Chemical ligands, genomics and drug discovery

George R. Lenz, Huw M. Nash and Satish Jindal

The sequencing of the human genome and numerous pathogen genomes has resulted in an explosion of potential drug targets. These targets represent both an unprecedented opportunity and a technological challenge for the pharmaceutical industry. A new strategy is required to initiate small-molecule drug discovery with sets of incompletely characterized, disease-associated proteins. One such strategy is the early application of combinatorial chemistry and other technologies to the discovery of bioactive small-molecule ligands that act on candidate drug targets. Therapeutically active ligands serve to concurrently validate a target and provide lead structures for downstream drug development, thereby accelerating the drug discovery process.

Opportunities and challenges for small-molecule drug discovery in the post-genomic era

The future of drug discovery holds great promise, largely because of the introduction of new technologies that are beginning to revolutionize the scale and success rate at which truly novel and high-value drugs can be brought to market. New biological (genomic) technologies now enable genome-wide DNA sequencing and genome-wide gene expression analysis, thereby providing an opportunity to identify the best drug discovery targets that will lead to the best small-molecule drugs for many major diseases. The accompanying challenge is how to be the first to identify these targets from thousands of candidate macromolecular targets (primarily proteins, but also including nucleic acids). Given the ubiquitous nature of most primary genomic information, drug discovery is an increasingly dynamic competition that will favor developers and early adapters of new technologies.

Genomic technologies

Currently, biological technologies (both computational and empirical) at the level of the genome are unable to provide unique solutions to the question of which protein is the best small-molecule drug target for a given disease. Instead, these genomic technologies identify intermediate sets of partially characterized, disease-associated proteins. These technologies can be grouped into two general categories:

- Global strategies to identify proteins that are associated with a particular disease
- Target-specific strategies to provide partial characterization of the disease-associated proteins.

Global strategies for identification of candidate drug targets include computational homology alignments (i.e. between host and pathogen genomes)[1], differential gene expression analyses (i.e. cDNA or mRNA quantitation)[2] and whole proteome analyses [i.e. 2-D PAGE (two-dimensional polyacrylamide gel electrophoresis) or LC–MS/MS (reverse-phase liquid chromatography coupled to two-dimensional mass spectrometric analysis)][3] to identify disease-specific changes in protein expression levels or isoforms. Both transcriptional and proteomal technologies suffer from an inability to discriminate between the root molecular cause(s) of a disease and the much greater number of molecular effects (or symptoms), which is especially important in the analysis of human diseases.

Target-specific strategies include targeted gene disruption (gene knockouts), antisense and ribozyme inhibition of

George R. Lenz*, **Huw M. Nash** and **Satish Jindal**, NeoGenesis, 840 Memorial Drive, Cambridge, MA 02139, USA. *tel: +1 617 868 1500, fax: +1 617 868 1515, e-mail: grlen@neogenesis.com, nash@neogenesis.com, satjin@neogenesis.com

 1359-6446/00/$ – see front matter PII: S1359-6446(00)01468-9

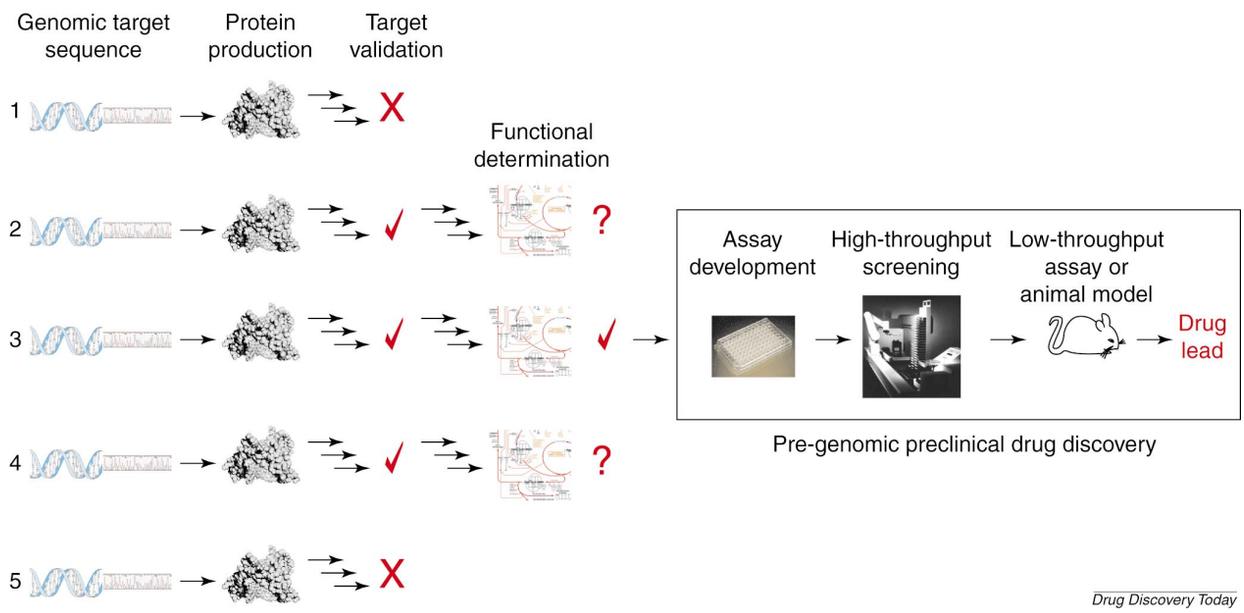

Figure 1. *The post-genomic drug discovery pathway, where target validation and detailed functional determination of candidate target proteins are placed upstream of assay development and high-throughput screening. This pathway cannot be scaled rapidly and economically to address the large sets of putative targets emerging from genomics.*

mRNA function, and computational modeling to predict the structure and/or function of gene products. Gene knockouts are laborious to perform in mammalian cells, thus limiting the scalability of this approach for studying human diseases, and this technology relies on the mouse as a model for human biology. In addition, the resultant knockout mice can yield inconclusive results because of developmental compensation or other forms of functional redundancy.

By comparison, knockouts in pathogens are relatively simple and rapid, and can efficiently identify the essential genes for survival or pathogenicity of the microorganism. Such studies have indicated that, on average, one-third of the genes within a microbial genome are essential for viability, usually in the order of hundreds to thousands[4]. Furthermore, knockout strains do not provide information on the relative susceptibility of each essential gene product to functional antagonism by a small molecule (sometimes referred to as the drugability of the protein).

