions. There are candidate markers of prostate cancer such as proto-oncogene PIM1. Dhanasekaran et al. (2001) showed that diminished PIM1 expression on the immunohistochemistry of prostate tumour samples conferred an increased risk of recurrence after surgery.

Treatment response prediction and patient selection

The microarray approach has a potentially greater predictive power than currently used approaches, and it needs to be validated in more prospective clinical studies. Unfortunately current prognostic markers do not adequately identify the most effective therapy for patients. It has been shown in a small sample of breast tumour that pre-treatment expression profiles predict clinical response to chemotherapy (Sotiriou et al., 2002). Sorlie et al. (2001) demonstrated that tumour subclasses defined by expression profiling can predict disease free and overall survival.

Proteomics data analysis

Complete proteomic analysis involves measuring the abundance, modification, activity, localisation and interaction of all the proteins in a given sample. Different detection, identification and quantisation methods have been developed for proteomic analysis, often with emphasis on those proteins with altered abundance relative to the reference sample. These include 2-color, 2-dimensional gel electrophoresis and mass spectrometry (matrix assisted laser desorption/ ionisation time of flight, surface enhanced laser desorption/ ionisation time of flight (Roepstorff, 1997; Fenn et al., 1989; Karas and Hillenkamp, 1988; Hillenkamp et al., 1991).

Mass spectrometry instruments have three components which are; an ion source to volatize and ionise the analyte, a mass analyser to separate ions based on their mass to charge ratio (m/z) and a detector to detect ions after separation. Mass spectrometry analyses biopolymers such as peptides, proteins and polynucleotide as ions. Matrix assisted laser desorption and surface enhanced laser desorption are an alternative to chip based proteome analysis and are useful for capturing and analysis of specifically labelled proteins. They are commonly used mass spectrometry techniques in translational research. With this technique the investigator does not pre-select the proteins to be examined but searches for changes in any proteins that are identified. Analysis is done with mass analysers of time-of-flight (TOF). Protein identification can also be done by electrospray ionisation tandem mass spectrometry (ESI-MS-MS). Where the proteins of interest have already been identified, antibody based affinity methods or multiple reactions monitoring tandem mass spectrometry techniques can be used (MRM-MS-MS) (Hillenkamp et al., 1991; Lahm and Langen, 2000; Yates, 1998; Ge, 2000).

Applications of clinical outcomes of proteomics

Classification of diagnosis

Using tissue microarrays, it seems some relevant molecular changes and clinical end points may be detected on an array containing just a single specimen per tumour. Estrogen receptor, progesterone receptor, p53, HER2 and S6-kinase expression/amplification in breast cancer were found to have a prognostic significance (Nocito et al., 2001). A study done by Simon et al. (2001) on urinary bladder cancer showed that cyclin

E amplification/over expression and Ki67 labelling index had prognostic significance in bladder cancer. Torhorst et al. (2001) showed that Vimentin expression in kidney cancer also has prognostic significance.

Prediction of metastasis

A protein profile that predicts the presence of metastasis in non-small cell lung cancer was identified and had an accuracy of about 75 to 85 analysis for this was done using MALDI-TOF-MS. Again using MALDI-TOF MS brain tumour tissue sections were analysed. It was found that classification based on mass spectra was more reliable than traditional histological examination (Schwartz et al., 2004).

COMBINED ANALYSIS FOR GENOMIC AND PROTEOMIC MICRO-ARRAYS

A number of tools and processes have been used or suggested for analysis of genomic and proteomic microarray data. One of the most useful and popular methods for analysing and identifying microarray patterns is the clustering technology. There are two classes of cluster algorithms, *viz* hierarchal and non hierarchal. Hierarchal clustering has been extensively used for the analysis of microarray expressions (Eisen et al., 1998; Wen et al., 1998; Sneath, 1973). The hierarchal clustering algorithm analysis produces a representation of data with a binary tree, with most patterns clustered in a hierarchy of nested subsets. This analysis method has already been applied to the study of gene expression patterns, for example Eisen et al. (1998) used the analysis to cluster two spotted DNA microarray data sets, Wen et al. (1998) used the analysis to cluster the central nervous system of gene expression data from rats, as well as Iyer et al. (1999).

