Drugs of the Future 2002, 27(12): 1165-1180

CCC: 0377-8282/2002

Review Article

Small molecule lead generation processes for drug discovery

Robert Goodnow, Jr. Ph.D.
New Leads Chemistry Initiative, Hoffmann-La Roche, Inc.,
340 Kingsland Street, Nutley, New Jersey 07110-1199, USA

CONTENTS

Abstract 1165
Introduction 1165
General strategies to generate small molecule leads 1167
Definition of quality leads: “The value of a good lead cannot be overstated” 1167
Drug-likeness *versus* lead-likeness for *de novo* lead generation 1168
Solubility as a principal component of ADME properties 1168
The hit-to-lead process for lead generation 1169
Design concepts for evolving combinatorial chemistry as a lead generation platform lead: library targeting and size 1169
Guiding template selection and library design by computational methods 1170
Appropriate library size 1172
Leveraging the advantages of high-throughput chemistry for building a lead generation collection 1173
Selected examples of combinatorial chemistry-based lead generation 1174
An outlook for lead generation: the chemical genomics concept 1176
Acknowledgements 1179
References 1179

Abstract

In response to the growing economic pressure to accelerate drug discovery processes, there have been numerous reports of attempts to improve the quality of leads through computational methods and combinatorial chemistry. With the reports of drug-likeness concepts and metrics, there has been a substantial increase in awareness of the types of molecules that are likely to be useful leads for drug discovery. The concepts of greater drug-likeness are used routinely to refine compound library designs for lead generation. Challenges still remain to develop chemistry amenable to the introduction of multiple diversity components while still creating molecules of lead-like and drug-like properties. Despite the common capability to synthesize and screen thousands of molecules, the current focus has shifted to smaller libraries targeted with greater computational sophistication according to putatively smarter target-drug premises. The need to target the right drug-like lead generation library will only grow in importance in order to exploit the information advantage of the chemical genomics approach. Organizations that develop efficient lead generation capabilities will likely find increased competitive advantage in the drug discovery field.

Introduction

The modern drug discovery process requires more “high quality” compounds to be assayed against multiple biological targets. Drug discovery can be depicted as a multistage process toward decisive milestone objectives (Fig. 1). Major milestones include the formulation of a new medicine proposal, the conceptual linking of a biological target with a disease indication, followed by efforts to validate this concept. Simultaneously, efforts are made to develop a high-throughput screening (HTS) assay that will permit rapid screening of compound collections. Assuming that a proper set of compounds is available for screening, a series of molecules having some preliminary indication of activity should be identified as hits. With a hit series in hand, the “hit-to-lead” process is pursued to understand which of the compounds identified by HTS will likely become promising lead compounds. This progress refines “hit” structures before beginning focused and resource-intensive medicinal chemistry-driven lead optimization toward a clinical candidate selection (CCS) milestone. Subsequently, *in vivo* efficacy and toxicity studies with animal models determine whether the entry into human (EIH) milestone will be realized.

The initiation of EIH studies is a major transition in the development process to transform the biologically active chemical entity into a drug. Final stages consist of full clinical development that may demonstrate the efficacy and safety to justify regulatory approval and ultimately market launch. Unfortunately, the majority of compounds moving along this pathway fail to achieve all of these milestones for various reasons (*vide infra*). In planning for this attrition rate, pharmaceutical organizations must also consider that each new drug requires an average of 10-15 years to develop and an expenditure of hundreds of millions of US dollars. A recent study estimates that the average cost to discover, develop and launch a new drug

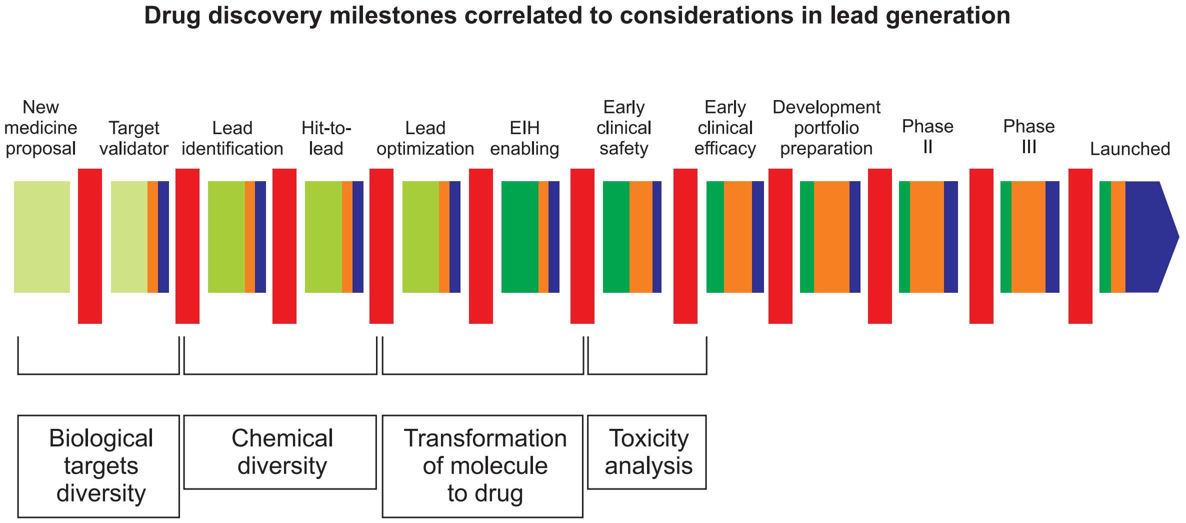


Fig. 1. The drug discovery pathway is a coordinated and sequential series of critical milestones through which biological targets and molecules active against those targets must pass.

has risen to approximately 800 million dollars (1, 2). Increases in these costs are due to longer and more complex clinical trials. Consequently, the pharmaceutical industry averages 0.5 major new drugs per company per year despite carrying an increasingly broad portfolio of therapeutic discovery projects. Study of historical financial growth rates within the pharmaceutical industry indicates that growth cannot be maintained if the overall average of 0.5 new drugs registered per annum per company is not increased (3).

While challenges and competition in drug discovery persist, recent completion of the human genome sequencing has provided a growing number of potential therapeutic targets. For example, from the sequence data of the human genome, it was estimated that there are encoded approximately 1000 G protein-coupled receptors (GPCR) and approximately 800 kinases (4). Continuing analysis of human genome sequence data has resulted in a reduction in the numbers of new targets; there may be only approximately 600 G protein-couple receptors for example (5). For these target classes alone, approximately only 10-20% of the proteins have been functionally characterized. The desire within the industry to functionally characterize the many newly identified genes for their therapeutic potential has led to an explosive growth in proteomics, the study of local and temporal protein products from gene expression. With the possibility of greater numbers of new therapeutic targets, pharmaceutical companies are searching for improved efficiency in all stages of drug discovery.

Given the costly and long-term nature of the process, the traditional drug discovery process has been characterized as having several liabilities. In many cases, the disease relevance of biological targets is not unequivocally defined until clinical trials, the point at which most new drugs fail. In contrast, the increasingly demanding health care market environment creates an evolutionary pressure for pharmaceutical companies to develop new practices and strategies to enhance success rates. Such strategies include the use of genomics, bioinformatics, improved computational ADME (absorption, distribution, metabolism and excretion) predictions and combinatorial chemistry as a means to improve compound design and the rate of compound synthesis. New strategies also include the use of small molecules with sufficient potencies and drug-likeness to validate a new target thus avoiding the investment of time and money on faulty therapeutic concepts. In order to tackle the liabilities associated with many small molecules and to provide more drug-like structures, increased attention to the physical properties of compounds as potential drugs is occurring earlier in the discovery process. The rationale is simple: *by initiating the lead optimization stage with compounds possessing greater drug-like character against therapeutic target proteins which have been validated with a small molecule, the chances of successful drug discovery are increased.* Such lead generation processes must take into account and integrate with the upstream and downstream elements in the drug discovery process. In this environment of evolutionary pressure, the primacy of efficient processes for the generation of good small molecule leads becomes evident.

