# Functional Genomics

We are in the midst of a genomics revolution. The first chapter of this revolution will end later this year with the completion of goal of the goal of the Human Genome Project: the sequencing of the entire human genome, about 100,000 genes in total. This endeavour has been the catalyst for the genomics revolution and has moved science into uncharted territories, which has necessitated the need to establish both new disciplines and new vocabularies. So, we know have pharmacogenomics, genotyping, pharmacogenetics, micro-arrays, biochips, differential display, bioinformatics, cheiminformatics.

On March 1, 2000, the Society for Medicines Research held a one-day meeting on *Functional Genomics* at the National Heart and Lung Institute of Imperial College of Science, Technology and Medicine, London, UK. The delegates, from both industry and academia, where from a range of backgrounds and disciplines, including molecular biology, biochemistry, pharmacology, and chemistry; there were even a number of analysts in attendance. The meeting provided both a review of some of the exciting progress that genomics has made of late, particularly in regard to its application to the discovery of new drugs, and a glimpse of what subsequent chapters of the genomics story hold in store for us.

**Genomics**

The application of microarrays (or DNA chips) to the drug discovery process was reviewed by Derk J Bergsma (SmithKline Beecham Pharmaceuticals, King of Prussia, USA). Microarrays are small squares of thin glass (about 1 x 1 cm) overlaid by a grid of approximately 100 x 100 smaller squares. Into each of these small squares is inserted a set of probes capable of detecting the presence or absence on of our 100,000 different genes. The probes (cDNAs encoding full or partial gene fragments) are robotically spotted onto the slide short fragments of single-stranded DNA. This is incubated with fluorescently labeled cDNA made from RNA isolated from cells or tissues. A confocal scanner is used to detect labeled cDNA hybridized to complementary DNA sequences on the chip. Fluorescent signal intensity represents relative expression of the corresponding genes when compared to their expression pattern in a control sample. Currently each microarray can detect between 7,000 and 10,000 different genes; it therefore provides a high throughput technology to study RNA expression in cells and tissues (e.g. normal and diseased) of tens of thousands of genes simultaneously. As part of a team effort, scientists at SmithKline Beecham developed a robust cDNA-based microarray platform, and have validated the

technology as a powerful tool for surveying RNA expression in human, yeast and bacterial systems.

**Toxicogenomics**

The use of microarrays for toxicological assessment was described by Peter Bugelski (SmithKline Beecham Pharmaceuticals, Harlow, UK). Using markers relevant to toxicity (e.g. P450 induction) it was argued that a particular profile of expressed sequence patterns could be used to provide a prediction of a toxicological liability in the clinic.

**Pharmacogenomics**

David Bailey (De Novo Pharmaceuticals, Cambridge, UK) defined pharmacogenomics as the science that combines pharmacology and genomics; this contrasts with the common meaning of this word, i.e. the tailoring drug treatment to in light of genomic variation, which he suggested would be better served by the name pharmacogenetics. He showed how conventional medicinal chemistry could be complemented by computer-aided drug design to provide both a virtual screen and de novo structure generation for both drug design and lead optimisation. To illustrate this, it was shown how 6 families of GDP-binding proteins were used and structures compared on the basis of either similarity or dissimilarity.

**Proteomics**

The application of proteomics in the drug discovery process was reviewed by Martin Page (Oxford GlycoSciences, Oxford, UK). Genomics fails to provide any information on the subsequent proteins, their cellular localisation, post-translational modifications and protein-protein interactions. The relationship between mRNA levels and protein levels are also unclear, and genomic/array analyses cannot easily study body fluids (such as serum, cerebrospinal fluid, synovial fluid) where novel targets and markers may reside. To complement the genomic approaches, there have been tremendous advances recently in the high throughput systematic analysis of proteins, termed proteomics. This involves the separation of proteins using 1-D and 2-D technology, detecting the proteins with highly sensitive fluorescent dyes, and then scanning these into computers such that they become digitised images. Using landmarks, and advanced suites of software, these are subsequently processed into a common geometry from which one is able to accurately interrogate the data, and identify which proteins are either differentially or uniquely expressed between sample sets. The proteins of interest are then subjected to

trypsinolysis and tandem mass spectrometry, and the resulting spectra searched against databases in order to identify the proteins. Much of this of this process is now automated, enabling the processing of hundreds of 2-D gels and proteins per week. By the careful choice of sample types, and fractions thereof, it is possible to rapidly identify key disease-specific or regulated proteins, and to study biological processes; the output of which can have major impact across the entire drug discovery process.