The hierarchal clustering method has been observed to present some draw backs when dealing with data that contains a non-negligible amount of noise (Luo, 2003). It has been claimed that the hierarchal clustering method suffers from lack of robustness and the solutions may not be dependent on the data order as well as be unique (Tamayo et al., 1999). Another problem with this method is the real difficulty when thousands of items are analysed as they have slow runtimes which are in the best case quadratic (Hartigan, 1975). With hierarchal clustering analysis another disadvantage is that some clusters of patterns end up being based on local decisions rather than on a global picture because of the impossibility of re-evaluating the results in light of the complete clustering data (Tamayo et al., 1999).

With non hierarchal clustering methods many algorithms have been developed and applied. The use of neural networks for analysis of microarray data was proposed as a convenient alternative to hierarchal clustering methods (Tamayo et al., 1999; Toronen et al., 1999; Herrero et al., 2001). Other different clustering methods have also been recently proposed (Heyer et al., 1999). Ben-Dor et al. (1999) used a graph theoretic algorithm to extract the high probability gene structures from gene expression data. Tamayo et al. (1999) used a self organising map to analyse the expression patterns of 6000 human genes, Tavazoie et al. (1999) used K-means to cluster 3000 yeast gene microarray data, De Smet et al. (2002) used a heuristic two step adaptive quality based algorithm and Yeung et al. (2001) used a clustering algorithm based on probability models.

Kohonen (1990) came up with the idea of self organising maps which use unsupervised learning. This learning method has the advantage that no previous knowledge about the system under study is required. This algorithm was used in previous sequence analyses by Ferran and Ferrara (1991, 1992); Ferran and Pflugfelder (1993) and Ferran et al. (1994) to classify protein sequences into groups based on their dipeptide compositions. These self organising neural networks generate a mapping from high-dimensional input signal spaces to lower dimensional output topological structures. The output presents an estimate of the probability density function of the input data. This model by Kohonen (1990) was found to have some severe limitations. Fritzke (1994) then proposed the unsupervised growing cell structures algorithm. Here the number of elements in the output map increases in those regions where the input space is denser and decreases in those regions where it is very low or null.

Dopazo and Carazo (1997) proposed a self organising tree algorithm (SOTA). This algorithm is based on both self organising maps (Kohonen, 1990) and growing cell structure algorithm (Fritzke, 1994). With SOTA the output space has been arranged following a binary tree topology which allows appropriate description of relationships amongst the sequences being studied, the resultant algorithm adapts the number of output nodes arranged in a binary tree to the intrinsic characteristics of the input data set. The output nodes grow until a complete classification of every sequence in the input data set is reached, the growth of the output nodes can also be stopped at a desired taxonomic level. SOTA uses neural network mechanisms and it is robust to noise data. It is important to note that this neural network classifies sequences with high accuracy whether they are protein or nucleotides, because of the way the neurons of the network interact amongst themselves.

This type of network is also advantageous because, since sequences are coded residue by residue, all the information contained in the homologous position of the alignment is used by the algorithm. Also an advantage of SOTA is that the process of growing can be stopped at any level because the tree structure grows as a function of the hierarchal relationships among the samples.

There are two main important limitations associated with SOTA namely, it does not properly represent a hierarchal relationship and that once the data is assigned improperly to a given cluster it cannot be re-evaluated and placed in another cluster. To overcome the limitations associated with SOTA, Luo et al. (2004) proposed an algorithm called the dynamically self organising tree algorithm (DGSOT) which is a self organising neural network designed to discover the proper hierarchal structure of the underlying data. The DGSOT algorithm combines with a K-level up mechanism (KLD) to improve clustering accuracy, which produces demonstrable, qualitative improvement over traditional solutions to the hierarchal clustering problem. The main purpose of having a K-level up distribution is that early hierarchal clustering stages with data that has been improperly clustered can later have a chance to be re-evaluated during the later hierarchal growing stages thus giving a more accurate final cluster result (Luo et al., 2004). Herrero et al. (2001) compared DGSOT with SOTA using 3000 normalised yeast gene cell cycle microarray expression profiles. The results showed that the clustering result of DGSOT is more statistically significant than that of SOTA. In addition the proper hierarchal structure of DGSOT makes the clustering result more reasonable for large clusters.

