New perspectives in lead generation I: Discovery of biological targets

Michael J. Ashton, Michael Jaye, and Jon S. Mason

Future success for most pharmaceutical companies will depend on innovation leading to the discovery of new drugs that provide clear medical benefit to patients when compared to current therapies. There are three key steps in the discovery process; discovery of *relevant* biological targets, generation of 'lead' compounds and the optimization of leads to give potent, efficacious and safe drugs. There are, increasingly, many 'biological' approaches to treatment, such as gene therapy or antisense therapy. This article is focused solely on key aspects of lead generation for compounds of low molecular weight[1]. Part 1 focuses on aspects of biological target identification. Part 2, to be published in the February issue of *Drug Discovery Today*, will address lead generation, with special emphasis on the measurement of diversity within and between compound libraries.

The initial step of uncovering a relevant new biological target is key to the whole drug discovery process. It is important that the target is linked directly to the disease under consideration, and is not solely a new protein/receptor found in, for example, the central nervous system or the cardiovascular system, with speculation as to its role in the pathology of disease. There has been much research in the past of this type, with some success, but many notable failures, and it can lead to 'a compound in search of a disease'. The consequence can be an exceedingly expensive and difficult clinical development program.

Traditionally, the identification of biological targets has involved a reductionist approach in which the pathological phenomenon is examined with increasing resolution, much like microscopy with increasingly powerful objective lenses. This process typically originates from an understanding of fundamental biological mechanisms in man or an animal model, followed by studies involving intact tissues or cells or preparations thereof, ultimately revealing molecular targets for therapy. This blueprint for the discovery of biological targets involves the combined expertise of numerous subdisciplines within biology, such as whole-animal physiology, biochemistry, enzymology and molecular biology. The key contributions of these disciplines are made more often in series than in parallel. Thus, identification of biological targets by this traditional route involves a considerable investment of time as well as human and financial resources.

The revolution in molecular biology within the past 20 years, in particular, the ability to identify disease-causing genes by such techniques as positional cloning, has generated new strategies for the identification of targets for therapeutic intervention. This powerful approach relies on the discovery of a linkage, or association, between a phenotype and a DNA marker in affected families or animals. Analysis with additional neighboring DNA markers gradually defines the smallest genomic DNA segment harboring the disease gene, which is ultimately identified by DNA sequencing of the

Michael J. Ashton*, **Michael Jaye**, and **Jon S. Mason**, Rhône-Poulenc Rorer, Pharmaceutical Research, 500 Arcola Road, Collegeville, PA 19426-0107, USA. *tel: +1 610 454 8516, fax: +1 610 454 3340, e-mail: jon.mason@rp.fr

candidate gene(s). Examples of the successful use of this technique include the identification of the genetic defects underlying human monogenic disorders, such as cystic fibrosis[2,3] and Huntington's disease[4], and the recent identification of the murine gene responsible for obesity[5]. These techniques should facilitate the more daunting challenge of discovery of genes underlying more complex polygenic disorders, such as atherosclerosis and hypertension.

The ultimate objective of the international Human Genome Project is to determine the DNA sequence of the approximately 3×10^9 base pair haploid human genome. Great effort is being expended to localize DNA markers, or 'tags' at regular intervals in the human genome, creating a physical map of the genome. Whether these tags correspond to one of the estimated 10^5 different genes that comprise the genome, or to intergenic DNA, is irrelevant, because the tag's purpose is to provide a signature for a particular segment of the genome.

Expressed sequence tags

Expressed sequence tags (ESTs)[6] represent a special class of DNA marker. Gene expression in eukaryotes involves the initial transcription of a gene by RNA polymerase II into an unprocessed mRNA precursor containing both exons (protein encoding sequences) and introns (intervening sequences). After nuclear excision of the introns which juxtaposes the exons, or splices them together, the mature protein encoding mRNA is exported from the nucleus to the cytoplasm, where it is translated into protein. In the molecular biology laboratory, DNA copies of a mixed mRNA population from cells and tissues can be generated using a series of enzymatic steps, inserted into a variety of cloning vehicles, and propagated (for example in the bacterium *Escherichia coli*). Such copy DNA (cDNA) libraries (often consisting of more than 10^6 clones) represent the entire repertoire of genes expressed in the tissue at the time of isolation (Figure 1). The nucleotide (base)