In this presentation, the application of proteomics for new target and marker discovery will be illustrated using examples from clinically derived breast cancer tissue and sera studies. These studies are largely collaborative with the Ludwig Institute for Cancer Research, and the Cancer Research Campaign laboratories. Proteome analysis has been used to examine the differences at the protein level, between the two major cell types comprising the normal human breast. From the interrogation of over 43,000 proteins in this study, 170 proteins were identified whose expression differed significantly between the two purified cell populations. The derivation of an advanced composite proteome image of the breast luminal cells is particularly important, since the majority of breast tumours are known to arise from this cell type. In parallel studies, we have also derived proteome images from a series of purified cell preparations of primary and metastatic breast tumours. Comparisons of the proteomes from normal and tumour breast are underway, and disease-specific proteins have been found.

In addition to the cell-based studies above, proteome studies have also been applied to fractionated material, and there is an extensive database of human breast membrane proteins. Some of these proteins could be strong candidates for therapeutic approaches involving antibodies and vaccines. These are areas where significant clinical data is emerging with encouraging results.

Our proteome studies have also been used to examine the proteins present in the sera from patients with primary breast cancer, metastatic breast cancer, or normal human volunteers. Using a depletion process to remove the common abundant proteins in the sera samples, an initial pilot study involving sera from 17 patients in each of these groups was undertaken. Subsequent analysis of the differences between the groups identified clusters of proteins whose level of expression are altered. This pilot study has now been extended to encompass a longitudinal study involving a time course, in which sera is examined from node-negative breast cancer patients who

have either benefited from a period of extended remission, or who have relapsed early with recurrent advanced disease.

**Bioinformatics**

With the enormous amount of information now being generated from both genomic an proteomic approaches, it is essential that we have an adequate approach to both analyse and make sense of this data. Jean-Michel Claverie (CNRS-AVENTIS, Marseille, France) described use of non-sequence similarity-based bioinformatics approaches in functional genomics. There are three basic challenges for bioinformatics: (i) finding the genes; (ii) locating their coding regions; and (iii) predicting their functions. However, our capacity for interpreting vertebrate genomic and transcript (cDNA) sequences using experimental or computational means very much lags behind our raw sequencing power [1]. If the performances of current programs in identifying internal coding exons are good, the precise 5'-->3' delineation of transcription units (and promoters) still requires additional experiments. Similarly, functional predictions made with reference to previously characterized homologues are leaving >50% of human genes unannotated or classified in uninformative categories ('kinase', 'ATP-binding', etc.). I will discuss the implication of these limitations for the annotation of the human genome sequence [2].

To overcome the insufficiency of sequence similarity-based bioinformatics, radically new approaches are being developed to take advantage of large-scale gene expression as well as comparative genomics studies, capable of providing genome-scale experimental information at a pace consistent with the progress of sequencing. Academic and industrial researchers are increasingly relying on those methods to prioritize their studies and choose their targets. For instance, the study of expression patterns can provide some insight into the function, reveal regulatory pathways, indicate side effects of drugs or serve as a diagnostic tool [3].

**Functional genomics**

Extrapolating systemically from gene sequence to function is undoubtedly the major challenge facing industry and academia alike at the start of the millennium, especially as a result of completion of the human genome sequence. Igor B. Roninson (University of Illinois, Chicago, IL, USA) reviewed a general approach to the identification and functional analysis of genes that are involved in almost any cellular phenotype approach. The Genetic Suppressor Element (GSE)