DGSOT has a number of advantages. By setting different vertical growth stop thresholds its algorithm can be terminated at any hierarchal stage. DGSOT can easily display its cluster result as a dendrogram for visualisation. Biological functionality enrichment in the clustering result of DGSOT is considerably higher than the clustering of SOTA.

For large data sets the DGSOT algorithm can display high level main patterns of the data set only if visualisation of the whole hierarchal structure is difficult. Luo et al. (2004) believe DSGOT to be a robust and accurate framework for the study of microarray expression data. Khan and Luo (2005) used an existing 112 genes expression data of a rat central nervous system (CNS) to analyse with DGSOT. They used raw 9-dimensional expression data. Data was normalised to the maximal expression level among the 9 time points for each gene.

The results showed a very good hierarchical structure. Herrero and Dopazo (2002) used 6120 genes to show how they can be separated into a few class patterns *via* analysis with SOTA. SOTA was able to find the different average patterns of gene expression despite the enormous differences in the number of members in the clusters. When applying the average linkage with SOM, different classes of activation patterns, were not resolved in similar detail to SOTA. Now because DGSOT has been found to be more significantly accurate than SOTA, this would imply that the study done by Herrero and Dopazo (2002) if analysed with DGSOT will yield more accurate results than with SOTA.

CONCLUSION

From reviewing analysis methods used for analysing microarray data, it was found that currently the DGSOT is the most efficient method. It has a combination of horizontal and vertical growth, optimizing the number of sub clusters which helps the algorithm to find the right hierarchical structure of the underlying data set. Experimentally, the DGSOT algorithm has been used to cluster benchmark data sets, and has demonstrated impressive results. It was also found that DGSOT is the most suitable algorithm which can be used in an attempt to analyse both genomic and proteomic microarrays.

REFERENCES

Adam BL, Vlahou A, Semmes OJ, Wright GL, Jr. (2001). Proteomic approaches to biomarker discovery in prostate and bladder cancers. Proteomics 1:1264-1270.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, Mccombie WR, Venter JC (1991). Complementary-DNA Sequencing - Expressed Sequence Tags and Human Genome Project. Science. 252:1651-1656.

Anderson L, Seilhamer J (1997). A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18:533-537.

Armstrong SA, Staunton JE, Silverman LB, Pieters R, de Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genet. 30:41-47.

Barlund M, Forozan F, Kononen J, Bubendorf L, Chen YD, Bittner, ML, Torhorst J, Haas P, Bucher C, Sauter G, Kallioniemi OP, Kallioniemi A (2000). Detecting activation of ribosomal protein S6 kinase by complementary DNA and tissue microarray analysis. J. Natl. Cancer Inst. 92:1252-1259.

Bauer KD, de la Torre-Bueno J, Diel IJ, Hawes D, Decker WJ, Priddy C, Bossy B, Ludmann S, Yamamoto K, Masih AS, Espinoza FP, Harrington DS (2000). Reliable and sensitive analysis of occult bone marrow metastases using automated cellular imaging. Clin. Cancer Res. 6:3552-3559.

Ben-Dor A, Shamir R, Yakhini Z (1999). Clustering gene expression patterns. J. Comp. Biol. 6:281-297.

Bielas JH, Loeb KR, Rubin BP, True LD, Loeb LA (2006). Human cancers express a mutator phenotype. Proc. Natl. Acad. Sci. U S A. 103:18238-18242.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406:536-540.

Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, Quackenbush J, Yeatman TJ (2004). Multi-platform, multi-site, microarray-based human tumor classification. Am. J. Pathol. 164:9-16.