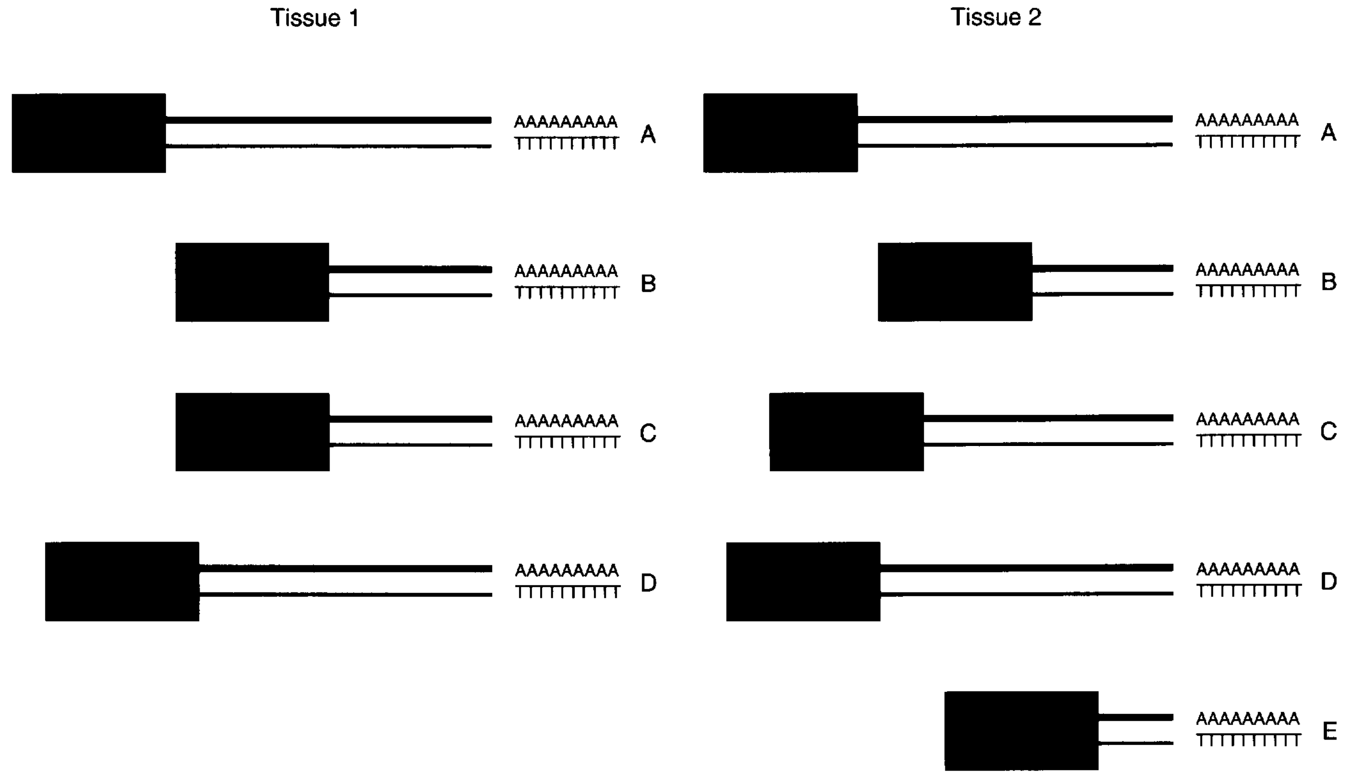

Figure 1. *Expressed sequence tagging strategy. For each sample, mRNA populations (thick lines) are copied into cDNA (thin lines) using a single downstream oligo d(T) primer. Standard techniques are then used to convert the single-stranded cDNA into a cDNA library of clones representative of the original mRNA population. The cDNA inserts of several thousand individual clones from each library are partially sequenced from one or both extremities (regions of sequence from the 5′ extremities of cloned cDNAs are indicated by the boxes), each sequence representing an expressed sequence tag (EST). In the example shown here, comparison of ESTs from tissues 1 and 2 reveals an EST (E in the figure) that is unique to tissue 2.*

sequence can be determined for short regions of individual cDNAs and this provides a unique identifier or EST for that cDNA. Since cDNAs are derived from expressed genes, each EST represents a specific gene.