Computational methodologies for structure and function prediction include sequence homology alignments [i.e. PSI-BLAST (position-specific iterated basic local alignment search tool)][5] and threading or *ab initio** techniques for structure generation. PSI-BLAST and threading algorithms rely on mining large databases that are currently incomplete in their population and annotation that, unfortunately, are unlikely to be filled for several years. For example, the function of the genes that encode 40–60% of all pathogen genomes sequenced and analyzed to-date and probably most of the 140,000 genes that encode the human genome are still unknown[6,7].

From a structural perspective, the Protein Data Bank (PDB) contains high-resolution structural representatives of only a few hundred-fold protein-fold topologies, with 1000–3000 of these topologies estimated in nature[8]. Although *ab initio* methods for structure prediction do not rely on the PDB, current algorithms cannot reliably predict accurate structures at a meaningful resolution[9,10]. In addition, objective evaluations of current database-driven structure and function prediction algorithms have concluded that these methods cannot predict the structure and function of most novel proteins at the necessary precision to confidently assign function and develop HTS assays[11,12].

Strategic limitations of genomic technologies

Genomic technologies are therefore yielding intermediate sets of putative targets that can overwhelm the downstream discovery capabilities of even the largest pharmaceutical

* *Ab initio* techniques are strictly defined as structure-prediction algorithms that rely exclusively on first principles of physical chemistry, and are more loosely defined by some in the field to describe structure-prediction algorithms that use pattern-recognition algorithms in the absence of statistically significant homology to any protein of known structure.

companies. As shown in Figure 1, target validation and detailed functional determination of candidate target proteins are placed ahead of assay development and HTS in the conventional drug discovery pathway. For reasons already discussed, target validation is costly and time-consuming for human proteins, and detailed functional determination is equally challenging for microbial and human proteins with <20% homology to proteins of known function[13,14]. In both cases, these steps usually require the equivalent of 1–2 researchers (one PhD level scientist plus one research associate) for 18–24 months, making it economically prohibitive to analyze every candidate target protein in parallel. This disconnection between the scales at which biological technologies can identify putative targets and at which high-resolution structural and functional determination and HTS assay development can be performed is a serious bottleneck that is holding back the full impact of genomics on drug discovery.

Exacerbating this situation is the fact that, prior to the advent of genomics, functional determination and target validation studies were performed primarily in academic and medical research institutions, with pre-clinical research within the pharmaceutical industry usually being initiated at the assay development stage (Fig. 1). These studies involve both the genomic technologies already described and basic research methodologies, such as site-directed mutagenesis, *in situ* hybridization and immunocytochemical assays. Hence, genomics is increasing the cost and time of pre-clinical drug discovery.

There is therefore a requirement for new strategies to initiate small-molecule drug discovery with intermediate sets of incompletely characterized, disease-associated proteins. Ideally, this should rationally and economically focus pre-clinical drug discovery on the best drug targets emerging from genomics. One such strategic model is the early and extensive application of combinatorial chemistry and other technologies to the discovery and analysis of biologically active small-molecule ligands that act on these candidate drug targets.

A role for small-molecule ligands in target validation and lead discovery

Ligand-directed drug discovery model

An emerging strategic model for initiating small-molecule drug discovery with novel proteins from genomics involves the early discovery of biologically active ligands that act on candidate drug targets. As shown in Figure 2, the ligand-directed drug discovery pathway comprises three basic steps. In step 1, genes encoding a set of putative target proteins identified by any of the aforementioned genomic technologies are cloned into standard expression systems and a few milligrams of each recombinant protein is expressed and purified. In step 2, affinity selection experiments screen the purified proteins against large libraries of biopolymers (modified peptides) or small organic molecules to identify those molecules that bind with high affinity and specificity to each protein. In step 3, the best ligands for each protein are tested for biological activity in one or more low-throughput biological assays (i.e. pathogen-specific toxicity assays, viral proliferation assays, general immune response assays).

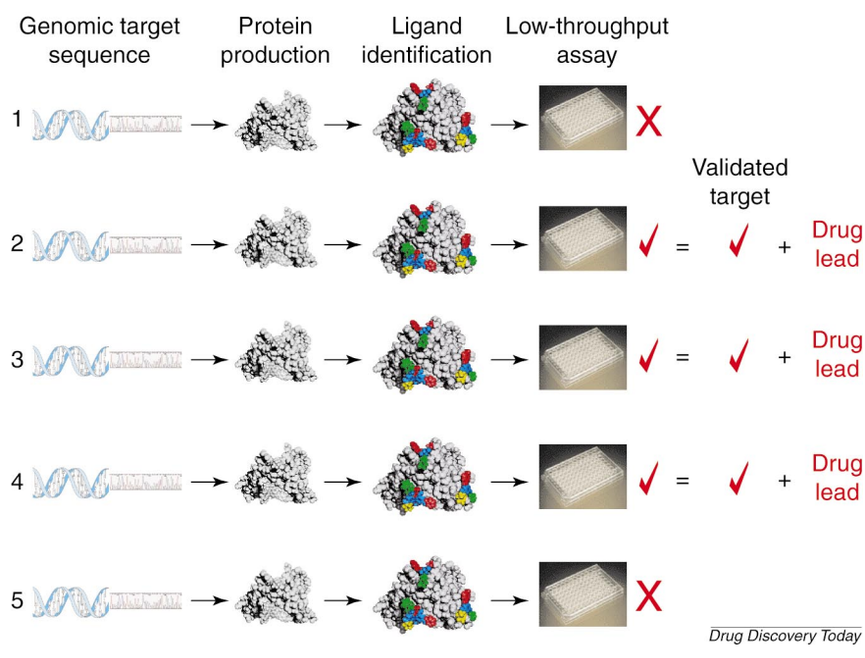

Figure 2. *The ligand-directed drug discovery pathway, where the identification of a therapeutically active ligand to a target protein can serve to concurrently validate the target and provide a lead structure for downstream drug development.*

Advantages of ligand-directed drug discovery

A distinct advantage of ligand-directed drug discovery is the concurrent analysis of both the importance of a protein in a disease process and a protein's amenability

to functional modulation by a small molecule. This latter feature differentiates ligand-directed discovery from all other genome-level biological approaches. Furthermore, ligand discovery technologies rely on affinity at the primary screening level and, thus, assay development can be initiated with a partial understanding of the protein's function or structure. This attribute is highly significant when considering the vast number of incompletely characterized, disease-associated genomic targets.

Affinity-based screens also enable the placement of one or more non-HTS biological assays downstream of the large-scale primary screen to study the activities of the small subset of molecules that possess requisite affinities and specificities. For many proteins, an affinity-based screen can provide a set of ligands that exhibit multiple biological activities, owing to the existence of multiple binding modes and even multiple distinct binding sites. The exploration of these binding sites is especially relevant to human protein targets, as these proteins frequently possess more than one functional domain. In addition, new biological technologies, such as yeast two-hybrid screens, have supported a growing appreciation that protein–protein interactions are key components of most, if not all, biological systems, and that modulation of these interactions represent an under-exploited area for drug discovery[15,16]. The ligand-directed drug discovery model provides perhaps the best mechanism for attacking this difficult area, and indeed some notable successes in disrupting and enforcing protein–protein interactions have been achieved with affinity-based technologies[17,18].