Blume-Jensen P, Hunter T (2001). Oncogenic kinase signalling. Nature 411:355-365.

Bowtell DDL (1999). Options available - from start to finish - for obtaining expression data by microarray. Nature Genet. 21:25-32.

Briasoulis E, Pavlidis N (1997). Cancer Unknown Prim. Origin. Oncologist 2:142-152.

Brown PO, Botstein D (1999). Exploring the new world of the genome with DNA microarrays. Nature Genet. 21:33-37.

Bubendorf L KJ, Koivisto P (1999). Survey of gene amplifications during prostate cancer progression by high-throughout fluorescence *in situ* hybridization on tissue microarrays. Cancer Res. 59:803-806.

Bubendorf L, Kononen J, Koivisto P, Schraml P, Moch H, Gasser TC, Willi N, Mihatsch MJ, Sauter G, Kallioniemi OP (1999). Survey of gene amplifications during prostate cancer progression by high throughput fluorescence *in situ* hybridization on tissue microarrays. Cancer Res. 59:803-806.

Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, Dohner H, Pollack JR (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. New Engl. J. Med. 350:1605-1616.

Cahill DJ (2001). Protein and antibody arrays and their medical applications. J. Immunol. Methods 250:81-91.

Carter D, Douglass JF, Cornellison CD, Retter MW, Johnson JC, Bennington AA, Fleming TP, Reed SG, Houghton RL, Diamond DL, Vedvick TS (2002). Purification and characterization of the mammaglobin/lipophilin B complex, a promising diagnostic marker for breast cancer. Biochemistry 41:6714-6722.

Chung CH, Levy S, Chaurand P, Carbone DP (2007). Genomics and proteomics: emerging technologies in clinical cancer research. Crit. Rev. Oncol. Hematol. 61:1-25.

De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y (2002). Adaptive quality-based clustering of gene expression profiles. Bioinformatics 18:735-746.

Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM (2001). Delineation of prognostic biomarkers in prostate cancer. Nature 412:822-826.

Dopazo J, Carazo JM (1997). Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J. Mol. Evol. 44:226-233.

Duggan DJ, Bittner M, Chen YD, Meltzer P, Trent JM (1999). Expression profiling using cDNA microarrays. Nature Genet. 21:10-14.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95:14863-14868.

Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen YD, Simon R, Meltzer P, Trent JM, Boguski MS (1998). Data management and analysis for gene expression arrays. Nature Genet. 20:19-23.

Fadlelmola FM, Zhou M, de Leeuw RJ, Dosanjh NS, Harmer K, Huntsman D, Lam WL, Banerjee D (2008). Sub-megabase resolution tiling (SMRT) array-based comparative genomic hybridization profiling reveals novel gains and losses of chromosomal regions in Hodgkin Lymphoma and Anaplastic Large Cell Lymphoma cell lines. Mol. Cancer 7:2.

Fejzo MS, Slamon DJ (2001). Frozen tumor tissue microarray technology for analysis of tumor RNA, DNA, and proteins. Am. J. Pathol. 159:1645-1650.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64-71.

Ferran EA, Ferrara P (1991). Topological maps of protein sequences. Biol. Cybern. 65:451-458.

Ferran EA, Ferrara P (1992). Clustering Proteins into Families Using Artificial Neural Networks. Comp. Appli. Biosci. 8:39-44.

Ferran EA, Ferrara P, Pflugfelder B (1993). Protein classification using neural networks. Proc. Int .Conf. Intell. Syst. Mol. Biol. 1:127-135.

Ferran EA, Pflugfelder B (1993). A hybrid method to cluster protein sequences based on statistics and artificial neural networks. Comput. Appl. Biosci. 9:671-680.

Ferran EA, Pflugfelder B, Ferrara P (1994). Self-Organized Neural Maps of Human Protein Sequences. Protein Sci. 3:507-521.

Ferrando AA, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi, SC, Behm FG, Pui CH, Downing JR, Gilliland DG, Lander, ES, Golub TR, Look AT (2002). Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell 1:75-87.