At institutions such as Human Genome Sciences, Inc. (HGS), The Institute for Genomic Research (TIGR), Incyte, and the Washington University/Merck project, such efforts are coupled to powerful computational capacity for storage and analysis of sequence data. This strategy is implemented using robotic high-throughput technologies, so that nucleotide sequences of thousands of clones can be determined within days. Recently, a compilation and analysis of 83×10^6 nucleotides of EST sequence, representing over 266 000 ESTs, was published by HGS and TIGR[7]. This approach provides a mountain of information and an arsenal of tools toward the fulfilment of the aims of the international Human Genome Project, and, from the point of view of drug discovery, potential biological targets for unmet medical needs are inevitably revealed. For example, a novel DNA sequence showing convincing homology to an identified superfamily of genes such as G protein-linked, 7 membrane-spanning receptors, protein kinases and cytokines may be discovered.

The powerful informatic systems for storage and analysis of the EST sequences also permit the pattern of expression of any particular EST to be determined[7] – a virtual electronic northern blot. Thus, gene expression of a particular DNA sequence could be examined in a range of tissues and cell types, or within a single cell type, such as endothelial cells or lymphocytes, following biological or pharmacological perturbation. Such analysis may illuminate genes whose expression is restricted to specific tissues or cells and/or those that are controlled by biological or pharmacological agents, such as genes turned on or off in response to T-cell activation, or growth factor or cytokine addition or withdrawal.

The main disadvantage of the EST strategy for drug discovery is the requirement for specialized expertise and a substantial investment of resources. Furthermore, in contrast to positional cloning, which identifies genes underlying phenotypes, an EST by itself frequently cannot be ascribed a precise function. This is because the process of ascribing a function to the protein encoded by the EST relies on comparison of its nucleotide sequence to sequences within a databank of known sequences. Therefore, sequences with no homology to those in the databank cannot be assigned even a speculative function. In contrast, for sequences that can be assigned to a superfamily of genes, such as tyrosine kinases, a tentative, but imprecise function can be assigned. Considerable additional work is required to ascertain the function of the encoded protein; an even more daunting task is the determination of its role in the biology of the whole organism. The difficulty of ascertaining the function of an EST is also an obstacle in alternative gene discovery techniques, such as the technique of differential display, which is described below.

For drug discovery, the EST strategy provides both the revelation of new gene sequences and the ability to perform electronic northern blots to study the pattern of expression of the gene of interest. Alternative techniques, such as subtractive hybridization and differential display, can also provide such information. The main difference between the EST strategy and the techniques of subtractive hybridization and differential display is that the latter techniques focus exclusively on genes that are differentially expressed. Thus, the EST strategy reveals differentially expressed genes as a consequence of cataloguing gene expression, whereas, by design, subtractive hybridization, differential display and related techniques illuminate only differentially expressed genes. It is important to add that physical mapping of ESTs will greatly facilitate efforts to identify disease-associated genes by positional cloning. This is because once a disease is mapped to a particular genomic region by positional cloning, the disease-associated gene will be more readily recognized from the physical map of ESTs with known or suspected biological functions.

Differential display

The technique of differential display[8] and a related technique known as RNA fingerprinting[9,10] use short oligonucleotides of arbitrary sequence to prime polymerase chain reactions (PCR) on cDNA generated from mRNA samples. The power of these techniques is that they permit a simultaneous comparison of differentially expressed genes (either upregulated or downregulated) within different mRNA samples. First described in 1992, differential display and RNA fingerprinting are evolving techniques that can be practised using different methodologies. The most significant variation between the strategies used is in the sequence and length of the primers[8–11].

Figure 2 shows an example of the differential display technique in which mRNA subpopulations to be compared are first converted into DNA subpopulations, which then serve as templates for PCR using a panel of oligonucleotide primers of arbitrary sequence. In this approach, each of nine different downstream oligonucleotides is used separately to prime DNA synthesis on the two mRNA populations to be