Table I: Several general strategies to identify small molecule leads.

	Potential advantages	Potential liabilities
Natural products	Diverse Often soluble structures Proven source of drugs	Synthetically complex Small quantities
HTS of random compound collections	Diversity space coverage Many drug-like structures	Many non-drug-like structures Sometimes costly
Following known drugs	Starting with a valid target-drug concept Starting with drug-like structure Potentially few changes necessary to create a drug Often successful	Limitations of intellectual property Not applicable to targets without a precedent
Target-focused lead generation libraries	Potentially drug-like Focused diversity Potential for novel structures Chemical space coverage	Requires HTC infrastructure/platform Potential for high MW, clogP compounds
Structure-based drug design	Rationally based on biostructure information	Requires target structure for target Limited history of success

General strategies to generate small molecule leads

Lead identification and lead generation in modern drug discovery relies to a great extent on finding leads from among several major strategies or sources including HTS of random and focused compound collections, natural products-based leads, structures which follow known drugs, structure-based drug design and design and synthesis of targeted compound libraries (Table I). Currently, many organizations are increasingly focusing on rationally targeted lead generation efforts. Like most lead generation strategies, success in designing targeted lead generation libraries depends to a great extent on the same concepts that make for success in the HTS-based approach. Thus, successful lead identification practices for HTS-based lead generation establish precedence directly applicable to other targeted lead generation approaches.

Substantial developments in assay biology and technology have made even complex, multicomponent bioassays amenable to the HTS approach (6), which is the mainstay of lead generation in modern drug discovery. With the increasing capacity of HTS technology, drug discovery organizations are creating collections composed of millions of compounds. These collections of small molecules with diverse structural properties have been acquired by various means including compound archives of previous lead optimization efforts, purchases from compound vendors and company mergers. Recently, many drug discovery companies have been increasing the size and quality of their compound collections through internal and/or outsourcing efforts of targeted combinatorial chemistry. In addition to increasing the size of a compound collection, there are increased quality expectations for high compound purity and for library design ideas.

While HTS of large, random collections is often the principal strategy for lead generation, many groups are experimenting with alternative strategies to focus the screening of compounds with smaller sets of compounds based on assumptions of what the active structures is likely to be. No matter which approach is taken to identify the primary hits, it is crucial to identify molecules with properties that are amenable to transformation into a drug molecule.

Definition of quality leads: “The value of a good lead cannot be overstated”

A properly constructed and well-managed compound inventory composed of quality molecules is an enduring treasure for any pharmaceutical company. Parameters calculated from molecules in the World Drug Index were used by Lipinski to define upper limits for values that give a qualitative measure for drug-likeness of a given molecule (7, 8). A drug-like molecule is, among other things, one that is likely to have both good absorption and cell permeation properties.

Increasing violations of these rules decreases the likelihood that a molecule in question will have drug-like absorption or cell permeation properties. This concept of drug-likeness is known as the Rule of Five (9) (Fig. 2). These 4 rules define the upper limits of easily calculable metrics; the upper limits are also multiples of the number 5 and thus make for a useful mnemonic. The Rule of Five is a common and convenient way to characterize molecules in terms of their potential pharmaceutical utility. Indeed, there are exceptions to these simple rules (particularly molecules that are substrates of cellular transporter proteins), but the majority of successful drugs are within these boundaries. Despite the evolution of more computationally sophisticated methods (10) to predict small molecule drug-likeness, none have exceeded the practical simplicity nor have been as widely employed as the Lipinski Rule of Five.

Lipinski's Rule of Five: Assessing Violations

- Molecular Weight >500 DA
- ClogP >5 (or Moriguchi logP > 4.15)
- Number of hydrogen bond donors >5
- Number of H-bond acceptors (o, N) >10

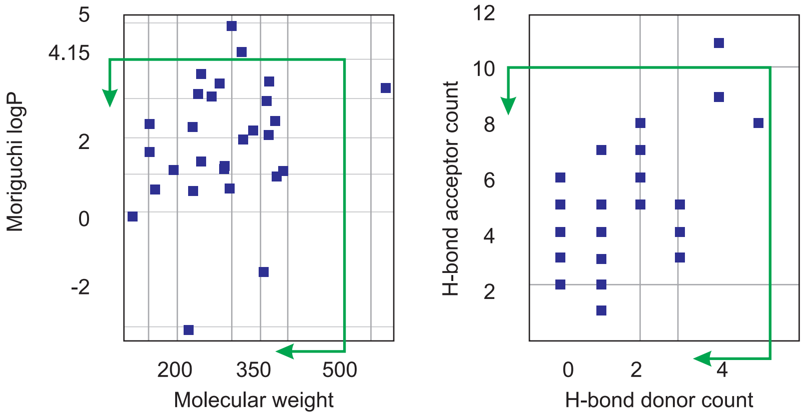


Fig. 2. Lipinski's Rule of Five: Definition of simple metrics for identifying compounds which are likely to have good absorption and cell permeation. The distribution of Rule of Five calculated values are shown for 25 recently launched drugs (9).

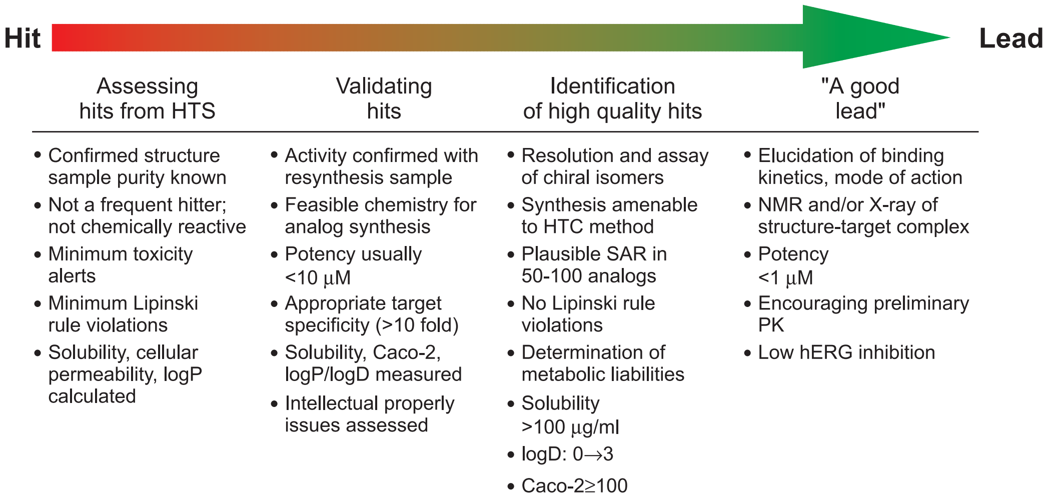
Others have built upon the Rule-of-Five concept noting that some compounds with molecular weights greater than 500 Da and with decreased molecular flexibility of low to moderate solvent-exposed, polar surface area may have acceptable cellular penetration properties. This should not be construed as saying that rigid, heavy compounds acceptable because they may be absorbed. Rather, this observation highlights a strategy to design and optimize unavoidably large molecules active for a particular target (11). Further, some have attempted to predict other drug-like features such as pharmacological and pharmacokinetic properties, thereby complementing the Lipinski Rule-of-Five (12).

