Obtaining value from the human genome: a challenge for the pharmaceutical industry

Paul Spence

In a relatively short period of time, the sequence of the entire human genome and the identity of the genes contained within it will be determined. Among this vast array of genes are the pharmaceutical targets of the future. To benefit fully from this explosion of data, the pharmaceutical industry must establish high-throughput target identification and validation processes to feed their development pipelines.

In the past few years, the pharmaceutical industry has been confronted by a series of challenges arising from the ever-increasing advances in technology. In a competitive industry, any technology that allows higher productivity needs to be adopted and effectively integrated into the production process as rapidly as possible. Within the pharmaceutical industry, this has meant the integration of high-throughput screening and combinatorial chemistry into the discovery process, together with the ever-increasing use of complex databases to organize and store the vast quantities of data such technologies create. The Human Genome Project, initiated largely as an academic exercise less than a decade ago, is adding another dimension to the equation. In order to utilize this new and immensely valuable database fully (it will soon contain the DNA and protein sequences of every drug target the industry will ever work on), the pharmaceutical industry must establish efficient processes to determine which genes or gene products have real therapeutic value.

Although the Human Genome Project is officially only six years old, we are already being informed in editorials and conference announcements that we are entering the 'postgenomics' era. For the pharmaceutical industry this means that attention is turning towards the biological characterization of the thousands of new genes that are catalogued in public and private databases. Furthermore, in most cases, the genes or gene products are being evaluated as the targets for small-molecule therapeutics. This review describes how the tools of bioinformatics, genetics, molecular biology and pharmacology may be applied to triage the data coming from the Human Genome Project and other genomic programs efficiently, and thus identify the most likely candidate genes for pharmaceutical development. An attempt has been made to describe the process as generically as possible, therefore, some of the subtleties of analysis that might be applied in a specific disease area have been ignored. However, an overview of both the process and the technologies necessary to 'cherry pick' the most valuable therapeutic targets in a timely manner is provided.

Analysis of the human genome and gene expression

The human genome comprises about 3×10^9 bp of DNA of which ~5% codes for genes. Various estimates have been given regarding the number of genes encoded by the human genome, but it is generally thought that there are about 50,000–100,000 (Refs 1,2). It is anticipated that the entire human genome will be sequenced by the year 2005 (Ref. 3). In parallel with these efforts, several academic and commercial centres are carrying out high-throughput

Paul Spence, CNS Disorders Division, Wyeth-Ayerst Research, CN8000 Princeton, NJ 08543-8000, USA. tel: +1 732 274 4027, fax: +1 732 274 4020, e-mail: spencep@war.wyeth.com

  1359-6446/98/$19.00. PII: S1359-6446(97)01162-8

Box 1. Transcriptional profiling and database mining

Two methods are currently in widespread use. The first involves the arraying of cDNAs, normally onto glass slides, followed by hybridization of fluorescently labelled probes derived from reverse-transcribed mRNA from the tissue of interest (Figure 2). Because of the small format and high density of cDNAs on the solid phase, it is claimed that such arrays can give both sensitive and quantitative data on changes in mRNA expression levels for many genes simultaneously[37,38]. Such arrays can be used, for example, to compare the gene expression profiles of normal and diseased tissues.

Differential display techniques do not rely on the availability of cDNAs but use quantitative polymerase chain reaction (PCR) amplification and primer combinations that can effectively amplify all sequences within a population of mRNAs (Ref. 39). These methods can be used to clone novel genes by virtue of their differential expression patterns in different tissues, or in the same tissue under different circumstances.

As the technologies for transcriptional profiling become more sophisticated – allowing highly quantitative analysis, accommodating a higher sample throughput and more-complete coverage of the entire expression repertoire of cells and tissues – it will soon be possible to generate complete transcriptional fingerprints[40,41]. Such information might enable the analysis of signal transduction pathways that are modulated in diseases such as cancer[42], as well as the 'whole-genome' activity of pharmaceuticals. It is likely that transcriptional profiling will be used less to isolate individual genes through their change in expression, but more to identify patterns of transcriptional modulation leading to a more complete understanding of the biology of disease processes.

Database mining is required at all stages of the gene discovery and validation process. It can be used for novel gene discovery, by homology searching of databases for sequences representing known gene sequences or subsequences[43]. Although often seen as a 'fast track' way of identifying new genes, care needs to be taken in designing search strategies and interpreting the results[44]. The reasonable assumption is that genes showing sequence homology are likely to show functional homology[45]. In addition, homology searching is capable of detecting relationships among genes that are separated by billions of years of evolution. This is well illustrated by the use of bioinformatics to characterize the function of the product of the ataxia telangiectasia gene in silico by its homology to the yeast phosphatidylinositol 3-kinase, in addition to FK506- and rapamycin-binding proteins[46,47]. Although not an experimental method for focusing gene discovery, database mining does allow the luxury of choosing the 'type' of gene one would like to study. Many valuable pharmaceutical targets such as the G protein-coupled receptors are present in the human genome as large superfamilies, and a great deal is known about the sequence motifs present in these proteins such that focused database mining can be used to identify novel members[48].

sequencing of expressed genes and generating databases of expressed sequence tags (ESTs). These databases are of particular value to pharmaceutical companies because they provide sequence data that are naturally filtered, albeit coarsely, for pharmaceutical evaluation (the databases only contain sequences from exons of expressed genes). They also allow comparisons of genes between species and the discovery of new gene families involved in human disease[4]. Although valuable (~50% of human genes have been sampled as ESTs; Ref. 5), EST databases are presently poorly annotated and largely provide a starting point for further work.