Finally, most ligand discovery technologies are highly scalable, and can handle the large numbers of proteins provided by genomic information. At the genome level, biologically active ligands can serve as empirical filters, and they offer a rational bridge between the scale of genomic target identification and the capacity of downstream discovery resources.

Biopolymer ligands

With all of its apparent advantages, the ligand-directed drug discovery model has failed to become the industry standard for the exploitation of new targets from genomics. The primary weakness of this model is the nature of the ligands used to date. Early ligand-directed functional genomic strategies have featured biopolymer ligands such as peptides and RNA aptamers, primarily because of the existence of several powerful methodologies for the screening of multimillion- to one billion-member libraries of these candidate ligands. Technologies that utilize extracellular presentation of candidate ligands take advantage of biological systems to synthesize the libraries and auto-amplify positive ligands between rounds of affinity selection[19]. Other related technologies employ intracellular presentation of candidate ligands coupled to phenotypic readouts and selectable markers[20]. However, biopolymer ligands have failed to completely bridge the gap between genomics and drug discovery for two reasons. Firstly, it is apparent that biopolymers do not fully cover drug diversity. Secondly, these ligands do not possess suitable biological stability and membrane permeability properties to be drug leads themselves and, thus, active biopolymer ligands are usually viewed as hits that must be reengineered to develop leads.

Unfortunately, there exists no reliable approach to directly convert biopolymer ligands into drug leads, as they are exquisitely sensitive to changes in their polymeric backbone that increase stability and permeability. Biopolymer ligands also possess numerous rotatable bonds, making it difficult to develop a three-dimensional model of the protein-bound conformation for designing non-biopolymer lead molecules.

Biopolymer ligands have been employed as surrogate markers in ligand displacement assays to discover small molecules that bind competitively. This strategy requires two cycles of method development and primary screening, with one cycle of screening against a biopolymer library and a second ligand displacement screen against a library of non-biopolymer small molecules. Obviously, it would be more efficient to use a one-step strategy to directly identify drug-like small-molecule ligands, and this approach is now redefining ligand-directed drug discovery.

Small-molecule ligands

It is becoming apparent that the optimal ligands for preclinical discovery will be small molecules that possess many features required for the final orally available small-molecule drug. Specifically, optimal ligands will possess high affinity and specificity for the target protein and will have reasonable membrane permeability to maximize the probability of significant biological activity in whole cell assays. Hence, therapeutically active small-molecule ligands can provide concurrent validation of novel targets and provide useful lead structures for downstream drug development, while accelerating the drug discovery process for targets that are incompatible with the conventional drug discovery model of full functional determination and validation, followed by HTS assay development and screening. As already discussed, this incompatibility arises from the nature and number of targets being provided by

genomics. In a sense, small-molecule ligand discovery represents a third generation of genomics discovery: DNA sequencing provides the primary data set, bioinformatics and mRNA/protein expression analyses provide a secondary data set, and bioactive small-molecule ligands provide a tertiary data set by mapping the interface between protein diversity and drug diversity.

Technologies for the discovery of small-molecule ligands

The promise of small-molecule ligand-directed drug discovery has catalyzed significant activity within the biotechnological and pharmaceutical industries to develop high-throughput affinity screening technologies and complementary small-molecule libraries for the discovery of drug-like small-molecule ligands. The most powerful ligand discovery platforms have incorporated recent advances in combinatorial chemistry, analytical instrumentation and data analysis technologies. Criteria that will ensure a high probability of finding chemical ligands for any potential binding site on a protein are:

- An inclusive set of highly diverse compounds that are significantly more diverse than at present. Although the number of compounds required is still open to discussion, it is undoubtedly large (millions), which will require the synthesis of large sets of compounds.
- Efficient methods of screening.
- Bioinformatics and cheminformatics. Although the least developed, this has the potential to make a significant contribution to the chemical ligand identification process.

As more targets are screened and the information becomes available in databases, compound libraries could be synthesized or prioritized based on protein family and selectivity could be improved by comparison with other family members. As genomic gene-chip information technology (IT) improves, the effect of a chemical ligand on a cell will be a quicker and more efficient method for facilitating functional genomics experiments and planning further pharmacological profiling. Several technologies for identifying chemical ligands are currently under investigation by entrepreneurial companies and academics, although few have been put into practice[21].

Schreiber approach to chemical ligands

Stuart Schreiber at the Harvard Institute of Chemistry and Cell Biology (Cambridge, MA, USA) has been investigating the concept of chemical genetics to discover small-molecule ligands for proteins[22]. In Schreiber's terminology, 'forward chemical genetics' is when large numbers of compounds are screened for a particularly biological activity, and the resultant bioactive ligand is used to identify the target and it's role in the pathway. In 'reverse chemical genetics', a small-molecule ligand is sought against a protein for which little, if any, information is known and the ligand is then used to probe the protein and determine its function.

This concept is based on combinatorial libraries derived from natural product templates known to bind to proteins, providing the best opportunity for discovering chemical ligands that target protein surfaces. Although significant insight into the requirements of small molecules binding to proteins is not available, most natural product-based protein ligands are relatively large, are functionally and stereochemically complex, and possess a fairly rigid structure. This rigidity is considered important in maintaining the required stereochemical relationships of the functional groups required for binding and increased affinity from reductions in entropy caused by fewer degrees of freedom. However, this has significantly increased the size of the screening libraries required to find ligands. Hence, Schreiber intends to prepare millions of compounds based on particular natural product templates.

The first reported example of such a library is based on naturally occurring shikimic acid [compound (**1**), Fig. 3] and contains 2.18 million discrete molecules[23]. The construction of this mixture combinatorial library converts shikimic acid into a tetracycle [compound (**2**), Fig. 3], which uses the free acid group to connect to a solid-phase resin suitable for split-and-pool synthesis[24]. The aromatic iodo-group is replaced by a series of alkynes using palladium coupling. A series of 54 amines then opens the lactone ring to form a hydroxy amide that is subsequently acylated with 44 different acids to form a library [see compound (**3**), Fig. 3]. Encoding was accomplished by a variation of Still's method[25]. This approach enables the use of a wide variety of synthetic reactions, together with multiple-connection chemistries that yield large compound libraries based on a complex natural product scaffold.