Drug-likeness versus lead-likeness for de novo *lead generation*

Analysis of multiple lead optimization efforts has highlighted the value of creating small molecule libraries with lead-like properties (13, 14) (*i.e.*, molecular weights < 350 Da, clogP < 3, acceptable solubilities, *etc.*). Several studies have shown that compounds of moderate potency for a particular target with lead-like properties move more quickly through the lead optimization process. This is in contrast to following a known drug closely; rather the lead-like concept applies particularly to *de novo* lead generation design strategies. Lead-like properties are particularly important to consider when initiating fragment-screening approaches to lead generation (15). In such efforts, small fragments that show even modest signs of binding affinity to a target are used as seeds to grow more complex structures. Traditional efforts to increase a compound's biological activity and selectivity often result in greater molecular complexity (16), molecular weight and compound lipophilicity, often resulting in less attractive ADME properties. Compounds with poor ADME properties may require special formulation for development into marketable drugs. A major hurdle for many small molecules to become drugs is often attributed to poor ADME properties in addition to lack of efficacy (17). Therefore, it stands to reason, that starting with smaller fragments allows for a greater number of options for increasing potency and selectivity through increases in molecular size and complexity.

Solubility as a principal component of ADME properties

The solubility of a compound has central importance with respect to ADME properties. Molecules with an aqueous solubility of less than 10 µg/ml are likely to have less than ideal ADME properties and require significant effort with special formulation procedures. There are several solubility prediction methods reported (18). However, for the complex phenomenon of solubility, predicted solubility results must be interpreted with caution; for example, the solubility of ionizable molecules varies dramatically whether the molecule is in its salt or free form. Solubility prediction algorithms are perhaps most useful for sorting molecules into broader categories such as “soluble”, “somewhat soluble” and “insoluble”. Interpreted in this way, solubility prediction methods are useful in the hit-to-lead phase and in library design to focus attention towards more soluble molecules.



Assessing hits from HTS	Validating hits	Identification of high quality hits	"A good lead"
• Confirmed structure sample purity known • Not a frequent hitter; not chemically reactive • Minimum toxicity alerts • Minimum Lipinski rule violations • Solubility, cellular permeability, logP calculated	• Activity confirmed with resynthesis sample • Feasible chemistry for analog synthesis • Potency usually <10 μM • Appropriate target specificity (>10 fold) • Solubility, Caco-2, logP/logD measured • Intellectual properly issues assessed	• Resolution and assay of chiral isomers • Synthesis amenable to HTC method • Plausible SAR in 50-100 analogs • No Lipinski rule violations • Determination of metabolic liabilities • Solubility >100 μg/ml • logD: 0→3 • Caco-2≥100	• Elucidation of binding kinetics, mode of action • NMR and/or X-ray of structure-target complex • Potency <1 μM • Encouraging preliminary PK • Low hERG inhibition

Fig. 3. The hit-to-lead process accumulates information with respect to molecules identified as primary hits which can lead to the identification of a high quality lead structure that is ready for rapid and efficient lead optimization. Although the exact criteria for a lead molecular series are specific to the target and project, certain features that will facilitate the identification of a high quality lead can be considered generally.

The hit-to-lead process for lead generation

Depending on the biological target, the assay of thousands of single compounds usually produces a number of primary hits. Primary hits are compounds for which there are limited data associated with an activity threshold and further validation is necessary to confirm the biological activity associated with such compounds. Routinely there is substantial attrition in the number of compounds that progress from primary hit to validated hit status. The process of primary hit filtration and assessment is part of a process commonly known as hit-to-lead (Fig. 3) (19). Experience has shown that of the many primary hits identified from a high-throughput screen, only a handful are likely to advance to the resource-intensive lead optimization phase. Of the reasons that a compound may fail during lead optimization, several may be generalized and may be used as a basis for consideration of the suitability of any molecule for a target. For example, recent reports have described methods to identify so called frequent hitters, molecules that appear active in several HTS assays unrelated as members of a single protein family (20). Because attempts to optimize frequent hitters as leads are nearly always futile, these molecules should be eliminated rapidly from consideration.

For a given molecule, failure to meet some hit-to-lead criteria does not necessarily eliminate the molecule from further consideration as much as it helps to direct focus on synthesis of other molecules to remedied the problem. Further, the exact criteria for a lead molecular series are specific to the target and project, but can highlight features that will often facilitate the identification of a high quality lead. Successful practices for the identification of good leads from HTS data are an efficient means to increase the chances for success in lead optimization. As encouraging information is accumulated, the molecule or series of molecules in question progresses from the status of an HTS primary hit to *bona fide* lead.

Design concepts for evolving combinatorial chemistry as a lead generation platform lead: library targeting and size

Although HTS has provided a great advantage in lead generation, many targets still require additional strategies to bring forward high quality leads. For example, the concept of a focused and targeted approach to lead generation is appealing. This strategy involves the design, synthesis and assay of targeted small molecule libraries. Although there may be significant information about the target, a library approach is often still required to exemplify the library design concept with a sufficient number of representative compounds. The type of information that can be encoded in a library are varied and diverse (Table II).

Early efforts in combinatorial chemistry promoted the concept that large numbers of compounds obviated the need for compound-by-compound design considerations. Although serendipity is an important advantage in the strategy of applying high-throughput chemistry for lead generation, increasingly, libraries are designed with a particular targeting focus. Alternatively, libraries have been targeted more generally around so called “privileged structures” (21) or motifs frequently associated with protein families such as G protein-coupled receptors, kinases and proteases. The computational association of particular library design core structures to targets can take at least two directions. First, where there is biostructural information, it is possible to test library design ideas by

Table II: General aspects incorporated in library design and corresponding factors that limit those aspects.

Library design aspect	Limitations to these features
Selection of template/core	Feasibility of chemistry for HTC methods
R group selection	Reagent availability
Diversity coverage	What chemical space density is necessary around a particular core to find a hit is unknown
Drug-likeness	Condensation of multiple diversity reagents often results in heavy and lipophilic molecules
Fitness to target	Accuracy of the design premise and computational metrics
SAR loading - increasing the density of information	Balancing chemistry space coverage: diversity *versus* density
Large library size	Complexity of synthesis process and assay of larger libraries

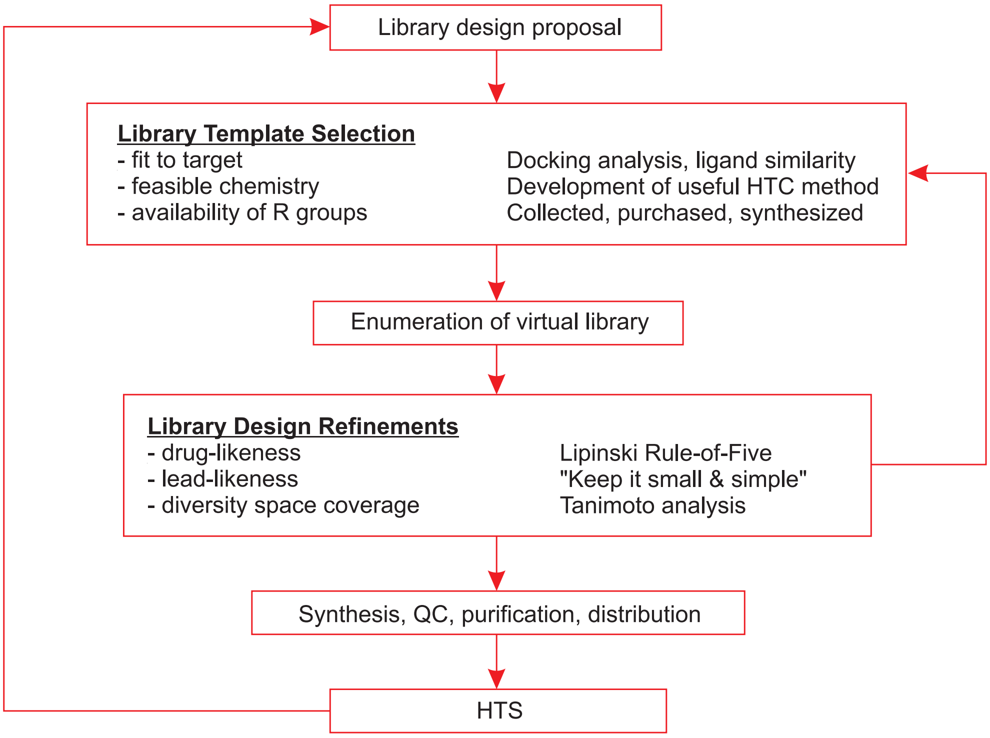

Library design proposal
Library Template Selection
- fit to target
- feasible chemistry
- availability of R groups
Docking analysis, ligand similarity
Development of useful HTC method
Collected, purchased, synthesized
Enumeration of virtual library
Library Design Refinements
- drug-likeness
- lead-likeness
- diversity space coverage
Lipinski Rule-of-Five
"Keep it small & simple"
Tanimoto analysis
Synthesis, QC, purification, distribution
HTS