The haploid human genome is divided among 23 chromosomes, each gene having a specific locus on its specific chromosome. A combination of the technologies and information that have been developed for the physical mapping of genes and family studies of inheritance has led to the identification of several disease genes over the past few years; for example, some genes have now been associated with Alzheimer's disease[6].

Recently, the work carried out by the groups mapping genes and identifying genes by EST sequencing was combined so that a gene map of increasing resolution could be constructed before the sequence of the entire human genome became available[3]. Such a map will greatly facilitate the identification of disease genes by the positional candidate approach, where large genomic regions identified in complex-trait studies can be rapidly analyzed based on their known gene content[7].

In addition to the above efforts, many pharmaceutical and biotechnology companies are generating large databases of ESTs focused around specific tissue types, diseased and normal tissues and tissues from various stages of embryonic development. However, as the genomic databases are increasing in size at a near exponential rate, the pharmaceutical industry is increasingly turning its attention to the elucidation of gene function, with the concern that the functional evaluation of genes is not yet a high-throughput science.

Focusing the search for disease genes

With the requirement to select the best targets from a pool of 100,000 candidates, it makes sense to reduce this

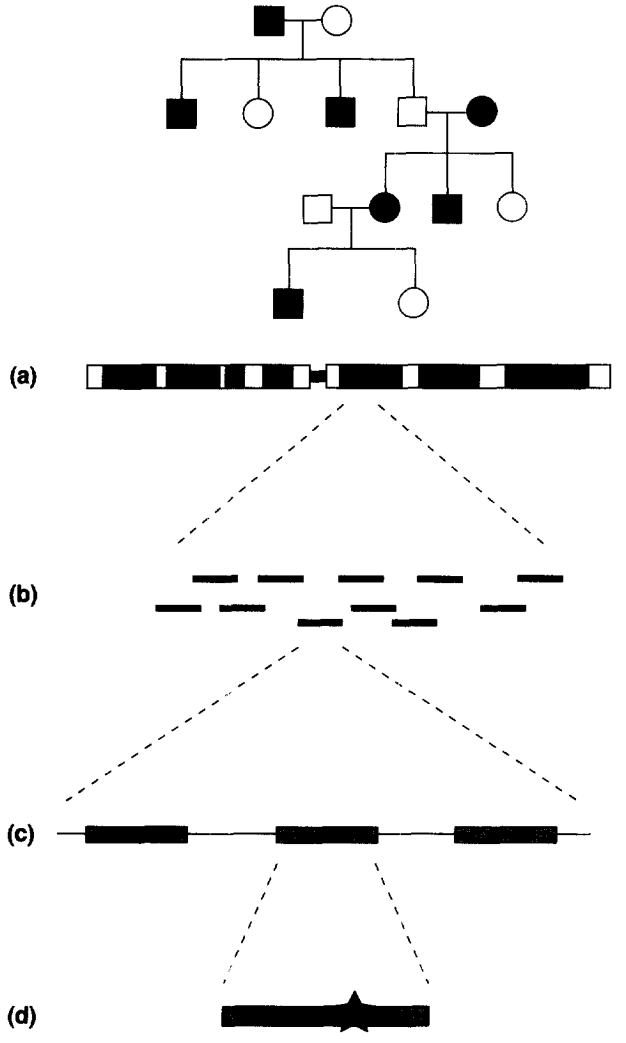

Figure 1. *Positional cloning. Families displaying a clear disease phenotype are collected for genetic mapping studies. From these studies, the disease loci are positioned to chromosomal regions using genetic markers (a). Clones containing large inserts (such as yeast artificial chromosomes, YACS) are identified by physical mapping (b). Candidate genes are identified through the application of transcript mapping and large-scale sequencing (c). The disease gene is identified through mutation analysis, comparing affected and unaffected populations (d).*

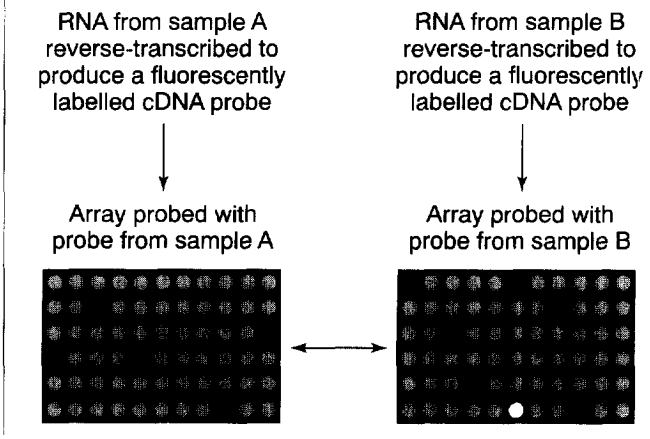

Figure 2. *Transcriptional profiling and database mining. Oligonucleotides or cDNAs are arrayed on a glass slide (the example here shows 66 separate arrays but many thousands are possible). The mRNA from the samples to be compared are fluorescently labelled during reverse transcription. The fluorescently labelled cDNA probes are then hybridized to the arrays and differentially expressed mRNAs are identified by the difference of fluorescence in equivalent sequences on the array. Grey spots represent genes expressed equally, while black and white spots represent genes that have increased and decreased expression levels, respectively. Alternative methods use different fluorescent labels for each sample and a single array for the hybridization of both samples. Differential expression is then detected ratiometrically.*

candidate pool size as soon as possible. The pharmaceutical industry, as well as many academic groups, has gone about this in several ways (see Box 1 and Figures 1–3). As will become evident, each approach has its advantages and disadvantages, which will require different emphases to be placed on the evaluation process (Box 2).