To miniaturize library screening, the ultimate objective is to produce the protein equivalent of a gene-chip where an array of proteins, potentially an entire genome, can be probed with library members. Initial progress has been made in nanoliter-scale assays using fluorescence as a read-out (for an animated description of the procedure, see http://iccb.med.harvard.edu/). A low-volume assay has been developed for whole cells where nanoliter droplets can be formed from a suspension of medium-containing beads and cells. A simple spray gun has then been devised

to produce ≈10,000 discrete nanodroplets on a small plastic dish[26]. Subsequently, the compounds can be released from the beads and the biological activity determined.

The screening methodology has been extended to a miniaturized array format, without spatial coding, facilitating the screening of large numbers of compounds against a single target[27]. Further progress has been made, with a preliminary experiment suggesting that the spatial array of potential chemical ligands might be possible that are capable of being probed by a (soluble) protein. Three known chemical ligands were attached using a sulfydryl-capped tether to a maleimide functionalized surface. Multiple robotic spotting on a glass slide yielded a spatial array of 11,000 spots. Interspersing a dye provided spatial orientation. Probing the array using the fluorescently labeled proteins and antibody then provided an accurate readout reflecting the affinity of the probe for the ligand[28].

The overall concept is to screen large libraries of arrayed compounds using a set of differentially labeled fluorescent proteins to identify their chemical ligands. Although early results are encouraging, the number of compounds requiring to be assayed based on the natural product scaffold concept would be vast, requiring >100 plates to screen the single library based on compound (**3**) (Fig. 3).

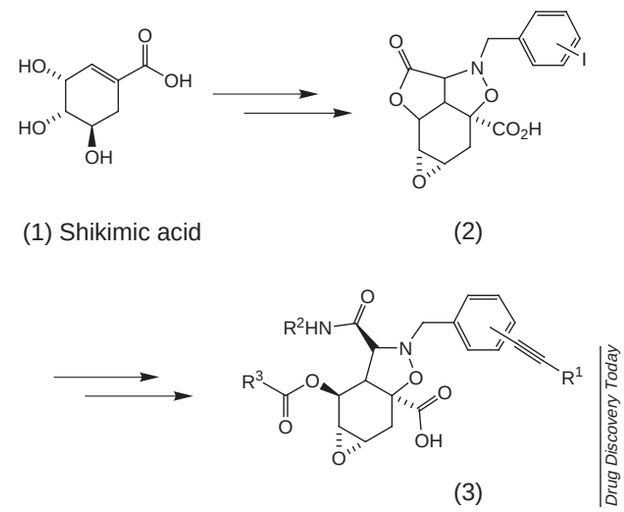

Figure 3. *The synthetic scheme of Schreiber and coworkers for the formation of a natural product-like combinatorial library (R^1 = 50 alkynes, R^2 = 87 amines, R^3 = 98 acids)[22]. Compounds (**2**) and (**3**) are bound to the resin through the free acid.*

Direct affinity approaches to chemical ligands

The discovery of chemical ligands for all potential binding sites on a protein generally involve an affinity selection process in which a library of compounds, usually as a mixture, is incubated with the protein. From there, the major challenge is the detection and identification of the chemical ligands[29], and key factors are:

- The type of libraries used
- The compound encoding/deconvolution mechanism
- The integrated data handling (IT) capability.

Phage display. The use of phage display for the identification of peptide-based chemical ligands is well established (Fig. 4)[30]. After identification, the peptide functions as a surrogate ligand for small non-peptidic molecules in competitive displacement HTS. Novalon (Durham, NC, USA) has pioneered this approach and developed several screening methods based on this technology.

Affinity screening. ATLAS (any target ligand affinity screen), developed by Scriptgen Pharmaceuticals (Waltham, MA, USA), involves an affinity screen using either single or small mixtures of compounds. Throughput is stated to be approximately 5000 individual compounds per week and can be increased using small mixtures. Detection is dependent on a change in protein conformation when complexed with a ligand. The patent literature indicates that there are several detection systems available for use in this technology depending on the target type[31]. These involve specifically recognizing either the free or the ligand-bound forms of the target. Antibody discrimination, differential proteolysis, differential binding of known ligands, and inhibition of aggre-

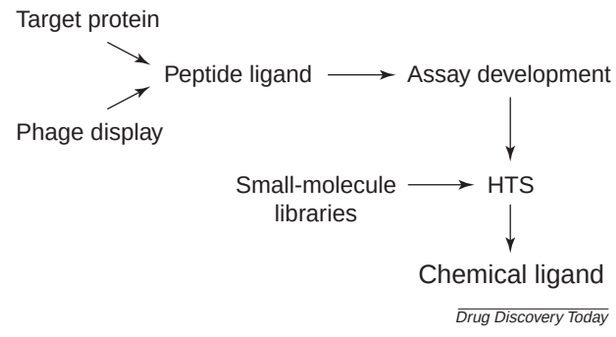

Figure 4. *A general scheme for using phage display in a two-step approach to discover drug-like, small-molecule, chemical ligands. Step one involves the discovery of a bioactive peptide ligand, and step two uses this peptide ligand to discover a bioactive small-molecule ligand via a high-throughput screening (HTS) peptide displacement assay.*

gation have been described. This chemical ligand identification procedure requires a biological readout. The affinity screening technologies for both proteins and nucleic acids have been combined into a technology designated RAPTV (rapid pharmacological target validation; Michael G. Palfryman, Scriptgen Pharmaceuticals, personal communication).

The percentage of hits obtained using ATLAS depends on the type of target screened and the size of the functional area being probed by the detection procedure employed. Scriptgen estimates that 10–30% of chemical ligands identified are biologically active in secondary functional assays.

Combined affinity selection with MS. Chiron (Emeryville, CA, USA) has reported a combined affinity selection and MS approach for finding biopolymer chemical ligands for protein targets[32] In this assay, a relatively small library of biopolymers with differing lengths and MWs is mixed with a target protein and subjected to size-exclusion chromatography, isolating any target–ligand complexes. The complex is desalted and concentrated on-line using reverse-phase chromatography to separate the biopolymer from the target protein. The MW, determined by MS, indicates the structure, which can be further confirmed using MS/MS.

Ultra-high-throughput chemical ligand identification. NeoGenesis (Cambridge, MA, USA) has developed an ultra-high-throughput chemical ligand identification process for small molecules termed ALIS (automated ligand identification system). In this automated, integrated system (Fig. 5), a library consisting of hundreds to thousands of compounds is incubated with a target protein, free in solution, and then passed through a micro-scale size-exclusion column that separates the protein and its bound ligands from the remaining library members. Dissociation and concentration occurs in a micro-reverse phase LC column, which feeds into the mass spectrometer for structural identification.