Fritzke B (1994). Growing Cell Structures - a Self-Organizing Network for Unsupervised and Supervised Learning. Neural Netw. 7:1441-1460.

Ge H (2000). UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. Nucleic Acids Res. 28:e3.

Geysen HM MR, Barteling SJ (1984). Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. Proc. Natl. Acad. Sci. USA. 81:3998-4002.

Gillespie DS (1965). A quantitative assay for DNA-RNA hybrids with DNA immobilised on a membrane. J. Mol. Biol. 12:829-842.

Giltnane JM, Rimm DL (2004). Technology Insight: identification of biomarkers with tissue microarray technology. Nat. Clin. Pract. Oncol. 1:104-111.

Giordano TJ, Shedden KA, Schwartz DR, Kuick R, Taylor JMG, Lee N, Misek DE, Greenson JK, Kardia SLR, Beer DG, Rennert G, Cho KR, Gruber SB, Fearon ER, Hanash S (2001). Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. Am. J. Pathol. 159:1231-1238.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing, JR, Caligiuri MA, Bloomfield, C.D., Lander, ES (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

Grunstein MHDS (1975). Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc. Natl. Acad. Sci. USA 72:3961-3965.

Hanahan D, Weinberg RA (2000). The hallmarks of cancer. Cell 100:57-70.

Hartigan JA (1975). Clustering algorithms. John Wiley & Sons, Inc., New York.

Herrero J, Dopazo J (2002). Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. J. Proteome Res. 1:467-470.

Herrero J, Valencia A, Dopazo J (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 17:126-136.

Heyer LJ, Kruglyak S, Yooseph S (1999). Exploring expression data: Identification and analysis of coexpressed genes. Genome Res. 9:1106-1115.

Hillenkamp F, Karas M, Beavis RC, Chait BT (1991). Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Anal. Chem. 63:1193A-1203A.

Hofmann WK, de Vos S, Tsukasaki K, Wachsman W, Pinkus GS, Said, JW, Koeffler HP (2001). Altered apoptosis pathways in mantle cell lymphoma detected by oligonucleotide microarray. Blood 98:787-794.

Humphery-Smith I, Wischerhoff E, Hashimoto R (2002). Protein arrays for assessment of target selectivity. Drug Discov. World 4:17-27.

Hunter T (2000). Signaling-2000 and beyond. Cell 100:113-127.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929-934.

Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO (1999). The transcriptional program in the response of human fibroblasts to serum. Science 283:83-87.

Kafatos FC, Jones, CW, Efstratiadis A (1979). Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. Nucleic Acids Res. 24:1541-1552.

Kallioniemi OP, Wagner U, Kononen J, Sauter G (2001). Tissue microarray technology for high-throughput molecular profiling of cancer. Hum. Mol. Genet. 10:657-662.

Kane M (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. 28:4552-4557.

Karas M, Hillenkamp F (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem. 60:2299-2301.

Khan L, Luo F (2005). Hierarchical clustering for complex data. Int. J. Artif. Intell. Tools 14:791-809.

Khrapko KR, Lysov YP, Khorlyn AA, Shick VV, Florentiev VL, Mirzabekov AD (1989). An Oligonucleotide Hybridization Approach to DNA Sequencing. Febs Lett. 256:118-122.

Kohonen T (1990). The Self-Organizing Map. Proc. IEEE 78:1464-1480.

Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton

S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. Natl. Med. 4:844-847.

LaBaer J, Ramachandran N (2005). Protein microarrays as tools for functional proteomics. Curr. Opin. Chem. Biol. 9:14-19.

Lahm HW, Langen H (2000). Mass spectrometry: A tool for the identification of proteins separated by gels. Electrophoresis 21:2105-2114.

Lal SP, Christopherson RI, dos Remedios CG (2002). Antibody arrays: an embryonic but rapidly growing technology. Drug Discov. Today 7:S143-149.

Lander ES (1999). Array of hope. Nat. Genet. 21:3-4.