Clearly, one of the most focused ways to look for disease genes is to use the tools of population genetics. The power of such an approach is that it can unambiguously identify the genes responsible for a particular disorder. However, the more common inherited disorders are the most difficult to study using human genetics, because combinations of different genes are often involved. Additional complications can be introduced by the involvement of environmental triggers. Genetic studies of complex disorders usually require large study populations and can often lead to the identification of several disease-associated loci, all of which might require a more-detailed analysis[8,9]. Even if one can identify the gene or genes associated with the disorder, such studies provide no information about the biochemical role of the gene products[10]. However, the genetic data may reinforce the knowledge already available, as has been the case for Alzheimer's disease[6].

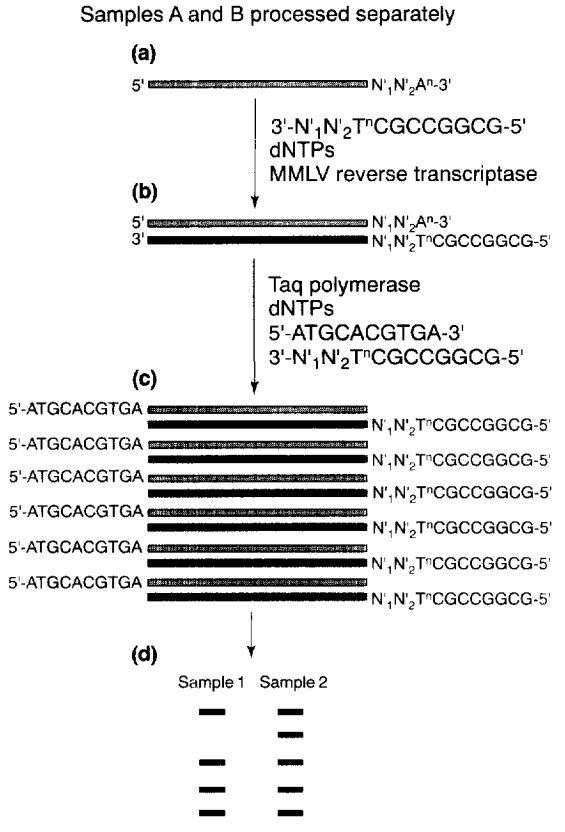

Figure 3. *Samples A and B are mRNAs from tissues or cell lines (e.g. normal and diseased or drug-treated and control) (a). The samples to be compared are first divided into 12 pools and reverse-transcribed with one of a set of anchored 3′ primers (N_1 is A, T, G or C; N_2 is A, G or C), containing poly(T) at their 3′ end, to produce copy DNA or cDNA (b). The products are then divided into a number of pools (depending on the extent of analysis planned) and amplified by polmerase chain reaction (PCR) with a selection of 5′ primers and the appropriate 3′ primer used in the cDNA reaction. The 5′ primers normally comprise ten or more bases of arbitrary sequence (represented in this example by the sequence ATGCACGTGA) (c). During the PCR, the samples are radiolabelled with [^{35}S]dATP or [^{33}P]dATP to allow autoradiographic detection after polyacrylamide gel electrophoresis. Differentially expressed mRNAs are identified by the presence or absence or change in signal on the autoradiograph (d). More recent adaptations of this technology make use of fluorescently labelled DNA and laser detection, which is much more quantitative than autoradiography. In addition, several methods have been developed that allow identification of the individual bands on the polyacrylamide gels by their size and end sequence properties*[40].

Human population studies are difficult to control, unlike mice colonies, which are much more amenable. Recently, several groups have used the correlation of mouse pathology to human disease to identify putative human disease genes[11–14]. Although genetic studies on mice are more efficient to carry out, one still needs to demonstrate the role of the human homologue (or orthologue) in the human disorder. Indeed, it is important to remember that the human homologue of the mouse mutant might not be the mutated gene in the human condition. Nevertheless, identification of the mouse mutation may point the way to the biochemical pathway involved in the disorder.

Transcriptional profiling, where the differential expression of genes is analyzed from various tissues sources (*in vitro* / *in vivo* disease models or biopsy samples), can supply a wealth of information regarding the biochemical processes being modulated (see Box 1 and Figures 2 and 3). Often these studies generate a large number of differentially expressed genes, challenging the researcher to make selections that differentiate cause from effect. Transcriptional profiling can be used not only to identify novel targets but also to check the validity of a target.

Database mining for novel genes from well-characterized structural classes, described by some as experiments 'in silico', has the advantage of allowing the researcher to choose the nature of the target being searched for (e.g. proteases or G protein-coupled receptors). The disadvantage is that the 'novel' gene is often identified with no associated data on disease relevance. Some focus can be introduced by selecting the database to be mined; for example, a database of brain ESTs or, better still, ESTs from specific brain regions might be mined for homologues of genes involved in CNS disorders.

Gene identification using technologies based on function include the yeast two-hybrid system, which has been used with great success to clone new genes by virtue of the association of their protein product with previously characterized proteins[15]. It provides a valuable tool for identifying new genes lying on defined biochemical pathways and for biochemically 'map-referencing' genes identified by other means. Other methodologies for gene identification, such as expression cloning and cloning by complementation, are also regularly used. These approaches, like the yeast two-hybrid system, benefit from the identification of genes within their biochemical context (e.g. as the receptor for an orphan ligand or by substituting for another gene product of known function)[16–18].

Candidate genes will arise from all of the above approaches, often in large numbers. In the next section, the

Box 2. Methods of gene identification: advantages and disadvantages

Human genetic studies
Genes identified by this approach will play a role in the etiology of the targeted disease. However, if novel, no information about the biochemical role will be available (unless inferred by homology to known genes).