The source chemistry is termed NeoMorph, and is a mixture-based combinatorial chemistry[33,34]. The libraries are synthesized in solution using a core plus building block approach (Fig. 6) where, for example, one core with three connection sites is coupled to a set of 15 diverse amine building blocks to furnish a mixture combinatorial library of 3375 compounds. Hence, the large numbers of compounds necessary can be prepared for simultaneous evaluation in the ALIS affinity screen. During library design, the compounds are encoded by mass using a proprietary algorithm such that all building blocks with core combinations in the library have different MWs. Libraries of several thousand compounds, differing by at least 0.05 amu (atomic mass units) can be prepared, enabling the MW to remain within a drug-like range, and facilitating direct identification of chemical ligands by MS using their MW. This mass coding algorithm can also be extended to

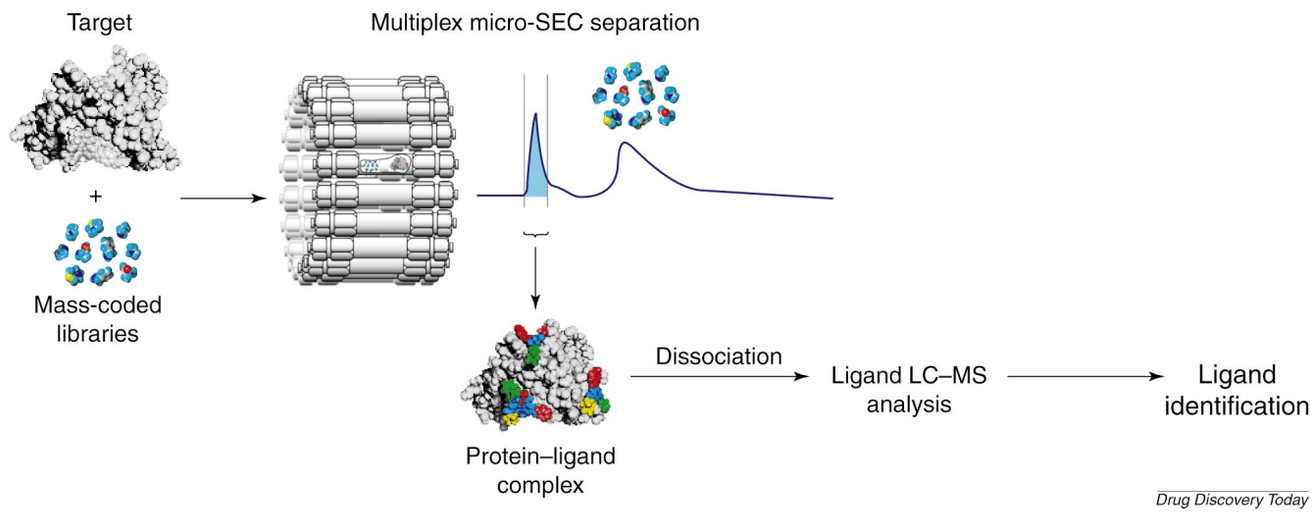

Figure 5. *NeoGenesis (Cambridge, MA, USA) integrated ultra-high-throughput small-molecule automated ligand identification system (ALIS). Combinatorial library mixtures are screened by size-exclusion chromatography coupled to reverse-phase HPLC and high-resolution electrospray mass spectrometry. Multi-thousand member library mixtures are designed to be self-encoded by the mass of each compound (mass coded), such that the mass of each ligand directly identifies its composition (building block plus core combination).*

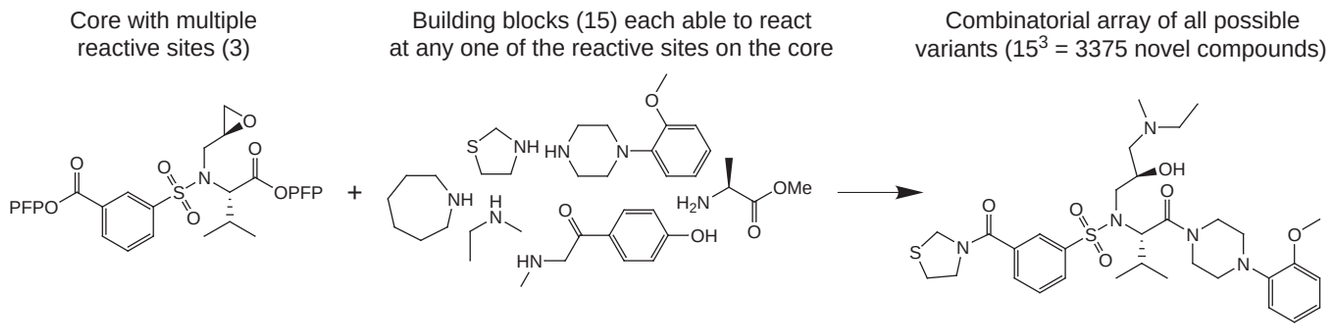

Figure 6. *NeoMorph (NeoGenesis, Cambridge, MA, USA) combinatorial libraries are synthesized as diverse, multi-thousand member mixtures. Each library mixture is mass coded, such that each core plus building block combination yields a unique mass, with a separation of at least 0.05 amu (atomic mass unit) between different combinations.*

combinatorial libraries prepared by solid-phase split-and-pool methods, as well as to single compounds in a screening deck. The MS data is analyzed in real time to identify chemical ligands. Structures of ligands that meet established criteria are then synthesized and confirmed as chemical ligands. In addition, the process enables quality control of the mixture libraries, and facilitates rapid optimization of affinity and additional pharmacokinetic parameters. Each platform can screen 300,000 compounds per day with minimal protein consumption, and has been used to find chemical ligand classes for many proteins, including enzyme and protein–protein interaction targets.

Capillary electrophoresis approach to chemical ligands

An alternative approach to discover chemical ligands, developed by Cetek (Marlborough, MA, USA), uses capillary electrophoresis (CE) to assess differences in mobility in an electric field between a free-tagged target protein and the same protein complexed with a ligand. The complex is then analyzed using fraction collection and LC/MS analysis, and deconvolution carried out using libraries of known compounds. However, with natural product extracts, this is only the first step in a challenging structural elucidation and synthesis problem.