Fig. 4. Iterative library design process for new lead generation involves the creation of virtual chemical structures and the analysis of fitness of these structures to a particular target, for their drug-likeness and lead-likeness.

in silico docking experiments. The structures of a virtual library, compounds yet to exist other than in a computer database, can be docked automatically into a particular target structure and evaluated for fitness to that target. It is assumed that a good fit in a computational simulation will correlate to an active compound in reality.

Another approach is required when the structure of the biological target is not known; one must base library design on similarity of ligands that are known to bind to the target in question.

While targeting library designs to particular target may increase the chances of finding a potent hit, it is critical to simultaneously refine the library design for optimal drug-like properties. In this way, the potent hits that may be created will have the right properties for efficient optimization to a drug molecules. This balance is best achieved in an iterative fashion prior to compound synthesis (Fig. 4).

Guiding template selection and library design by computational methods

The design, synthesis and assay of large compound libraries requires substantial planning, infrastructure, time

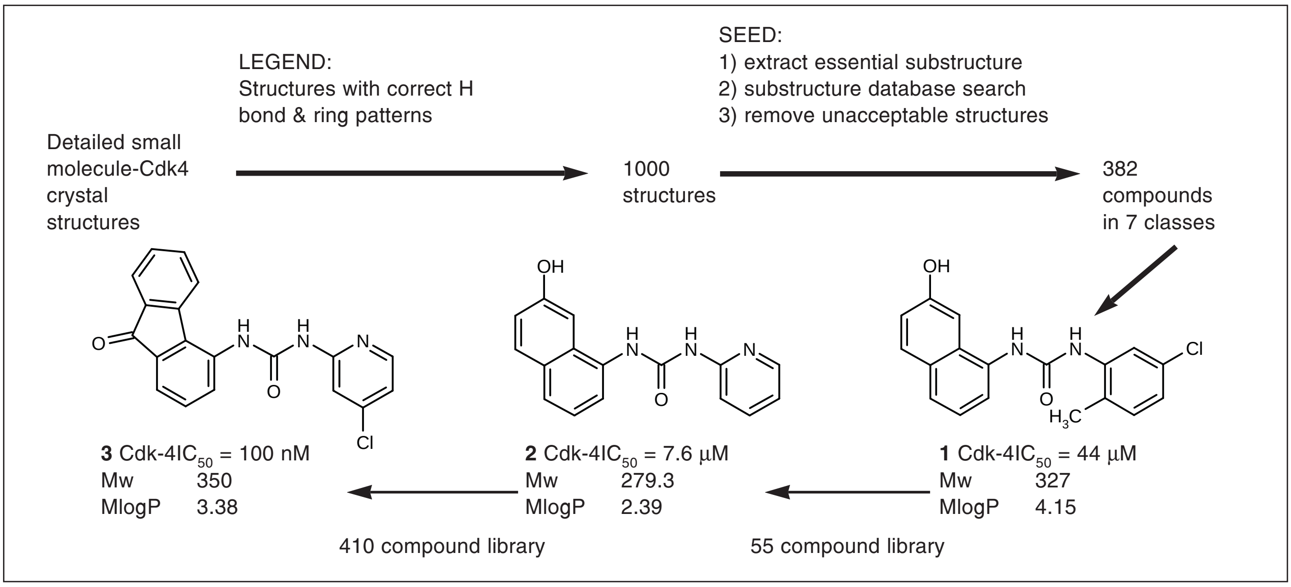

LEGEND:
Structures with correct H bond & ring patterns
SEED:
1) extract essential substructure
2) substructure database search
3) remove unacceptable structures
Detailed small molecule-Cdk4 crystal structures
1000 structures
382 compounds in 7 classes
O
H
N
H
N
N
O
Cl
OH
H
N
H
N
N
O
OH
H
N
H
N
Cl
O
H_3C
3 Cdk-4IC$_{50}$ = 100 nM
Mw 350
MlogP 3.38
2 Cdk-4IC$_{50}$ = 7.6 μM
Mw 279.3
MlogP 2.39
1 Cdk-4IC$_{50}$ = 44 μM
Mw 327
MlogP 4.15
410 compound library
55 compound library

Fig. 5. Structure-based generation of novel Cdk4 inhibitors.

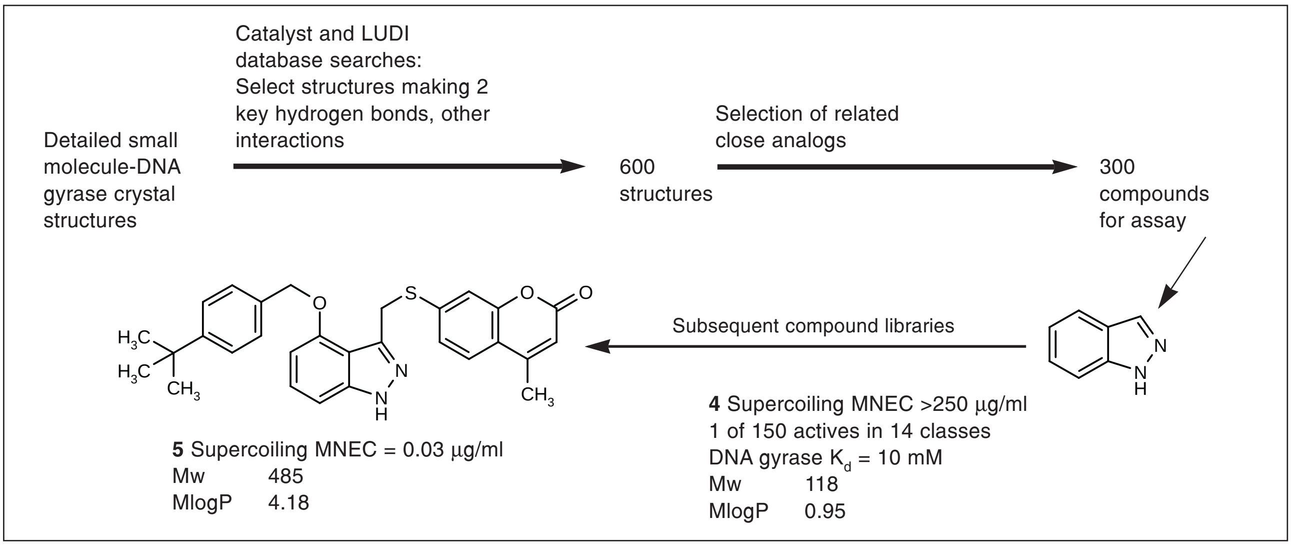
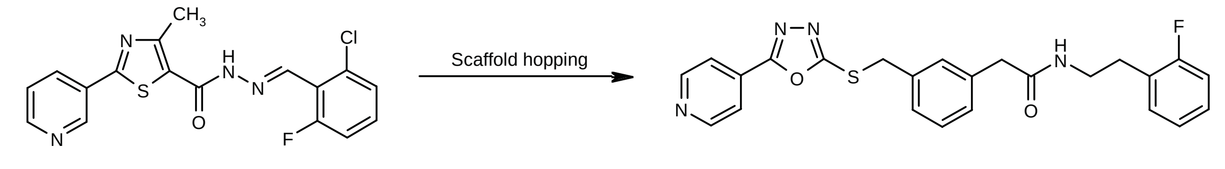