Transcriptional profiling
The paradigm from which transcriptionally modulated genes are identified will play a significant role in their value as candidate disease genes. Well-understood *in vivo* disease models or *in vitro* systems, thought to be representative of the disease process, would be the most useful sources. Additionally, the use of known disease/pathology modifiers (pharmaceuticals or proteins) in such models might well expose genes involved in therapeutic responses. However, the role, if any, that such modulated genes may play in the etiology or pathology of the disease will still need to be determined.

Database mining/homology cloning
Unless identified from very defined databases or libraries (e.g. regiospecific EST libraries and disease/normal subtracted libraries), no information regarding the role of these genes in human disorders will be available at the time of identification. Even the use of disease or regiodefined libraries or databases does not guarantee identification of genes with a disease association. However, the advantage of this approach is that it allows the researcher to choose the gene by the attributes of its product (e.g. presence of structural motifs or member of a known class of receptor).

Expression cloning
Genes identified by expression cloning will already have data associated with them; for example, they might be the receptor for a known ligand. Further value would be obtained by the use of cDNA libraries from defined sources (tissue-specific or disease/normal subtracted libraries).

Cloning by complementation
Genes cloned by complementation already have functional data associated with them. As with expression cloning, increased value can be obtained through the use of defined libraries.

sort of experimental filters that can be used to reduce these to a number more suitable for detailed study is outlined. In addition, many novel genes will initially be identified as ESTs rather than full-length cDNAs, and often with limited or no biological data. Because it is unclear what the patentable status of ESTs is[19,20], it is important to generate enough data to secure an intellectual property position as part of the initial validation process, and this will probably require full-length cDNA isolation and characterization.

The gene triage process

Prioritization by expression pattern

Before selecting the components of a target validation pipeline (outlined in Figure 4), it is worth stepping back and considering what properties the ideal therapeutic target would possess. Such a target might be a protein or nucleic acid, the function of which can be modulated to bring about the desired therapeutic effect with no undesirable physiological consequences. To achieve this, the target protein or nucleic acid should mediate a specific process that is directly associated with the biochemical mechanisms causing the disorder or its pathology. For many pharmaceutical companies, an ideal target might also function in a manner that can be modulated by small molecules rather than requiring the application of gene therapy, protein therapeutics or antisense technology. The target validation process should concentrate on determining which candidate genes most closely fit these requirements. It would seem likely that many genes will fail to meet the above criteria as they are essential for normal body function. Although it is certainly true that many genes, such as the genes for the *N*-methyl-D-aspartate receptor NR1 subunit[21] and apolipoprotein B (Ref. 22), are essential for embryonic development and beyond, some genes, such as *fosB* in mouse, that might have been considered essential, or at least important, for many biochemical activities are not[23]. Novel and potentially interesting genes are often identified in the absence of any definitive data concerning their role, vital or not, in the adult organism. In practice, this dilemma cannot be definitively resolved for the hundreds of genes that might be identified in a genomics program. Under these circumstances, it is necessary to be pragmatic. Thus, to reduce the number of genes entering the validation pipeline, a pre-screen can be established based on tissue expression patterns. Such a screen would not necessarily apply to all genes. For example, genes identified through human genetic studies should need no prioritization by this approach; however, the data obtained will be valuable (although lack of expression in the adult might indicate that attention should be focused on functions displayed earlier in embryological development). A more

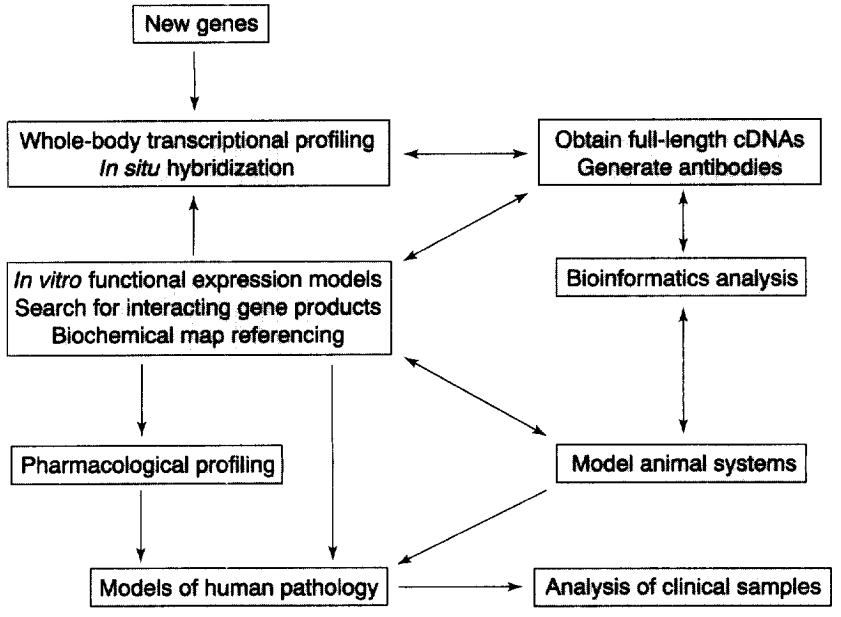

Figure 4. *New genes can enter the validation pipeline from several sources (see Box 2). To manage the large number of genes anticipated, the pipeline should incorporate prioritization assays. The expression pattern of the candidate genes within the whole body, and specifically within the organ or tissue of interest, can be used as the first level of prioritization, because such assays have a moderate-throughput capacity. The secondary validation assays, such as functional expression and bioinformatics analysis on full-length cDNAs, will be used depending on the particular gene under investigation, and may involve iterative cycles between systems to move along biochemical pathways to the eventual screening target. Eventually, pharmacological profiling of the candidate drug target can be determined (if compounds are available).*