methodology permits identification of target genes through expression selection. GSEs are short gene fragments that show biological activity when expressed in a cell. In most cases, GSEs act either by encoding antisense RNA that inhibits the expression of the gene from which it is derived or by being translated as dominant-negative peptide inhibitors of the corresponding protein. GSEs can be used as specific genetic inhibitors of gene function and as tools for identifying genes, inhibition of which has a desired cellular effect. GSEs are generated by random fragmentation of target DNA, which may comprise a single cDNA clone, a viral genome, a subtracted cDNA population, normalized cellular cDNA, or genomic DNA. The mixture of target DNA fragments is cloned into a suitable expression vector, which is then introduced into the target cell type, followed by expression selection of biologically active clones. GSE selection has been successfully used by different groups in bacterial, yeast and mammalian systems, using a variety of targets and selection strategies. The most efficient procedures for GSE selection in mammalian cells involve the use of retroviral expression vectors, which allow one to transduce up to $10^8$ or more cells with a random fragment library, prior to selection for the desired phenotype. The integrated retroviruses can be efficiently recovered from the selected cells by PCR and tested for functional activity. Since the same cell can be infected with different retroviruses, combinations of functionally cooperating GSEs can also be identified. GSE selection has benefited from the development of retroviral vectors for inducible gene expression. Such vectors allow one to select growth-inhibitory GSEs using, for example, the BuDR suicide selection technique. Other selection strategies that have been successfully utilized in mammalian cells are resistance to chemotherapeutic drugs, immortalization of primary fibroblasts, neoplastic transformation of 3T3 cells and flow cytometric selection for altered antigen expression. Some published examples of GSE selection in mammalian cells include the development of efficient inhibitors that interfere with different biological functions of p53, isolation of GSEs that protect $CD4^+$ human cells from HIV infection, identification of kinesin as a determinant of drug sensitivity in human cells, and cloning of the ING1 gene that cooperates with p53 in its tumor suppressor functions. The GSE approach continues to be utilized for the identification of potential therapeutic targets in cancer and other human diseases.

D. Xu (University of Glasgow, Glasgow UK) provided an excellent illustration of functional genomics. Dr Xu described the identification and functional studies of differentially expressed genes between Th1 and Th2 cells. The Th1 and Th2 T cell subsets have distinguished functions and play a central role in immunity. The development and balance of Th1 and Th2 cells can determine the outcome of many diseases. However, the mechanisms which control the differential

function and the balance between Th1 and Th2 cell in these diseases are still unknown. To identify the selectively expressed genes which control the induction and function of Th1 and Th2 responses is a key immunological objective with implications in many diverse areas such as vaccination and immunotherapy of infectious and autoimmune diseases. By using differential display PCR, a collection of Th1 and Th2 specific genes has been identified. One orphan receptor molecule, ST2L, a member of the IL-1 R family, was stably expressed on murine Th2 but not Th1 cells. An antibody against ST2L strongly labelled Th2 but not Th1 cells. Three-colour single cell flow cytometric analysis showed that cell surface ST2L was coexpressed with intracellular IL-4, but not with IFN-gamma. The antibody selectively lysed Th2 cells in vitro in a complement-dependent manner. In vivo, it enhanced Th1 responses by increasing IFN-gamma production and decreasing IL-4 and IL-5 synthesis. It induced resistance to Leishmania major infection in BALB/c mice and exacerbated collagen-induced arthritis in DBA/1 mice. We also found another member of the IL-1R family, IL-18R, was selectively expressed on Th1 but not on Th2 cells. An anti-IL-18R antibody inhibited IL-18- induced IFN-gamma production by Th1 clones in vitro. In vivo, anti-IL-18R antibody reduced local inflammation and lipopolysaccharide-induced mortality in mice. This was accompanied by shifting the balance from Th1 to Th2 responses, manifest as decreased IFN-gamma and proinflammatory cytokine production and increased IL-4 and IL-5 synthesis. Therefore, ST2L and IL-18R are the stable markers for Th2 and Th1 cells respectively and regulate Th1 and Th2 functions. These molecules may be targets for therapeutic intervention.

The use of model organisms for target validation in the pharmaceutical industry was considered Dr Sohaila Rastan (SmitheKline Beecham, Harlow, UK). Why should a unicellular organism such as yeast be useful when studying human genes? What relevance has a gene found in the nematode worm to discovering new molecular targets for the pharmaceutical industry? Is a fly with a defective gene and an abnormal circadian rhythm *really* going to tell us about sleep patterns in man? The answer to all these questions is predicated on the premise that biological function is conserved during evolution as are the functionally significant regions of the genome. We can experiment with yeast, fruit flies, nematode worms, and mice in a way that is not possible in humans and thus determine the molecular mechanism underlying the gene defect. Cross-species comparison and functional genomics in model organisms is thus the most useful approach to identify genes and study their function in normal and pathological conditions.