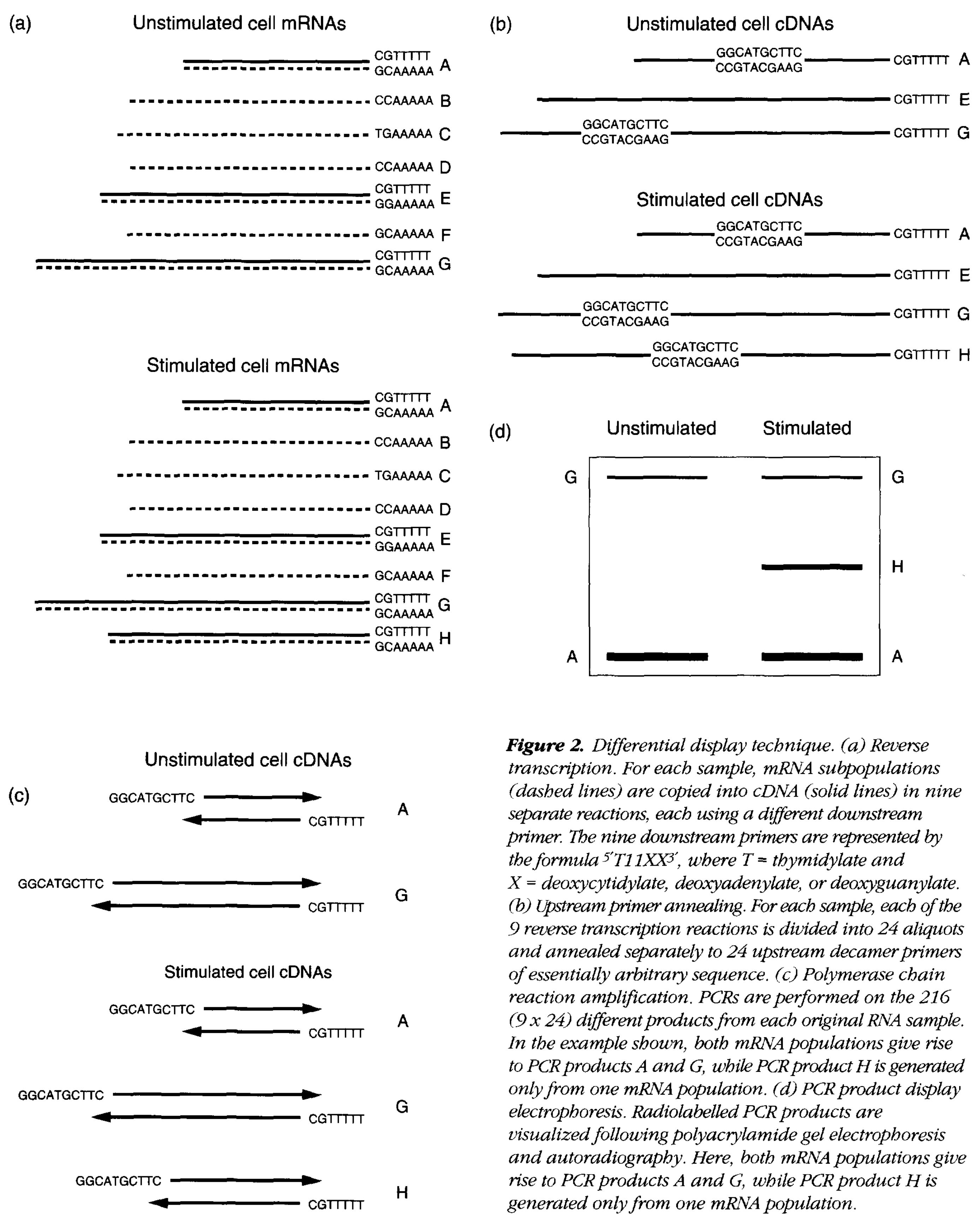

Figure 2. *Differential display technique. (a) Reverse transcription. For each sample, mRNA subpopulations (dashed lines) are copied into cDNA (solid lines) in nine separate reactions, each using a different downstream primer. The nine downstream primers are represented by the formula* $^{5'}T11XX^{3'}$*, where T = thymidylate and X = deoxycytidylate, deoxyadenylate, or deoxyguanylate. (b) Upstream primer annealing. For each sample, each of the 9 reverse transcription reactions is divided into 24 aliquots and annealed separately to 24 upstream decamer primers of essentially arbitrary sequence. (c) Polymerase chain reaction amplification. PCRs are performed on the 216 (9 x 24) different products from each original RNA sample. In the example shown, both mRNA populations give rise to PCR products A and G, while PCR product H is generated only from one mRNA population. (d) PCR product display electrophoresis. Radiolabelled PCR products are visualized following polyacrylamide gel electrophoresis and autoradiography. Here, both mRNA populations give rise to PCR products A and G, while PCR product H is generated only from one mRNA population.*