Catalyst and LUDI database searches:
Select structures making 2 key hydrogen bonds, other interactions
Selection of related close analogs
Detailed small molecule-DNA gyrase crystal structures
600 structures
300 compounds for assay
H_3C
H_3C
CH_3
O
S
O
O
N
N
H
CH_3
Subsequent compound libraries
N
N
H
5 Supercoiling MNEC = 0.03 μg/ml
Mw 485
MlogP 4.18
4 Supercoiling MNEC >250 μg/ml
1 of 150 actives in 14 classes
DNA gyrase K$_d$ = 10 mM
Mw 118
MlogP 0.95

Fig. 6. Novel inhibitors of DNA gyrase identified by biased needle screening. Assay results are expressed as the maximal noneffective concentration (MNEC) for a DNA gyrase inhibition assay.

and effort. Consequently, it is desirable to use computational chemistry to increase the targeting precision of lead generation efforts with small library designs. This strategy was well exemplified by the publication of scientists at Merck-Banyu with their discovery of Cdk-4 ligands (Fig. 5) (22). Based on several crystal structures of small molecules bound to Cdk2, important small molecule-protein interactions were identified for a Cdk4 homology model. Using proprietary software (LEGEND), Banyu scientists were able to identify 1000 candidate structures. With a second proprietary software application (SEED), critical substructures were extracted and used for a substructure-based search of the Available Chemical Directory (ACD). Filtration of these virtual hits led to the identification of seven structural classes of small molecules. A member of one class was shown to have an inhibitory potency of 7 μM. Elaboration of the hit resulted in the discovery of compounds with nanomolar inhibitory potency.

Screening for similarity of fragments or substructures within ligands, whether computationally or actual, is known as needle screening. A successful example of computationally identified starting points was reported by Roche scientists (Fig. 6) (23). Crystal structures of several natural products bound to DNA gyrase formed the basis of substantial SAR information for molecular



6 GSK-3 HTS hit
Mw 375
MlogP 2.78

7 GSK-3 IC$_{50}$ = 1.2 μM
Mw 448
MlogP 3.45

Fig. 7. Scaffold hopping for GSK-3 HTS hits.

interactions affecting enzyme inhibition. A virtual screen of compound databases using the LUDI (24) and CATALYST (25) programs resulted in the selection of some 6000 compounds sorted into 14 classes. Of these classes, indazole was identified as a millimolar hit. Further substitution of this heterocyclic core resulted in compounds active in the nanomolar range as inhibitors of DNA gyrase.

While it is important to balance the efforts of new lead generation with the potential advantages of starting library designs favoring privileged structures, it is equally important to consider the generation of *de novo* chemical structures. The concept of *de novo* ligand design is particularly important for “template hopping”, an attempt to design a different core structure with similar chemical and biological properties (26). Template hopping is particularly advantageous when looking to improve limitations (physical chemical properties, patent limitations, *etc.*) inherent in a chemical series with an different system. It is an obvious strategy to prepare large virtual libraries of compounds based on multiple core structures which are synthetically feasible using high-throughput synthesis methods. Various methods of virtual screening are then used to identify the relative fitness of a particular set of library compounds. Scientists at Novo Nordisk have reported a method for template hopping to identify selective glycogen synthase kinase 3 inhibitors (Fig. 7) (27). In this example, the stability of the hydrazone hit was a concern and a viable replacement was required. Virtual library compounds based on several templates were evaluated with proprietary CATS2 software for similarity to GSK-3 hits identified by an HTS campaign. A relatively higher virtual hit rate was detected for several library designs; these designs were selected for the synthesis of compound libraries. From these libraries, several compounds with lower micromolar and submicromolar inhibitory potency against GKS-3 were detected.

Other strategies for template selection in library design are reported involving the fragmentation of known drugs and the incorporation of fragments that confer a greater likelihood of drug action in new structures (28). Other efforts have reported the use of chemistry under extreme conditions to generate structures from unusual chemical pathways (29). Despite their complexity, natural products have been one of the best sources for diverse structures. Natural products are a proven source of many active hits for new lead generation and drug discovery. However, the practical difficulties with available quantities, structural characterization and resynthesis often limit the scaling of this source of lead generation. Perhaps ways to incorporate the diverse structures into synthetically convenient motifs will enhance natural products as a source of *de novo* ligand design.

Given limitations of time and expense for generating new lead structures, coupling library design with methods for virtual or *in silico* HTS have become commonplace (30). While rationally designing libraries around structural motifs previously shown to be active, it is important to include the development of new chemistries and new motifs. A balance between focused and protein-family targeted library design and new, chemistry-driven design is likely to ensure the discovery of useful and novel chemical entities. In response to these trends, many commercial compound library providers are increasing efforts to design libraries around protein-family targeted themes.

Appropriate library size

The necessary size of a compound collection presents an interesting theoretical and scientific problem. The number of structures included in the design of a focused library is the result of balancing several critical aspects such as chemical tractability, diversity coverage and design density. Design density refers to the number of molecules that are subtly different in substitution around a particular template or structural motif. For an active series, design density will provide very useful SAR information. There are reports that the presence or absence of a single methyl group will make the difference between an active and inactive hit structure; in such cases, it may be that critical functional groups must be viewed as parts of the core and not as diversity appendages. When balancing the density of diversity coverage by a particular library, one needs to focus on the correct fit of innovation, technology and feasible high-throughput chemistry thereby meeting the challenges of finding and exploiting new lead structures. However, when searching for lead structures, it may be necessary to sacrifice design density in favor of enhancing the diversity of a design to provide an initial hit. This balance exists against the background of the enormous numbers of compounds that are theoretically possible to synthesize. It seems that a greater potential for providing coverage across the field of useful diversity exists by creating multiple chemotype library designs as opposed to a fewer number of library designs composed of many more compounds. Such a concept has been described as the "Fewer of Many" as opposed to the "Many of Fewer".

There are other practical issues that impact on the question of appropriate library size. The capacity and cost of an organization's HTS technology will have influence on the number of compounds that can be routinely screened. Although it may seem appealing to screen as many compounds as possible for each target, the cost-per-well consideration becomes a significant factor for the assay of millions of compounds. Compound handling and tracking systems also influence the way an organization determines the size of the libraries that it wishes to acquire. With each 10-fold increase in the number of compounds, the distribution, archiving and data management of lead generation compounds presents significant additional challenges to process development and organizational structures. Organizations see advantages of focused screening to accelerate the discovery of new lead structures at reduced cost. In effect, there is an increasing interest to shift from the brute force HTS approach to more focused and targeted screening exercises.