detailed analysis of expression within the tissue or organ of interest would make use of *in situ* hybridization, although the utility of *in situ* studies depends on the anatomy of the target tissue. *In situ* data would be of great value for CNS genes and might supply a further level of prioritization. Both whole-body expression analysis, normally by Northern blot, and *in situ* hybridization can be carried out at moderate-throughput levels (tens to hundreds of genes) and do not require access to full-length cDNA clones, which often require an investment of time and resource to obtain, especially if the candidate gene was identified from database mining. As with other areas of pharmaceutical research, such analyses are not necessarily straightforward. The ubiquitous expression of a gene need not exclude its role in a tissue-specific disorder. For example, in Alzheimer's disease, the amyloid precursor protein is ubiquitously expressed and cleaved to produce β-amyloid, whereas the pathology associated with β-amyloid is only observed in the brain[24,25] (see also Box 3). Generally, candidate genes with expression restricted to the tissue or organ of interest would take priority over those having ubiquitous expression, while the genes not expressed in the target tissue could be dropped from consideration. Determining the expression pattern through development might also be used to assess the liabilities associated with subsequent gene-knockout studies and to gain some insight into function.

The pathway a gene follows through the validation process will vary, depending on the method by which it was identified (Box 2). However, for genes of unknown function, one of the main objectives would be to identify a function at both the cellular and the whole-organism levels. For genes identified through homology cloning, database mining, expression cloning, two-hybrid cloning and complementation cloning, a biochemical function can normally be assigned by virtue of the cloning method. For genes identified through human genetic studies or transcriptional profiling, even the biochemical function might be difficult to assign. An experimental evaluation of function requires a significant investment of resources, in terms of time and expertise. In addition, the elucidation of biological function will often not lead directly to the selection of the therapeutic target. The novel gene itself may not be the best target for therapeutic application; however, it may be the biochemical 'hook' for the identification and validation of the eventual therapeutic target. Depending on the degree of risk aversion within the particular organization, priority might be given to candidates that are likely therapeutic targets in their own right – these might be described as 'druggable' targets (my thanks to Geoff Duyk of Exelexis who introduced me to this useful and descriptive term). Examples of such candidate targets would be G protein-coupled receptors, enzymes and ion channels, and, more often than not, these would be the focus of database mining efforts.

Box 3. A gene for epilepsy

The identification of disease genes using linkage studies requires no knowledge of the disease gene function. However, given the significant amount of work required to identify such a gene, it is always hoped that its sequence might provide some clue as to its role in the disease. Unfortunately, this is not always the case. Recently, researchers in the USA and Finland have cloned the gene for progressive myoclonus epilepsy (*EPM1*)[49]; this disorder is a heterogeneous group of severe inherited epilepsies, characterized by myoclonic seizures, generalized epilepsy and progressive neurological deterioration. They found the gene coding for cystatin B, which is a cysteine protease inhibitor[50,51], and proposed that cystatin B might play a role in inactivating proteases leaking from lysosomes. Cystatin B is ubiquitously expressed, although EPM1 appears to be the only phenotype it produces in affected populations. This unexpected finding (i.e. genes directly involved with synaptic transmission were not identified) has led to speculation that there may be a functional commonality between lysosomes and synaptic vesicles[52]. However, more work is required to determine the role cystatin B plays in progressive myoclonus epilepsy.

Expression and knockdown studies

Once a candidate gene is selected for further evaluation, a full-length cDNA clone needs to be obtained. With the increasing need to identify and access the full coding region of novel genes regularly and efficiently, it is becoming routine to create high-quality cDNA libraries from the tissue sources of interest and to grid these libraries out on an individual clone basis, normally in 96-well plates. Gridded libraries can be probed with ESTs or oligonucleotide sequences from candidate genes and full-length positive clones rapidly identified for sequencing. In addition, the I.M.A.G.E. Consortium (http://www-bio.llnl.gov/bbrp/image/image.html) makes all characterized ESTs available as clones. Sometimes it is possible to use several ESTs to reconstitute a complete cDNA, although, more often than not, the ESTs represent sequences from the 3′ end of cDNAs and might not even contain a coding sequence. Full-length cDNAs will be required, not only for expression analysis, antisense-oligonucleotide design and designing antipeptide antibodies for *in vitro* and *in vivo* studies, but also for further structural analysis based on the full protein sequence. Because the analysis of protein expression, both *in vitro* and *in vivo*, is an essential part of functional evaluation, it is important to begin generating immunological reagents as soon as possible. Indeed, when using antisense knockdown as a tool for elucidating gene function, it is critical that the reagents are available to demonstrate reduction in protein expression. These reagents can be made using peptides, or through recombinant proteins expressed in bacteria. In addition, expression constructs can be made using epitope tags[26–28]. This is a valuable way of rapidly tracking the expression of proteins in cells, without the need to generate specific antibodies. It should be noted, however, that the tags might affect protein function and/or subcellular location. This concern would be especially acute with novel proteins having no known function. Immunological reagents produced against the unmodified gene product will enable tissue and subcellular protein expression to be determined immunohistochemically with more confidence (this is especially important in the brain, where the sites of protein synthesis and functional expression may be distant). In addition, these reagents will be important for *in vitro* expression studies. In parallel with antibody production, efforts to express the candidate gene in cell culture should be initiated, because a significant amount of data can be gained from *in vitro* expression studies in mammalian cells. For proteins having a potential role in signal transduction, expression systems can be designed to identify the pathways these proteins might affect. These systems are available commercially and comprise the various components required to reconstruct known signal transduction pathways and report their activity in cell culture[29]. Transient transfection assays may not be sensitive enough for phenotype analysis if the biochemical function of the protein is subtle. Creating permanent cell lines for expression of the candidate protein generates a tool with significantly more utility, but takes time to achieve. Often it is wise to embark on transient studies, because they require a limited time commitment.