compared (Figure 2a). In the second step, each of the nine reverse transcription reactions are divided into 24 aliquots and annealed separately to 24 different upstream decamer primers of essentially arbitrary sequence (b). After the annealing step, which is based on fortuitous homologies between individual upstream primers and cDNAs, PCR is performed on the 216 (9 × 24) different combinations of primers for each of the mRNA populations to be compared (c). Thus, a complete analysis of two mRNA populations, such as control endothelial cells or lymphocytes against biologically or pharmacologically perturbed cells would require 432 PCR. Since PCR is performed in the presence of $[\alpha\text{-}^{32}\text{P}]$dATP, the PCR products are visualized following polyacrylamide gel electrophoresis and autoradiography (d). Most of the PCR products generated with any pair of upstream and downstream primers will be identical between the two mRNA samples being compared (exemplified by PCR products G and A) while PCR products that appear unique to one sample (or whose abundance appears significantly different between samples) represent potentially differentially regulated genes. At this point, the PCR products representing potentially differentially expressed genes are more extensively characterized; for each, this involves reproduction of original results with the relevant pair of upstream and downstream primers, followed by confirmation of differential expression by northern (RNA) blotting, and finally, determination of the nucleotide sequence of the differentially expressed PCR product. Comparison of the nucleotide sequence with those in electronic databases may reveal the identity or homology of the differentially expressed gene to published sequences. However, as in the EST strategy, considerable work is required to determine the function of the encoded protein.

Although differential display would theoretically appear to offer a short cut, requiring little starting mRNA for the identification of differentially expressed genes, the technique consumes significantly larger amounts of mRNA in confirmatory steps, such as northern blotting. It is often difficult to confirm differential expression, and when successful, the extent of differential expression as revealed by northern blotting is frequently less dramatic than that suggested from the intensities of the original PCR products. Furthermore, because differential display relies on fortuitous homologies between individual upstream primers and cDNAs, the technique fundamentally samples gene expression but is not exhaustive. Thus, some genes may be missed. Finally, as with any technique focused on revealing genes which are differentially expressed, it must also be demonstrated whether differential gene expression is a cause or consequence of the original biological or pharmacological differences or perturbations.

In spite of the limitations, there have been numerous reports of the successful use of the differential display technique[12–15], as well as continuing evolution in the original technique and supporting methodologies that have increased its speed and reproducibility, and thus its application to the elucidation of biological targets. Thus, by the use of positional cloning, ESTs and differential display techniques, a vast array of new biological targets relevant to disease states will emerge in the near future, giving much opportunity for the discovery of new drugs that modulate the course of a disease rather than ameliorate the symptoms.

ACKNOWLEDGEMENTS

We thank R.A. Lewis, K. Lynch, I.M. McLay and S.D. Pickett for their valued contributions to this article.

REFERENCES

1 Ashton, M.J. First Presented at S.C.I. Conference on Lead Generation, Robinson College, Cambridge, UK, June 1995
2 Rommens, J.M. *et al.* (1989) *Science* 245, 1059–1065
3 Riordan, J.R. *et al.* (1989) *Science* 245, 1066–1073
4 The Huntington's Disease Collaborative Research Group (1993) *Cell* 72, 971–983
5 Zhang, Y. *et al.* (1994) *Nature* 372, 425–432
6 Adams, M.D. *et al.* (1991) *Science* 252, 1651–1656
7 Adams, M.D. *et al.* (1995) *Nature* 377, (6547S suppl.) 3–174
8 Liang, P. and Pardee, A.B. (1992) *Science* 257, 967–971
9 Welsh, J. *et al.* (1992) *Nucleic Acids Res.* 20, 4965–4970
10 McClelland, M., Mathieu-Daude, F. and Welsh, J. (1995) *Trends Genet.* 11, 242–246
11 Zhao, S., Ooi, S.L. and Pardee, A.B. (1995) *Biotechniques* 18, 842–850
12 Sager, R. *et al.* (1993) *FASEB J.* 7, 964–970
13 Autieri, M.V. *et al.* (1995) *Lab. Invest.* 72, 656–661
14 Kozian, D.H. and Augustin, H.G. (1995) *Biochem. Biophys. Res. Commun.* 209, 1068–1075
15 Utans, U. *et al.* (1995) *Proc. Natl Acad. Sci. USA* 91, 6463–6467