Scanning calorimetry affinity approaches to chemical ligands

An interesting approach to chemical ligand discovery has been developed by 3-Dimensional Pharmaceuticals (Exton, PA, USA; Weiss, P.M., 3-Dimensional Pharmaceuticals, personal communication), which uses the stabilization of protein structure that occurs when it binds to a ligand (Fig. 7). Proteins, as with other polymeric materials, have differing degrees of organization or structure and, as a result, undergo an unfolding transition at a temperature (T_m) that is characteristic to each protein. As all these steps involve energy changes, the difference in the T_m will be indicative of the degree of stabilization by the ligand and simple thermodynamic calculations can then determine the affinity constant. The readout is fluorometric and occurs when the T_m is reached. The unfolding that occurs exposes lipophilic surfaces into which a fluorescent dye can dissolve and the emitted fluorescence detected using a charge-coupled device (CCD). The screening process is currently run in a 384-well plate format and on individual compounds, processing approximately 5000 compounds per week and consuming approximately 1 mg of target protein. The process can also be used for nucleic acids.

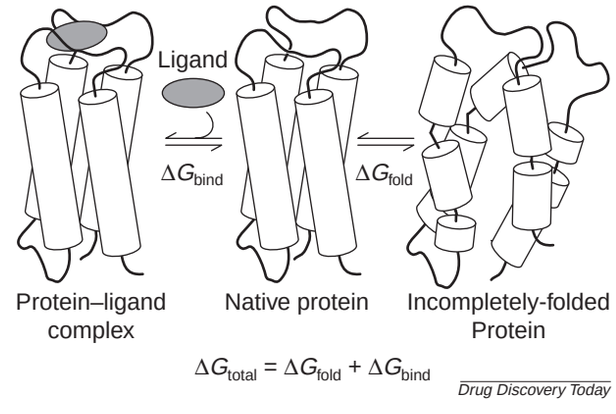

Figure 7. *Chemical ligand discovery through scanning microcalorimetry screens for the energetic differences between a target that is unfolded in the absence and presence of ligands. Abbreviations: ΔG_{bind}, change in Gibbs free energy on binding of the ligand; ΔG_{fold}, change in Gibbs free energy on folding of the protein.*

High-resolution NMR identification of chemical ligands

The use of NMR to study ligand–protein interactions is an established procedure, and has been used for screening and for structure–activity relationships (SAR) development of known ligands. To use this methodology, several requirements must be met:

- Large quantities of protein must be uniformly labeled with ^{15}N
- Both the protein and the ligands must possess sufficient solubility characteristics
- The protein must be sufficiently stable in solution and in the presence of significant quantities of dimethylsulfoxide (DMSO).

To enable the search for chemical ligands, all the ^{15}N resonances must be determined, which is not an inconsequential task. This has restricted the approach to small proteins or protein fragments (≈35 kDa). In addition, the methodology to assign ^{15}N resonances at the 100 kDa level does not currently exist[35]. Because of instrumentation limitations relating to signal resolution, only small mixtures of compounds (approximately ten) can be screened, placing significant restrictions on its use in getting from a gene to the screening stage. Recent improvements in cryogenic probe technology for NMR (Cryoprobe, Bruker Instruments, Billerica, MA, USA) have improved signal-to-noise ratios and enabled the use of mixtures of approximately 100 compounds per target. Assuming 100 runs per day per high field NMR instrument, approximately 10,000 compounds can be screened. Identification is by deconvolution, where sub-libraries are prepared and tested[36]. While this is an improvement, the technology still uses significant quantities of ^{15}N-labeled protein and all the other requirements already listed still apply.

Chemical ligands for RNA targets

An intriguing approach in getting from a gene to the screening stage is to move back one step to target the RNA rather than the proteins[37]. This was prompted by the realization that RNA possesses significant (three-dimensional) structure and that some natural product-based antibiotics interact selectively with bacterial RNA. By contrast to the helical structure of DNA, RNA can form three-dimensional shapes consisting of bulges, hairpins, loops and stems, which are essential for both recognition by proteins and for function. These are ordinarily found in RNA domains that regulate message stability and transport, splicing, or translational efficiency. These overall characteristic and distinct shapes provide the opportunity and ability for small molecules to specifically interact with a particular domain on an individual RNA.

The initial target class for this technology is antibacterials, as RNA structures have a high degree of conservation across species that are either absent or different in humans, allowing for potential broad-spectrum antibiotics. In addition, the macrolide and aminoglycoside families of antibiotics interact with domains on bacterial RNA.

The overall RNA drug discovery process begins with a bioinformatics exercise to recognize conserved domains across species, followed by a determination of whether an implicated domain possesses sufficient structure to serve as a target and whether there is a corresponding human equivalent. For instance, the aminoglycosides bind to a specific 27-mer domain of the *Escherichia coli* 16S ribosomal RNA but do not bind to the closely related human 18S site (Fig. 8). As the domain is small, the target(s) RNA is readily synthesized. The screening technology is based on Fourier Transform-ion cyclotron MS and affinity selection[38]. In this process, the target RNA is incubated with a chemical library for ligand affinity selection. The resultant non-covalently bonded complex, together with the free target, is analyzed by high-resolution MS and, as the mass of each library compound is known, the combined MW identifies the chemical ligands. In addition, MS–MS enables the determination of the binding domain of the ligand–target complex[39]. The screening assay can be multiplexed and it is estimated that up to ten RNA targets could be screened against libraries of up to 200 compounds per assay[40].

Scriptgen (Waltham, MA, USA) has developed an affinity screening system for nucleic acids called SCAN (screen for compounds with affinity for nucleic acids)[41]. As SCAN is an affinity-based approach, it can find ligands that bind to various sites on the nucleotide sequence of interest. The process can screen thousands of compounds per week in individual assays, and throughput can be increased by using small mixtures. Identification of the chemical ligand is then straightforward, as the structures of the compounds in the library are known.

Chemical diversity and chemical ligands

The reasons for, and the methods of, finding chemical ligands have already been discussed. However, the actual ability to find the ligands for all the possible binding sites on any protein is also very important, especially when little, if anything, is known about the structure of the target. The total compound library then has to include all potential binding surfaces with associated functionality on any protein. This means the libraries have to be molecularly

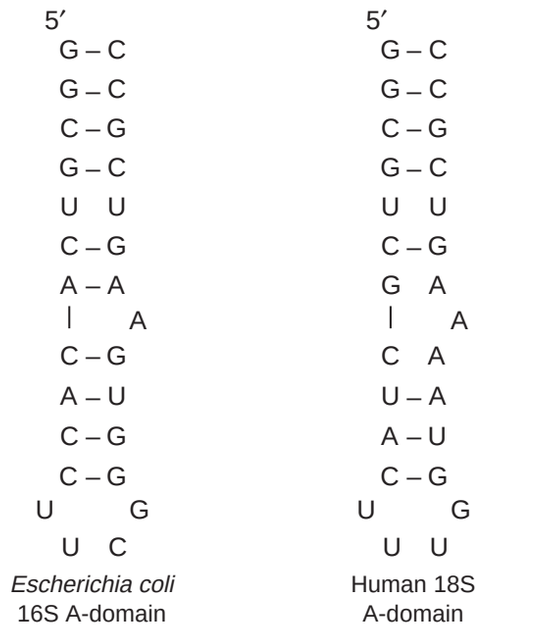

Figure 8. *A comparison of the bacterial (*Escherichia coli*) and human aminoglycoside binding domains on the target stem-loop RNA fragments. The differences between the two RNA targets are sufficient to yield specific small-molecule ligands.*

diverse.