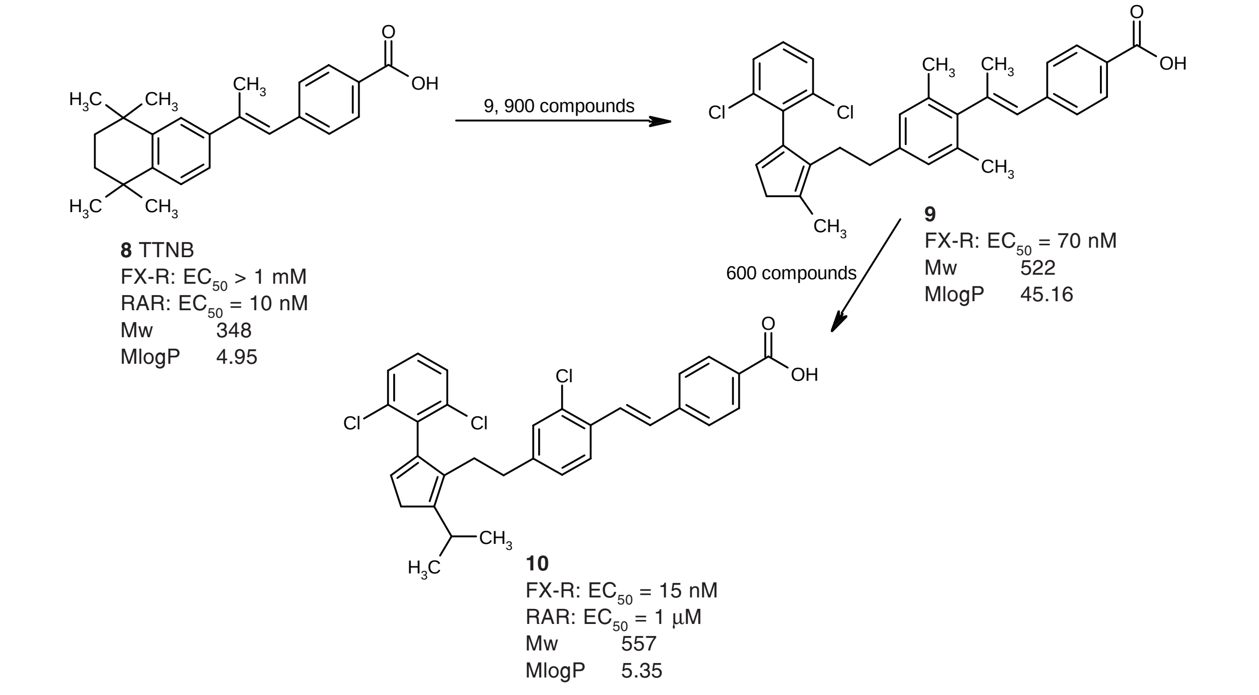
It is in part for the cost and difficulty in establishing and maintaining large compound collections that several companies continue the use of mixture screening. Although trends in HTS have favored the screening of high purity single compounds, the use of mixtures of compounds in conjunction with deconvolution routines continues to be an integral part of new lead generation.

We have observed a critical determinant of optimal library size is actually the chemistry itself. A diversity of reagents results in a diversity of reactivity. There are few reactions in organic chemistry that consistently provide products in high yields for a wide diversity of chemical functionality. Further, the commercial availability of many reagents is quite limited with the exception of a few key, chemical classes. The response to these limitations has been to focus on library chemistry methodology and to incorporate an increasing number of advanced building blocks.

Leveraging the advantages of high-throughput chemistry for building a lead generation collection

The capabilities of high-throughput chemistry are sometimes limiting with respect to synthesizing large, structurally diverse collections of high substance quality (compound purity and correct identity) small molecules with drug-like and/or lead-like properties. Although significant progress in the sophistication and complexity of structures made possible by high-throughput chemical methods is (31-33), expediency often requires shorter high-throughput syntheses using advanced reagents and building blocks. The availability of appropriate building blocks facilitates faster library chemistry development and faster synthesis times of compounds in higher quantity and purity. Building blocks also allow for flexibility in selection of the appropriate high-throughput synthetic methods, such as solution-phase parallel synthesis, solid-phase combinatorial synthesis or solution-phase synthesis using solid-phase reagents. Frequently, building block acquisition is determined by availability and presumed diversity; large collections of these building blocks have been stockpiled without much consideration of library designs for which they would be useful. Currently, the strategy to synthesize protein-family targeted libraries dictates the synthesis and acquisition of building blocks according to particular design themes and motifs. In short, library design strategies should determine which building blocks to make and/or acquire.

The principal appeal of high-throughput synthesis is the production of multiple, defined small molecule structures in a short period of time. Compound characteristics for compound synthesis are often stated as diverse, drug-like, small molecule compound libraries. Strategies differ as to the number of useful compounds per library design from hundreds to tens of thousands. Practical realities of compound handling capabilities and HTS costs, as well as the limitations of a particular chemistry scheme, often impact on the question of appropriate number of compounds per library design. In-house drug discovery processes also determine the necessary quantity of each small molecule; necessary minimum quantities vary from less than 1 mg/compound to more than 20 mg/compound. Synthesizing greater quantities of each compound allows for the archiving of solid sample after distribution of the compounds to HTS format. If a particular compound proves active, follow-up activities are facilitated with the ready availability of that compound. The purity level of each compound is important to most organizations. In order to reduce the false positive rate due to compound impurities (*i.e.*, the rate that screening samples appear active during HTS, but later prove inactive), many organizations define minimum purity criteria (*e.g.*, purity 85% or greater as measured according to UV_{214nm} absorption in a validated LC/MS analytical system). The overall value of the library is increased if compounds which fail to meet these criteria are removed or purified before plating, distribution and archiving. The value of a library is maintained over time if the stability of the compounds contained therein is verified periodically. There are increasingly stringent standards for lead generation libraries with respect to high substance quality, compound quantities and numbers. The explicit determination of these numbers is different for each lead generation operation. The days of one lead generation library to fit all drug discovery operations is an antiquated concept. In summary, although the value of a "good lead" cannot be overstated, the challenges of creating a large and diverse collection of "quality" potential lead molecules should not be understated.



8 TTNB
FX-R: EC_{50} > 1 mM
RAR: EC_{50} = 10 nM
Mw 348
MlogP 4.95

9
FX-R: EC_{50} = 70 nM
Mw 522
MlogP 45.16

10
FX-R: EC_{50} = 15 nM
RAR: EC_{50} = 1 μM
Mw 557
MlogP 5.35

Fig. 8. Identification of FXR selective ligands through a lead generation library approach.

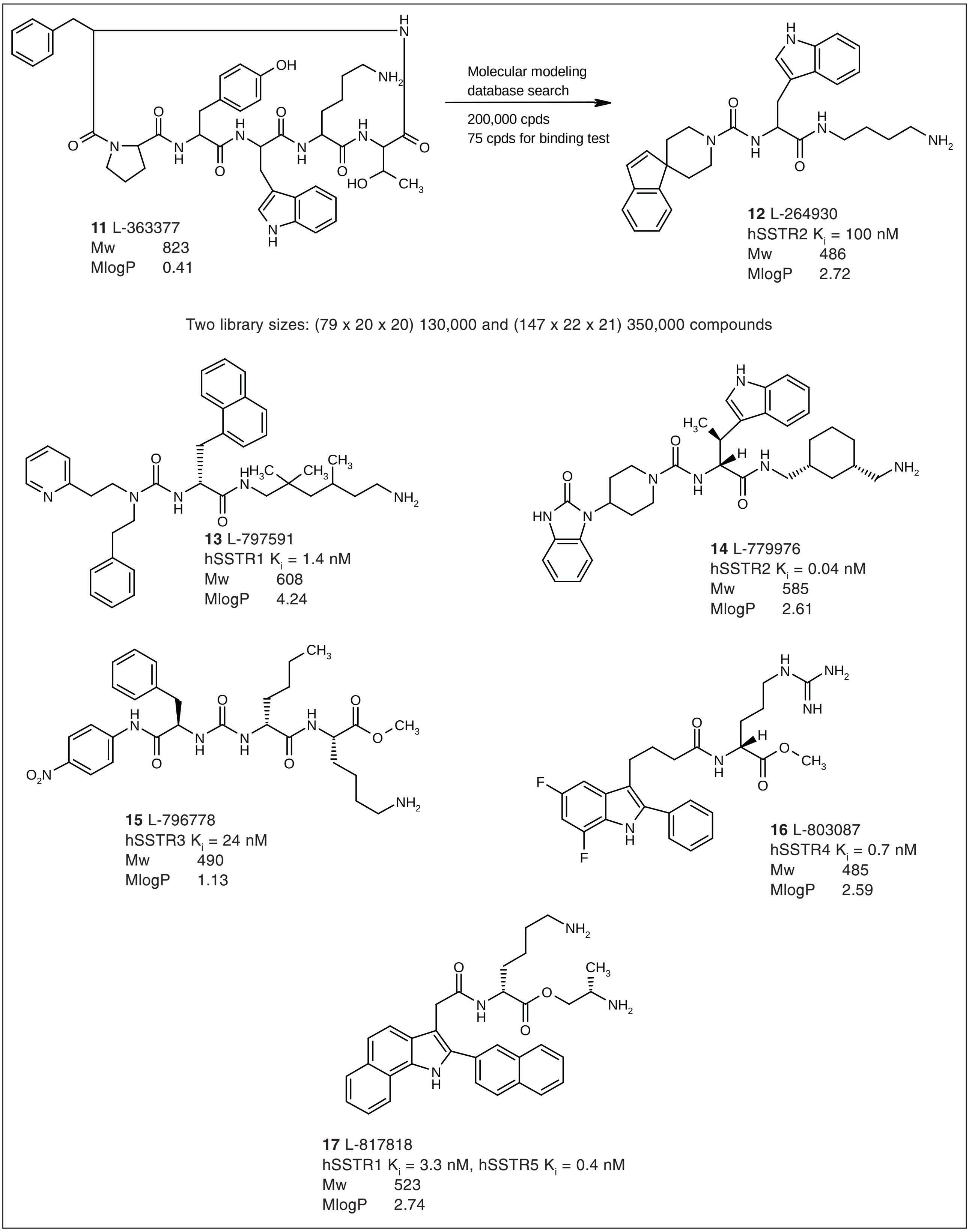
Selected examples of combinatorial chemistry-based lead generation