Legrain P, Jestin JL, Schachter V (2000). From the analysis of protein complexes to proteome-wide linkage maps. Curr. Opin. Biotechnol. 11:402-407.

Leung SY CX, Chu KM, Yuen ST, Mathy J, Ji J, Chan AS, Li R, Law S, Troyanskaya OG (2002). Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis. Proc. Natl. Acad. Sci. USA. 99:16203-16208.

Levine AJ (1993). The tumor suppressor genes. Annu. Rev. Biochem. 62:623-651.

Liotta L, Petricoin E (2000). Molecular profiling of human cancer. Natl. Rev. Genet. 1:48-56.

Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D, Munson PJ, Young L, Wulfkuhle J, Petricoin EF (2003). Protein microarrays: meeting analytical challenges for clinical applications. Cancer Cell 3:317-325.

Liotta LA, Kohn EC, Petricoin EF (2001). Clinical proteomics - Personalized molecular medicine. Jama J. Am. Med. Assoc. 286:2211-2214.

Lipshutz RJ FS, Gingeras TR and Lockhart DJ. (1999). Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays. Natl. Genet. 21:20-24.

Luo F, Khan L, Bastani F, Yen I L (2003). A Dynamical Growing Self-Organizing Tree (DGSOT). Technical Report. University of Texas, Dallas.

Luo F, Khan L, Bastani F, Yen IL, Zhou JZ (2004). A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles. Bioinformatics 20:2605-2617.

MacBeath G (2002). Protein microarrays and proteomics. Nat. Genet. 32 Suppl:526-532.

MacBeath G (2002). Recombinant antibody microarrays - a viable option? Natl. Genet. 32 Suppl:526-532.

MacBeath G, Schreiber SL (2000). Printing proteins as microarrays for high-throughput function determination. Science 289:1760-1763.

McAdams HH, Shapiro L (1995). Circuit simulation of genetic networks. Science 269:650-656.

Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP, Sauter G (1999). High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. Am. J. Pathol. 154:981-986.

Nocito A, Bubendorf L, Tinner EM, Suess K, Wagner U, Forster T, Kononen J, Fijan A, Bruderer J, Schmid U, Ackermann D, Maurer R, Alund G, Knonagel H, Rist M, Anabitarte M, Hering F, Hardmeier T, Schoenenberger AJ, Flury R, Jager P, Fehr JL, Schraml P, Moch H, Mihatsch MJ, Gasser T, Sauter G (2001). Microarrays of bladder cancer tissue are highly representative of proliferation index and histological grade. J. Pathol. 194:349-357.

Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K (1992). Large-Scale cDNA Sequencing for Analysis of Quantitative and Qualitative Aspects of Gene-Expression. Nat. Genet. 2:173-179.

Paweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW, Emmert-Buck MR, Roth MJ, Petricoin EF, Liotta LA (2001). Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene 20:1981-1989.

Pawson T (1995). Protein modules and signalling networks. Nature 373:573-580.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000). Molecular portraits of human breast tumours. Nature 406:747-752.

Petach H, Gold L (2002). Dimensionality is the issue: use of photoaptamers in protein microarrays. Curr. Opin. Biotechnol. 13:309-314.

Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA (2002). Clinical proteomics: Translating bench side promise into bedside reality. Nat. Rev. Drug Discov. 1:683-695.

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415:436-442.

Popper HH, Kothmaier H (2008). Proteomics-Tissue and Protein Microarrays and Antibody Array What Information Is Provided? Arch. Pathol. Lab. Med. 132:1570-1572.

Ramachandran N, Hainsworth E, Bhullar B, Eisenstein S, Rosen B, Lau AY, Walter JC, LaBaer J (2004). Self-assembling protein microarrays. Science 305:86-90.

Ramaswamy S TP, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA 98:15149-15154.

Ramaswamy S, Ross KN, Lander ES, Golub TR (2003). A molecular signature of metastasis in primary solid tumors. Nat. Genet. 33:49-54.

Relogio A SC, Richter A, Ansorge W, Valcarcel J (2002). Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Res. 30:e51.