When planning *in vitro* expression studies, the following questions should be considered:

- Based on the candidate gene under evaluation, should a specific cell host be selected (such as neurone, haemopoietic cell or vascular endothelial cell)?
- Is expression of the candidate gene likely to be detrimental to the cell during long-term culture?
- Is it possible that the host cell line displays endogenous expression of the candidate gene?

Many potential host cell lines are available from the ATCC (American Type Culture Collection; http://www.atcc.org/atcc.html) and these might make useful starting points. If the candidate gene shows activity in the *in vitro* expression

system (which would be the hope), this might lead to toxicity. If a permanent cell line is the eventual aim, it would be worth considering the use of an inducible expression system to allow control over when the gene is expressed. Recently, two inducible systems have been developed that not only give low background expression (mammalian inducible systems have been notoriously leaky for expression from so-called inducible vectors) but can also be used in a dose–response mode with the inducing agents, allowing some control of expression levels[30,31]. In addition, both systems can be used in transgenic animals. If the gene of interest is found to be endogenously expressed in cultured cells, consideration should be given to examining the effects of knockdown studies using antisense oligonucleotides. In cell culture, antisense sequences can be delivered as oligonucleotides[32] or even in antisense expression constructs. The combination of antisense and overexpression *in vitro* may be highly informative because the analysis of activity over a range of expression levels is possible.

If the function of the gene of interest is known, and especially if it represents a known class of pharmaceutical targets, expression in mammalian cells will enable a rapid determination of its pharmacological properties. Expression of the candidate genes in cell culture will often be the first opportunity for a pharmaceutical company to make use of its compound library (arguably its most valuable resource) in the target validation process. The most obvious value for screening a selection of proprietary, as well as public domain compounds, would be to evaluate the pharmacological profile of candidate genes from database mining efforts, as these would have been selected based on their homology to known pharmacological targets. In addition, one could also evaluate the effect novel genes had on known targets or biochemical pathways through the modification of the pharmacology (e.g. altering desensitization of G protein-coupled receptors). Systems such as PathDetect[29] and other reporter systems would be valuable in this regard. Finding pharmacological tools early in the evaluation process will enable rapid *in vivo* analysis of function and may also aid in the construction of disease models.

Animal models

The functional evaluation of many novel genes will not easily succumb to *in vitro* analysis, either because they only function in specialized cell systems or because they require complex multicellular systems to reveal the phenotype. In these cases, functional analysis will require the application of animal model systems (Box 4). The choice of animal model will depend on the expected biology of the candidate gene under consideration.

The use of less complex organisms such as yeast, worms or flies will, in many cases, be valuable in determining function, especially for genes whose function is required during development. However, the determination of function will not necessarily lead to the identification of the appropriate therapeutic target. For this, a mammalian animal model will be required. Several mammalian species can be used for transgenic models of gene overexpression, but at present only the mouse is available as a model for gene knockouts or mutagenesis via homologous recombination[33]. Many natural mouse mutant strains have been described[34] and these might provide a valuable resource for disease models associated with novel genes, in addition to being the route for new gene discovery. However, the dilemma remains that a mouse model of a disorder cannot predict the human disease process with certainty. In addition, the use of animal models that were developed and 'validated' using clinically efficacious compounds might not be appropriate for the novel targets that emerge from genome research. Whenever possible, reference to human disease biology should be made either through the availability of surrogate markers or through the use of human polymorphism detection – not only of the genes coding for the potential therapeutic target, but also of genes whose products fall on the same biochemical pathways. As more data on mapped genes become available, the latter approach should become more feasible. Some targets, such as those involved in disorders of behaviour, will be harder if not impossible to validate in animals through the use of surrogate markers, as they probably involve subtle alterations in neurochemical pathways[35] and may necessitate a human genetics approach.

Conclusion

This review has attempted to give an indication of the challenge the pharmaceutical industry is facing in turning the information content of the human genome into therapeutically valuable and viable commodities. The magnitude of this challenge is clearly illustrated when we consider that, to date, this industry has produced drugs acting against just over 400 human molecular targets[36], and, with screening and combinatorial chemistry technologies increasing the efficiency of the discovery process, the demand for suitable targets is rapidly outstripping the supply. It has been estimated that between 3,000 and 10,000 new targets for

Box 4. The use of animal models for gene function studies

It is fortunate that it is not only the human genome that is subjected to intense sequencing efforts. The genomes of several model organisms are presently being sequenced. For *Saccharomyces cerevisiae* it has been completed[53] and the sequence of the nematode worm *Caenorhabditis elegans* is >50% complete. At present there is limited coverage of other model organisms such as the fruit fly, *Drosophila melanogaster*, and the mouse.

The value of model organisms such as the mouse requires little justification. As a mammal, the mouse is closely related to humans and displays many pathologies similar to those of humans, obesity being a good example[11,12,14]. The value of lower animals relates more to their utility in the determination of gene function rather than pathology. How yeast and *C. elegans* aid the functional characterization of novel genes is described below. However, other model organisms, such as *D. melanogaster* and the puffer fish, *Fugu rubripes*, are also likely to play significant roles in the functional characterization of novel genes[54,55].