As the generation of new leads for drug discovery has great potential to provide advantage in the competitive pharmaceutical environment, few detailed examples of successful lead generation exist in scientific literature (34). However, several examples provide insight into the necessary scale and strategies that have been undertaken to provide such new lead structures.

Scientists at Pfizer have reported the identification of a selective, nonsteroidal agonist for a recently identified nuclear receptor family member, the FX receptor (Fig. 8) (35). After synthesis and assay of a 9900 member lead generation library, Structure **9** was identified having a potency of EC_{50} = 70 nM (partial agonist). A focused library of 600 structures was initiated to optimize the potency, efficacy and selectivity with respect to farnesoid x receptor (FXR) activity. This resulted in Structure **10** having an EC_{50} = 15 nM. Drug-like characteristics were demonstrated in oral dosing in rats. The compound showed significant lowering of serum triglyceride levels. This example is helpful in illustrating the number of compounds necessary to find a diverse initial hit starting from a known active. The necessity of an additional 600 compounds to enhance potency 4-fold while maintaining drug-like characteristics is also noteworthy.

A seminal paper by scientists at Merck described the discovery of potent and selective agonists for 5 subtypes of human somatostatin receptors (36). Cyclic peptide-based pharmacophore information was used for virtual screening of 200000 library compounds; 7500 compounds were predicted to be hits. The predicted hits were assayed in a binding assay with membranes from CHO cells expressing the human sst2 receptor. The most potent compound had a K_i of 100 nM (Fig. 9). Based on this result and the identification of privileged fragments contained in both the small molecule and in the peptide, several mixture libraries were synthesized using split and pool combinatorial chemistry methods. Assay of 100 compound pools followed by deconvolution resulted in the identification of several lower nanomolar, selective ligands for 5 human somatostatin receptor subtypes. Increases in potency and selectivity were achieved with some increases in molecular weight and lipophilicity. Although further elaboration of these ligands as drugs has not been reported, the potential for the discovery of highly active and selective structures with libraries of this size is exemplified.

Scientists at Affymax used, thiol-containing diketopiperazine **18**, a nonselective MMP inhibitor of double-digit micromolar inhibitory potency as a design starting point (Fig. 10) (37). A 1225-member compound library was useful in improving potency with respect to MMP-1 roughly 2 orders of magnitude while enhancing protein family subtype selectivity. A slightly smaller library was necessary to further improve potency against MMP-1 while maintaining selective inhibitory activity. It seems that with the critical thiol functionality contained in the initial structure, modest potency was observed. Additional



Molecular modeling
database search

200,000 cpds
75 cpds for binding test

11 L-363377
Mw 823
MlogP 0.41

12 L-264930
hSSTR2 K_i = 100 nM
Mw 486
MlogP 2.72

Two library sizes: (79 x 20 x 20) 130,000 and (147 x 22 x 21) 350,000 compounds

13 L-797591
hSSTR1 K_i = 1.4 nM
Mw 608
MlogP 4.24

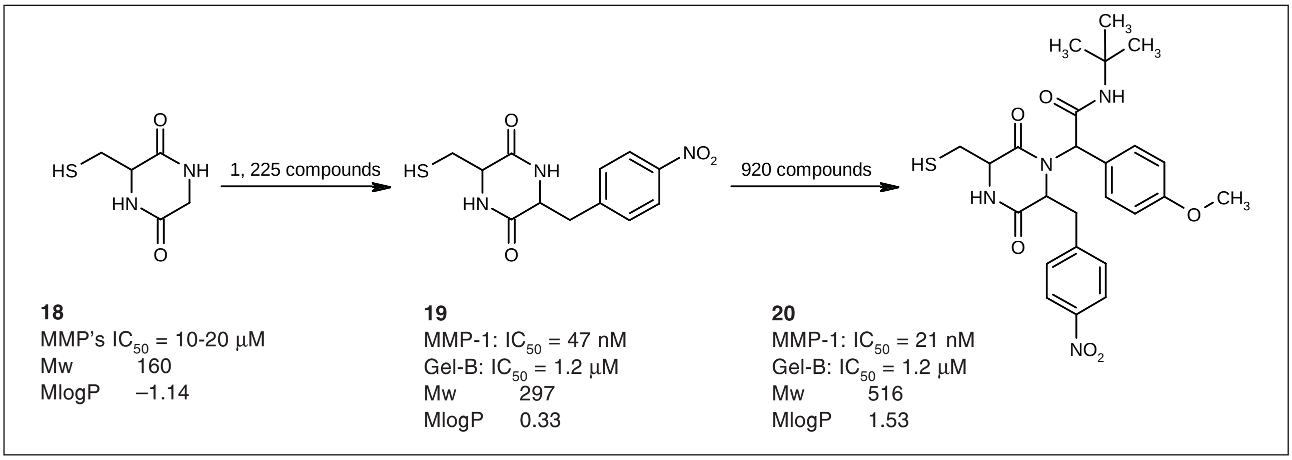
14 L-779976
hSSTR2 K_i = 0.04 nM
Mw 585
MlogP 2.61

15 L-796778
hSSTR3 K_i = 24 nM
Mw 490
MlogP 1.13

16 L-803087
hSSTR4 K_i = 0.7 nM
Mw 485
MlogP 2.59

17 L-817818
hSSTR1 K_i = 3.3 nM, hSSTR5 K_i = 0.4 nM
Mw 523
MlogP 2.74

Fig. 9. Identification of selective human somatostatin receptor ligands through screening of mixture libraries generated through combinatorial chemistry. Results are expressed as K_i values (nM) for compounds in the presence of 3-[^{125}I]-iodotyrosine-25-somatostatin and somatostatin subtype-expressing CHO-K1 cells.