Richter J, Wagner U, Kononen J, Fijan A, Bruderer J, Schmid U, Ackermann D, Maurer R, Alund G, Knonagel H, Rist M, Wilber K, Anabitarte R, Hering F, Hardmeier T, Schonenberger A, Flury R, Jager P, Fehr JL, Schraml P, Moch H, Mihatsch MJ, Gasser T, Kallioniemi OP, Sauter G (2000). High-throughput tissue microarray analysis of cyclin E gene amplification and overexpression in urinary bladder cancer. Am. J. Pathol. 157:787-794.

Ritossa F, Malva C, Boncinelli E, Graziani, F, Polito L (1971). The first steps of magnification of DNA complementary to ribosomal RNA in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA. 68:1580-1584.

Roepstorff P (1997). Mass spectrometry in protein studies from genome to function. Curr. Opin. Biotechnol. 8:6-13.

Rosty C (2002). Identification of hepatocarcinoma-intestinepancreas/pancreatitis- associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology. Cancer Res. 62:1868-1875.

Russo G, Zegar C, Giordano A (2003). Advantages and limitations of microarray technology in human cancer. Oncogene 22:6497-6507.

Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN (2000). A gene expression database for the molecular pharmacology of cancer. Nat. Genet. 24:236-244.

Schraml P, Kononen J, Bubendorf L, Moch H, Bissig H, Nocito A, Mihatsch MJ, Kallioniemi OP, Sauter G (1999). Tissue microarrays for gene amplification surveys in many different tumor types. Clin. Cancer Res. 5:1966-1975.

Schulze A, Downward J (2001). Navigating gene expression using microarrays - a technology review. Nat. Cell Biol. 3:E190-E195.

Schwartz SA, Weil RJ, Johnson MD, Toms SA, Caprioli RM (2004). Protein profiling in brain tumors using mass spectrometry: Feasibility of a new technique for the analysis of protein expression. Clin. Cancer Res. 10:981-987.

Schwikowski B, Uetz P, Fields S (2000). A network of protein-protein interactions in yeast. Nat. Biotechnol. 18:1257-1261.

Sheehan KM, Calvert VS, Kay EW, Lu YL, Fishman D, Espina V, Aquino J, Speer R, Araujo R, Mills GB, Liotta LA, Petricoin EF, Wulfkuhle JD (2005). Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. Mol. Cell. Proteomics 4:346-355.

Shimada H, Ichikawa H, Ohki M (2002). Potential involvement of the AML1-MTG8 fusion protein in the granulocytic maturation characteristic of the t(8;21) acute myelogenous leukemia revealed by microarray analysis. Leukemia 16:874-885.

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 8:68-74.

Simon R, Nocito A, Hubscher T, Bucher C, Torhorst J, Schraml P, Bubendorf L, Mihatsch MM, Moch H, Wilber K, Schotzau A, Kononen J, Sauter G (2001). Patterns of her-2/neu amplification and overexpression in primary and metastatic breast cancer. J Natl. Cancer Inst. 93:1141-1146.

Sneath PHAaS, R.R. (1973). Numerical Taxonomy. W. H. Freeman & Co., San Francisco, California.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale AL (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA. 98:10869-10874.

Sotiriou C, Powles TJ, Dowsett M, Jazaeri AA, Feldman AL, Assersohn L, Gadisetti C, Libutti SK, Liu ET (2002). Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer. Breast Cancer Res. 4:R3.

Southern EM (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503-517.

Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR (2001). Chemosensitivity prediction by transcriptional profiling. Proc. Natl. Acad. Sci. USA. 98:10787-10792.

Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Hampton GM (2001). Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res. 61:7388-7393.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky, E, Lander ES, Golub TR (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA. 96:2907-2912.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999). Systematic determination of genetic network architecture. Nat. Genet. 22:281-285.

Templin MF, Stoll D, Schrenk M, Traub PC, Vohringer CF, Joos TO (2002). Protein microarray technology. Trends Biotechnol. 20:160-166.