Molecular diversity is both a word and a concept that is much discussed and usually ignored in practice. Diversity in combinatorial chemistry is a measure of the differences between library members, and is important as each compound represents the molecular complement of a potential binding site on a target. Therefore, increasing molecular diversity of a library increases the number of potential target surfaces for scanning and, thus, increases the probability of finding chemical ligands for a target. While chemical ligand generation is, perhaps, easier using targets that are well understood and characterized, it becomes crucially important when ligands are sought against a novel target. The lack of molecular diversity is probably a main reason for the bimodal distribution of hits seen in HTS campaigns.

Although a high-quality chemical library provides useful leads for most targets of interest, the question of how to prepare such a library is rarely addressed. As the number of potential targets increases dramatically as the results of the genomic initiatives become available, the requirement for diverse libraries to discover chemical ligands for these targets becomes imperative.

Molecular diversity among compounds in a library is based on descriptors familiar to all medicinal chemists. All the chemical software companies provide diversity modules as part of their overall cheminformatics packages. However, they all suffer from the same drawback: they provide only relative information about diversity and, thus, only return predictions that relate the diversity of one compound or compounds to an existing set of compounds. Although software packages can give valuable insights into the best method to screen an internal library or whether to purchase a new set of combinatorial compounds to supplement an existing database, they cannot answer more general questions concerning the absolute nature of diversity and how to fill the gaps. There is a necessity for a computational approach that is based on comprehensive molecular diversity and, by inference, is able to address all potential binding sites on a target, together with a method for systematically filling in the spaces to assure a thorough coverage[41].

Current concepts on diversity are based primarily on drugs found in the Current Medicinal Compounds or related databases. This is a biased diversity set[42,43]. Using current diversity for combinatorial library design and evaluation will continue to provide hits for the types of targets represented in these databases. However, this type of diversity is unlikely to provide reasonable chemical ligands for new types of genomic targets. This is supported by recent reports on a granulocyte-colony-stimulating factor (G-CSF) mimetic and an insulin receptor sensitizer, where the ligands look unlike anything to be expected in the screening libraries[44,45]. Effective screening libraries for finding chemical ligands will require a substantial increase in diversity while retaining descriptors for important parameters such as membrane permeability and other bioavailability parameters.

There is a real need for a new strategy that emphasizes the importance of approaching diversity from a biological perspective. This method would use a full set of theoretical protein surfaces for defining 'absolute' diversity, and would eliminate the inherent problems using current computational methods. The evolving treatment of diversity from an absolute, theoretical standpoint will provide a framework in which molecules are not initially examined relative to other compounds and protein surfaces are not examined relative to known proteins[41].

This type of diversity approach (termed quantized surface complementarity diversity) would enable:

- Absolute numerical prediction of diversity

- Clear prediction of which molecule sets would provide useful leads for most targets of interest (a 'universal' library in the absolute sense)
- Prediction of molecular binding to protein surfaces as yet unexplored (e.g. protein–protein interactions and novel genomic proteins)
- Tuneable resolution for hit optimization and differentiation between similar surfaces.

A true diversity approach would define the size and nature of the library necessary for efficient screening of a target protein at a predetermined molecular resolution, and would be an essential component of the most powerful ligand discovery technologies.

The search for chemical ligands is an integrated combination of compounds, diversity, screening and IT. A successful search would use carefully designed and diverse libraries of vast numbers of compounds, preferably in a mixture format, coupled with an ultra-HTS system. The IT would enable automation of the process and therefore dramatically increase efficiency, as well as generate databases that would enable data mining to increase efficiency overall. With intelligent design and execution, a chemical ligand approach to drug discovery would provide methods for efficiently evaluating the enormous range of potential targets provided by genomics and substantially decreasing the time taken to get from a gene to a drug.

Conclusion

There is great and justifiable optimism that genomics will provide new opportunities to identify the best drug discovery targets and, hence, the best small-molecule drugs for many major diseases. However, the extraordinary dividend that is predicted for post-genomic drug discovery will be realized by the companies that are the first to identify the best targets from the thousands of macromolecular targets. Conventional drug discovery can currently handle neither the nature nor the number of potential drug targets offered by genomics and, thus, pre-clinical drug discovery will have to be significantly reinvented to flourish in the post-genomic era. Genomics has changed the rules, and there is no going back. The post-genomic challenge for the pharmaceutical industry lies in finding the best targets in the fastest possible way without breaking the bank.

The most profound and fundamental advantage of ligand-directed drug discovery is its unique ability to provide the most simplifying and most valuable data set from genomics, namely, a validated drug target and its associated small-molecule drug lead. Simplicity cannot be underestimated, as many other genomic technologies convert a finite set of primary sequence information into ever expanding, multi-dimensional sets of secondary information. Many of these transcriptional, proteomal or computational data sets contain speculative or unknown data points and, in the worst case, these new technologies add complexity, cost, time and even error to the pre-clinical decision-making process. The ligand-directed discovery strategy features a much simpler decision path, and represents the shortest, straightest route from gene to lead. Properly implemented, it will make the pre-clinical discovery process simpler, cheaper and faster. This strategy is the most efficient mechanism for converting genomic information into an ever-expanding development pipeline and, perhaps, is the only strategy that offers the potential of increasing the scale at which pre-clinical drug discovery can be initiated while reducing the cost per target.