Single-cell systems

Yeasts are single-celled eukaryotic organisms that are almost as easy to culture and genetically manipulate as *Escherichia coli*. The genome comprises 13,105,020 bp arranged on 16 chromosomes[56]. Considering the evolutionary separation between humans and yeast, it is striking that so many yeast genes are homologous to human genes[57]. Determining gene function in yeast is comparatively easy, so it is anticipated, based on the knowledge of its genome sequence and its facile genetics, that yeast will be a productive tool for a subset of therapeutic targets, including genes involved in tumorigenesis or hyperplasia where mutagenesis followed by clonal amplification often initiates the pathology.

Multicellular systems

Many genes display their function only within the context of a multicellular system, or at a specific point in the developmental process. Therefore, well-characterized multicellular model organisms are likely to become valuable tools for the dissection of gene function. *C. elegans* has been used as a model organism to analyze the genetics of development since its utility was first recognized by Sidney Brenner in the 1960s. The adult organism contains 959 cells, and the origin of each of these cells has been tracked through the developmental process[58]. Rather like yeast, many *C. elegans* genes show functional homology to mammalian genes. An example of where our understanding of a mammalian system has been greatly enhanced by studying *C. elegans* is the phenomenon of programmed cell death, or apoptosis[59].

Apoptosis is used to remove unwanted or damaged cells in an orderly fashion. It is a vital process for the normal development of multicellular organisms and is likely to play a role in disorders such as Alzheimer's disease and Parkinson's disease[60]. Mutations in 14 different genes have been shown to affect apoptosis in *C. elegans*. Most are involved in the death of small cell populations; however, three genes (*ced-3*, *ced-4* and *ced-9*) are involved in the 'death' decisions of all cells; *ced-3* and *ced-4* are pro-apoptotic, while *ced-9* is anti-apoptotic. Genetic studies using *C. elegans* double mutants have demonstrated the epistatic relationship between these genes – *ced-9* functioning upstream of *ced-3* and *ced-4*; *ced-9* is homologous to the mammalian anti-apoptotic *Bcl2* gene and *ced-3* is homologous to the cysteine protease family of genes[61]. However, until recently no *ced-4* homologue had been identified in mammalian cells. It has been shown that the *ced-4* gene product can interact with both Bcl-x_L (a member of the Bcl-2 family) and interleukin-1β converting enzyme (ICE, a member of the cysteine protease family)[62]. It is clear, based on these observations, that a mammalian *ced-4* functional homologue exists that biochemically couples the activities of the cysteine protease and Bcl-2 family of proteins.

Transgenic mouse models have been available to the research community for some time[33,63]. The value of these animals lies in their use both for the analysis of gene function and for the creation of disease models. However, the use of mouse transgenic models can be limited by the time it takes to generate the appropriate strains and other factors, such as regiospecific expression of the transgene, as well as the impact of expression or gene knockout on the developmental process. The use of antisense technology in the context of gene 'knockdown' has the potential to overcome some of these problems[64]. In addition, antisense oligonucleotides designed and validated for *in vitro* studies can be used *in vivo* to examine gene knockdown effects rapidly. Although effective oligonucleotide design has been a problem, new empirical methods of antisense oligonucleotide selection are becoming available[32].

therapeutic intervention will emerge from genomic research[30], and it is these new targets that will fill discovery pipelines in the future. It is likely that the limiting phase of the drug discovery process will soon become the validation of the therapeutic target. This process therefore needs to be made as efficient as possible, not only to fuel the discovery/development engine, but also to ensure a solid proprietary position for each new target. Clearly, many technologies, some of which are not normally associated with pharmaceutical research, will be required to characterize and develop the new therapeutic targets, and it is likely that no single organization will be able, or even want, to develop all

these in-house. This is clearly illustrated by considering the unpredictability of the validation process that is required for genes identified through human genetic studies. The role these genes play in the disease process can be obscure, and it will be impossible to predict what the validation pathway will be. For this reason, alliances between major pharmaceutical organizations and genome-based biotechnology companies based on specific technologies, rather than therapeutic focus, are likely to become more common. It is also likely that multiple smaller collaborations or alliances will be required to facilitate the efficient discovery and development of the next generation of pharmaceuticals. A model for collaborations that might gain favour in the future, as it would allow the maximal use of cutting edge technologies, would be the linear organization of alliances for gene discovery, target validation and ultra-high-throughput screening coupled with combinatorial chemistry.

The industry must also face the fact that novelty, by its very nature, creates a degree of uncertainty as compounds against these new targets move towards the clinic. Some of the alternative approaches for gene discovery discussed above offer the advantage of target-type selection, either up-front, as in database mining, or through prioritization, as in transcriptional profiling. However, the role these genes might play in the human disease process may eventually require validation in the clinic. The industry will have to rely more heavily on novel animal models of disease that cannot be 'back-validated' with known drugs. However, genomics research offers the opportunity of streamlining the development process, by increasing the tools available for drug metabolism and toxicology studies, as well as patient selection for clinical trials. Genomics is therefore set to change all aspects of the drug discovery and development process.

Acknowledgements

I thank my colleagues at Wyeth-Ayerst Research, especially Jim Barrett, Maria Betty, Mark Cockett, Andrew Wood and Ken Rhodes, for their contributions to the preparation of this article.