Templin MF, Stoll D, Schrenk M, Traub PC, Vohringer CF, Joos TO (2002). Protein microarray technology. Trends Biotechnol. 20:160-166.

Templin MF, Stoll D, Schwenk JM, Potz O, Kramer S, Joos TO (2003). Protein microarrays: promising tools for proteomic research. Proteomics. 3:2155-2166.

TJ Y (2003). The future of cancer management: translating the genome, transcriptome, and proteome. Ann. Surg. Oncol. 10:7-14.

Torhorst J, Bucher C, Kononen J, Haas P, Zuber M, Kochli OR, Mross F, Dieterich H, Moch H, Mihatsch M, Kallioniemi OP, Sauter G (2001). Tissue microarrays for rapid linking of molecular changes to clinical endpoints. Am. J. Pathol. 159:2249-2256.

Toronen P, Kolehmainen M, Wong C, Castren E (1999). Analysis of gene expression data using self-organizing maps. Febs Lett. 451:142-146.

Tothill RW, Kowalczyk A, Rischin D, Bousioutas A, Haviv I, van Laar RK, Waring PM, Zalcberg J, Ward R, Biankin AV, Sutherland RL, Henshall SM, Fong K, Pollack JR, Bowtell DDL, Holloway AJ (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res. 65:4031-4040.

Ullmann R, Morbini P, Halbwedl I, Bongiovanni M, Gogg-Kammerer M, Papotti M, Gabor S, Renner H, Popper HH (2004). Protein expression profiles in adenocarcinomas and squamous cell carcinomas of the lung generated using tissue microarrays. J. Pathol. 203:798-807.

Valk PJM, Verhaak RGW, Beijen MA, Erpelinck CAJ, van Doorn-Khosrovani SBV, Boer JM, Beverloo HB, Moorhouse MJ, van der Spek PJ, Lowenberg B, Delwel R (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. New Engl. J. Med. 350:1617-1628.

van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Mao, M., Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002). A gene-expression signature as a predictor of survival in breast cancer. New Engl. J. Med. 347:1999-2009.

van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995). Serial analysis of gene expression. Science 270: 484-487.

Vogelstein B KK (2004). Cancer genes and the pathways they control. Nat. Med. 10:789-799.

Vogelstein B, Kinzler KW (1993). The multistep nature of cancer. Trends Genet. 9:138-141.

Wang S, Saboorian MH, Frenkel EP, Haley BB, Siddiqui MT, Gokaslan S, Wians FH Jr., Hynan L, Ashfaq R (2001). Assessment of HER-2/neu status in breast cancer. Automated Cellular Imaging System (ACIS)-assisted quantitation of immunohistochemical assay achieves high accuracy in comparison with fluorescence in situ hybridization assay as the standard. Am. J. Clin. Pathol. 116:495-503.

Weinberg RA (1995). The molecular basis of oncogenes and tumor suppressor genes. Ann NY Acad. Sci. 758:331-338.

Wen XL, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998). Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. USA. 95:334-339.

Wilson DS, Nock S (2003). Recent developments in protein microarray technology. Angew Chem. Int. Ed Engl. 42:494-500.

Wu TD (2001). Analysing gene expression data from DNA microarrays to identify candidate genes. J. Pathol. 195:53-65.

Yates JR (1998). Mass spectrometry and the age of the proteome. J. Mass Spectrom. 33:1-19.

Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz, R., Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou XD, Li JY, Liu HQ, Pui CH, Evans WE, Naeve C, Wong LS, Downing JR (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell. 1:133-143.

Yeung KY, Haynor DR, Ruzzo WL (2001). Validating clustering for gene expression data. Bioinformatics 17:309-318.

Zajchowski DA, Bartholdi MF, Gong Y, Webster L, Liu HL, Munishkin A, Beauheim C, Harvey S, Ethier SP, Johnson PH (2001). Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells. Cancer Res. 61:5168-5178.

Zhu H, Snyder M (2003). Protein chip technology. Curr. Opin. Chem. Biol. 7:55-63.