REFERENCES

1 Guild, B.C. (1999) Genomics, target selection, validation, and assay considerations in the development of antibacterial screens. *Annu. Rep. Med. Chem.* 34, 227–236
2 Duggan, D.J. *et al.* (1998) Expression profiling using cDNA arrays. *Nat. Genet.* 21 (Suppl. 1), 10–14
3 Lottspeich, F. (1999) Proteome analysis: A pathway to the functional analysis of proteins. *Angew. Chem., Int. Ed. Engl.* 38, 2476–2492
4 Ackerly, B.J. *et al.* (1998) Systematic identification of essential genes by *in vitro* mariner mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8927–8932
5 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
6 Rychlewski, L. *et al.* (1999) Functional insights from structural predictions: Analysis of the *Escherichia coli* genome. *Protein Sci.* 8, 614–624
7 Pawlowski, K. *et al.* (1999) The *Helicobacter pylori* genome: From sequence analysis to structural and functional predictions. *Proteins* 36, 20–30
8 Šali, A. (1998) Arise, go forth, and solve structures. *Nat. Struct. Biol.* 5, 1029–1032
9 Shortle, D. (1999) Structure prediction: The state of the art. *Curr. Biol.* 9, R205–R209
10 Sternberg, M.J.E. *et al.* (1999) Progress in protein structure prediction: Assessment of CASP3. *Curr. Opin. Struct. Biol.* 9, 368–373
11 Zhang, B. *et al.* (1999) From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* 8, 1104–1115
12 Wei, L. *et al.* (1999) Are predicted structures good enough to preserve

functional sites. *Structure* 7, 643–650
13 Park, J. *et al.* (1998) Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods. *J. Mol. Biol.* 284, 1201–1210
14 Brenner, S.E. *et al.* (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073–6078
15 Mendelsohn, A.R. and Brent, R. (1999) Protein interaction methods – towards an end game. *Science* 284, 1948–1950
16 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753
17 Wrighton, N.C. *et al.* (1996) Small peptides as potent mimetics of the protein hormone erythropoietin. *Science* 273, 458–463
18 Chirinos-Rojas, C.L. *et al.* (1998) A peptidomimetic antagonist of TNF-alpha-mediated cytotoxicity identified from a phage-displayed random peptide library. *J. Immunol.* 161, 5621–5626
19 Schatz, P.J. (1994) Construction and screening of biological peptide libraries. *Curr. Opin. Biotechnol.* 5, 487–494
20 Norman, T.C. *et al.* (1999) Genetic selection of peptide inhibitors of biological pathways. *Science* 285, 591–595
21 Eliseev, A.V. (1998) Emerging approaches to target-assisted screening of combinatorial mixtures. *Curr. Opin. Drug Dis. Dev.* 1, 106–115
22 Tan, D.S. *et al.* (1999) Synthesis and preliminary evaluation of a library of polycyclic small molecules for use in chemical genetic assays. *J. Amer. Chem. Soc.* 121, 9073–9087
23 Tan, D.S. *et al.* (1998) Stereoselective synthesis of over two million compounds having structural features both reminiscent of natural products and compatible with miniaturized cell-based assays. *J. Amer. Chem. Soc.* 120, 8565–8566
24 Lenz, G.R. (1998) Optimizing small-molecule drug targets: Focus on combinatorial chemistry. *Spectrum Reports: Drug Discovery and Design Decision Resources* 16, 1–15
25 Nestler, H.P. *et al.* (1994) A general method for molecular tagging of encoded combinatorial libraries. *J. Org. Chem.* 59, 4723–4724
26 Borchardt, A. *et al.* (1997) Small-molecule dependent genetic selection in stochastic nanodroplets as a means of detecting protein–ligand interactions on a large scale. *Chem. Biol.* 4, 961–968
27 You, A.J. *et al.* (1997) A miniaturized arrayed assay format for detecting small molecule–protein interactions in cells. *Chem. Biol.* 4, 969–975
28 MacBeath, G. *et al.* (1999) Printing small molecules as microarrays and detecting protein–ligand interactions *en masse*. *J. Amer. Chem. Soc.* 121, 7967–7968
29 Jindal, S. and Lenz, G.R. (1998) Affinity selection: An emerging technology for drug discovery. *Spectrum Reports: Drug Discovery and Design Decision Resources* 20, 1–13
30 Kay, B.K. *et al.* (1998) From peptides to drugs via phage display. *Drug Discovery Today* 3, 370–378
31 Bowie, J.U. and Pakula, A.A. (1996) Scriptgen Pharmaceuticals screening method for identifying ligands for target proteins. US 5585277
32 Kaur, S. *et al.* (1997) Affinity selection and mass spectrometry-based strategies to identify lead compounds in combinatorial libraries. *J. Protein. Chem.* 16, 505–511
33 Carell, T. *et al.* (1994) A novel procedure for the synthesis of libraries containing small organic molecules. *Angew. Chem., Int. Ed. Engl.* 33, 2059–2061
34 Dunayevskiy, Y.M. *et al.* (1996) Application of capillary electrophoresis-electrospray ionization mass spectrometry in the determination of molecular diversity. *Proc. Natl. Acad. Sci. U. S. A.* 93, 6152–6157
35 Shapiro, M.J. and Wareing, J.R. (1999) High resolution NMR for screening ligand/protein binding. *Curr. Opin. Drug Dis. Dev.* 2, 396–400
36 Hajduk, P.J. *et al.* (1999) High-throughput nuclear magnetic resonance-based screening. *J. Med. Chem.* 42, 2315–2317
37 Ecker, D.J. and Griffey, R.H. (1999) RNA as a small-molecule drug target: Doubling the value of genomics. *Drug Discovery Today* 4, 420–429
38 Griffey, R.H. *et al.* (1999) Determinants of aminoglycoside-binding specificity for rRNA by using mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10129–10133
39 Griffey, R.H. *et al.* (1999) Targeted site-specific gas-phase cleavage of oligoribonucleotides. Application in mass spectrometry-based identification of ligand binding sites. *J. Amer. Chem. Soc.* 121, 474–475
40 Hofstadler, S.A. *et al.* (1999) Multiplexed screening of neutral mass-tagged RNA targets against ligand libraries with electrospray ionization FTICR MS: A paradigm for high-throughput affinity screening. *Anal. Chem.* 71, 3436–3440
41 Wintner, E.A. and Moallemi, C.C. Quantized surface complementarity diversity: A model based on small molecule-target complementarity. *J. Med. Chem.* (in press)
42 Drews, J. (1997) Strategic choices facing the pharmaceutical industry: A case for innovation. *Drug Discovery Today* 2, 72–78
43 Drews, J. (1996) Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* 14, 1516–1518
44 Tian, S.S. *et al.* (1998) A small, nonpeptidyl mimic of granulocyte-colony-stimulating factor. *Science* 281, 257–259
45 Zhang, B. *et al.* (1999) Discovery of a small molecule insulin mimetic with antidiabetic activity in mice. *Science* 284, 974–977

Collaboration...

The Scripps Research Institute (TSRI; La Jolla, CA, USA) have announced a functional genomics research collaboration with **Lexicon Genetics** (The Woodlands, TX, USA). This will involve the use of Lexicon's proprietary homologous recombination technology to generate knockout mice for a gene identified by TSRI, and will be followed by collaborative research to define the function of the gene and its potential role in human disease. The mice are generated by recombinase technology and can contain a point mutation in the gene of interest. Determination of the function of the gene in disease and the validation of gene products as drug targets are then carried out using an *in vivo* mammalian system.