REFERENCES

1 Antequera, F. and Bird, A. (1994) *Nat. Genet.* 8, 114
2 Fields, C. *et al.* (1994) *Nat. Genet.* 7, 345–346
3 Schuler, G.D. *et al.* (1996) *Science* 274, 540–546
4 Papadopoulos, N. *et al.* (1994) *Science* 263, 1625–1629
5 Hillier, L.D. *et al.* (1996) *Genet. Res.* 6, 807–828
6 Selkoe, D.J. (1997) *Science* 275, 630–631
7 Collins, F.S. (1995) *Nat. Genet.* 9, 347–350
8 Murphy, K.C., Cardno, A.G. and McGuffin, P. (1996) *J. Mol. Neurol.* 7, 147–157
9 Pulver, A.E. *et al.* (1995) *Am. J. Med. Genet.* 60, 252–260
10 Roelfsema, J.H. and Breuning, M.H. (1996) *Adv. Nephrol. Necker Hosp.* 25, 131–145
11 Fan, W. *et al.* (1997) *Nature* 385, 165–168
12 Lee, G.H. *et al.* (1996) *Nature* 379, 632–635
13 Tartaglia, L.A. *et al.* (1995) *Cell* 83, 1263–1271
14 Zhang, Y. *et al.* (1994) *Nature* 372, 425–432
15 Fields, S. and Song, O. (1989) *Nature* 340, 245–246
16 Ko, C.H. and Gaber, R.F. (1991) *Mol. Cell. Biol.* 11, 4266–4273
17 Seed, B. (1995) *Curr. Opin. Biotechnol.* 6, 567–573
18 Simonsen, H. and Lodish, H.F. (1994) *Trends Pharmacol. Sci.* 15, 437–441
19 Auth, D.R. (1997) *Nat. Biotechnol.* 15, 911–912
20 Eisenberg, R.S. (1996) *Trends Biotechnol.* 14, 302–307
21 Forrest, D. *et al.* (1994) *Neuron* 13, 325–338
22 Farese, R.V., Jr *et al.* (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 1774–1778
23 Brown, J.R. *et al.* (1996) *Cell* 86, 297–309
24 Selkoe, D.J. (1991) *Neuron* 6, 487–498
25 Selkoe, D.J. (1993) *Trends Neurosci.* 16, 403–409
26 Kolodziej, P.A. and Young, R.A. (1991) *Methods Enzymol.* 194, 508–519
27 Wang, L.F. *et al.* (1996) *Gene* 169, 53–58
28 Sells, M.A. and Chernoff, J. (1995) *Gene* 152, 187–189
29 Xu, L., Sanchez, T. and Zheng, C-F. (1997) *Strategies* 10, 1–3
30 No, D., Yao, T.P. and Evans, R.M. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 3346–3351
31 Gossen, M. *et al.* (1995) *Science* 268, 1766–1769
32 Ho, S.P. *et al.* (1996) *Nucleic Acids Res.* 24, 1901–1907
33 Jacobson, D. and Anagnostopoulos, A. (1996) *Trends Genet.* 12, 117–118
34 Darling, S. (1996) *Curr. Opin. Genet. Dev.* 6, 289–294
35 Campbell, I.L. and Gold, L.H. (1996) *Mol. Psychiatry* 1, 105–120
36 Drews, J. (1996) *Nat. Biotechnol.* 14, 1516–1517
37 Schena, M. *et al.* (1995) *Science* 270, 467–470
38 Schena, M. *et al.* (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 10614–10619
39 Liang, P. and Pardee, A.B. (1992) *Science* 257, 967–971
40 Prashar, Y. and Weissman, S.M. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 659–663
41 Lockhart, D.J. *et al.* (1996) *Nat. Biotechnol.* 14, 1675–1680
42 DeRisi, J. *et al.* (1996) *Nat. Genet.* 14, 457–460
43 Fickett, J.W. (1996) *Trends Genet.* 12, 316–320
44 Bork, P. and Bairoch, A. (1996) *Trends Genet.* 12, 425–427
45 Boguski, M.S. (1995) *New Engl. J. Med.* 333, 645–647
46 Savitsky, K. *et al.* (1995) *Hum. Mol. Genet.* 4, 2025–2032
47 Savitsky, K. *et al.* (1995) *Science* 268, 1749–1753
48 Watson, S. and Arkinstall, S. (1994) *The G-Protein Linked Receptor Facts Book*, Academic Press
49 Pennacchio, L.A. *et al.* (1996) *Science* 271, 1731–1734
50 Turk, V. and Bode, W. (1991) *FEBS Lett.* 285, 213–219
51 Jarvinen, M. and Rinne, A. (1982) *Biochim. Biophys. Acta* 708, 210–217
52 O'Brien, C. (1997) *Science* 271, 1997
53 Mewes, H.W. *et al.* (1997) *Nature* 387 (Suppl. 6632s), 7–65
54 Osborne, K.A. *et al.* (1997) *Science* 277, 834–836
55 Brenner, S. *et al.* (1993) *Nature* 366, 265–268
56 Goffeau, A. *et al.* (1996) *Science* 274, 546 and 563–567
57 Bassett, D.E., Boguski, M.S. and Hieter, P. (1996) *Nature* 379, 589–590
58 Sulston, J.E. *et al.* (1983) *Dev. Biol.* 100, 64–119
59 Hengartner, M.O. (1996) *Curr. Opin. Genet. Dev.* 6, 34–38
60 Spence, P. *et al.* (1996) *Expert Opin. Ther. Pat.* 6, 345–360
61 Steller, H. (1995) *Science* 267, 1445–1449
62 Chinnaiyan, A.M. *et al.* (1997) *Science* 275, 1122–1126
63 Gordon, J.W. *et al.* (1980) *Proc. Natl. Acad. Sci. U. S. A.* 77, 7380–7384
64 Christoffersen, R.E. (1997) *Nat. Biotechnol.* 15, 483–484