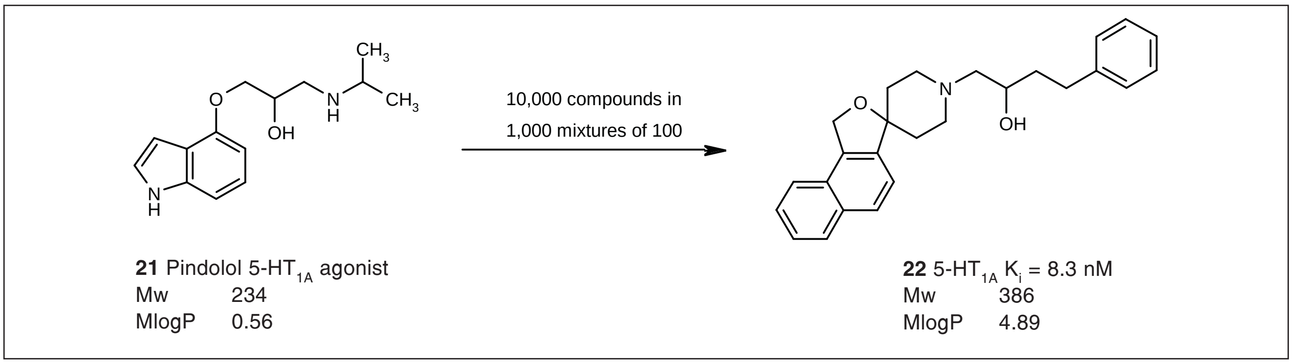


18
MMP's IC_{50} = 10-20 μM
Mw 160
MlogP –1.14

19
MMP-1: IC_{50} = 47 nM
Gel-B: IC_{50} = 1.2 μM
Mw 297
MlogP 0.33

20
MMP-1: IC_{50} = 21 nM
Gel-B: IC_{50} = 1.2 μM
Mw 516
MlogP 1.53

Fig. 10. Identification MMP-1 selective ligand built up from thiol containing diketopiperazines.



21 Pindolol 5-HT_{1A} agonist
Mw 234
MlogP 0.56

22 5-HT_{1A} K_i = 8.3 nM
Mw 386
MlogP 4.89

Fig. 11. Discovery of a nanomolar 5-HT_{1A} inhibitor through screening of mixture libraries generated by combinatorial chemistry.

potency and selectivity were achieved by filling adjacent protein pockets with lipophilic aryl moieties. This is a good example of the importance of selecting a template with the minimum, critical functionality to which is then added diversity elements to enhance potency and selectivity. It is important to note the size of the first and second libraries and the resulting gains of potency and selectivity. The first library compounds resulted in rather dramatic increases in potency and selectivity, whereas the second library resulted in modest gains.

In the discovery of 5-HT_{1A} receptor antagonists, Merck scientists exemplified the successful application of the privileged fragment concept (38). The amino alcohol moiety of pindolol **21** is a fragment frequently found in ligands of biogenic amine GPCRs. Chemistry for the assembly of such compounds from a diversity of phenols, epichlorohydrin and amine nucleophiles is particularly suited to a high-throughput chemistry approach (Fig. 11). Substantial potency and selectivity gains were achieved with large, focused compound libraries. These gains have come at the expense of increased molecular weight and lipophilicity; however, Merck scientists report that Structure **22** penetrates the blood-brain barrier.

An outlook for lead generation: the chemical genomics concept

The chemical genomics process is the next stage in the evolution of greater efficiency in drug discovery; processes for the generation of small molecule leads are central to this concept (39). With the sequencing of the human genome, a new understanding of the relationships among gene products makes it possible to consider grouping biological targets in families according to similarities in gene sequence. Central to the chemical genomics concept is the grouping of protein families not only by their gene sequence similarities, but also by the similarity among the ligands that show activity for those biological targets (40). In the chemical genomics approach to drug discovery, instead of single targets treated one at a time, protein target families are treated in parallel with protein-targeted compound libraries (Fig. 12). The parallel nature takes into account the likely attrition rate of molecules moving through this process and thus offers the potential for accelerating the discovery of new lead structures.

A central concept of the chemical genomics process is bringing together diverse sets of information to facilitate

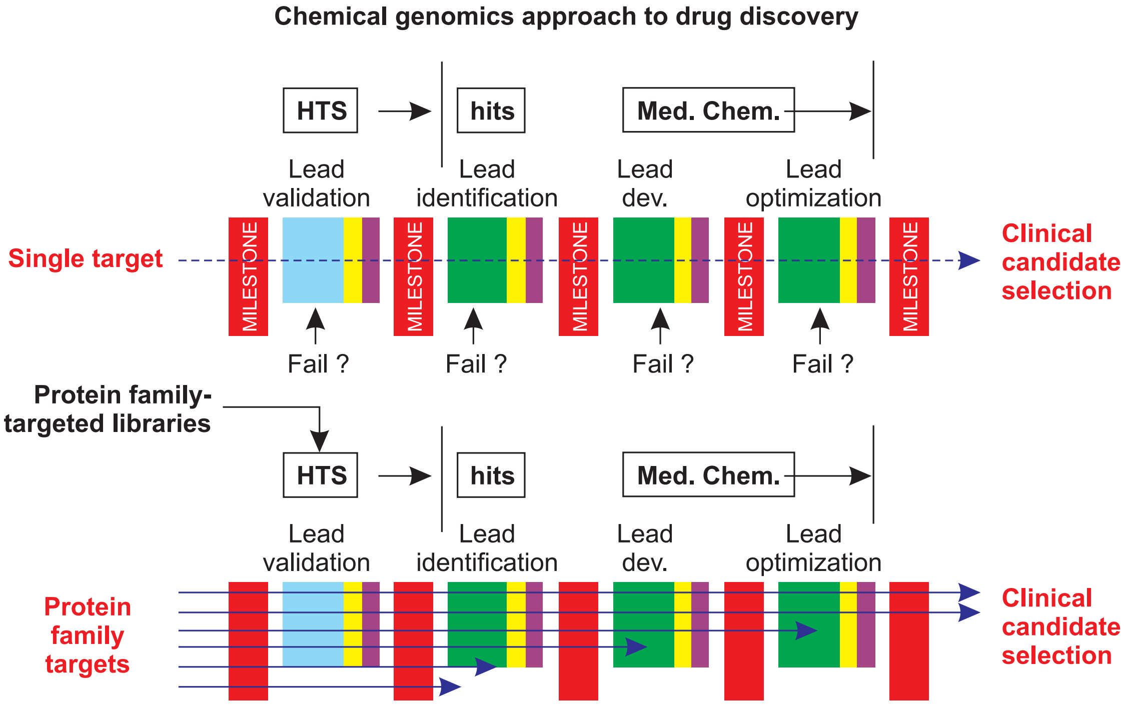


Fig. 12. Enhancing efficacy of the drug discovery process through chemical genomics. Instead of single targets treated one at a time, protein target families are treated in parallel with protein-targeted compound libraries. This method allows for the attrition rate of molecules moving through this process.

better and more informed decision-making in library design. Chemical genomics includes annotation of interactions amongst diverse collections of small “lead” compounds and gene products from the human genome. Pertinent data to be annotated are more than just a compound’s potency with respect to a single target; rather a compound’s potency against multiple targets, as well as information about toxicity, metabolism and various similarity metrics to known drugs should be considered. A useful chemical genomics database would include information to gauge the potential for finding a small molecule drug for a particular target or family of targets. Finally, the chemical genomics approach requires the development of ways to identify trends or signals of association of biological targets according to chemistry and vice versa. This information process is designed to be a scalable approach to leverage annotation in finding potent, selective, less toxic disease-modulating molecules against the many new targets that have been elucidated following the human genome decoding. The goal of the process is not only the evaluation of the best molecules to make, but also to provide information which will facilitate target validation earlier in the discovery process (Fig. 13). Recent progress has been reported by scientists at De Novo Pharmaceuticals in developing consensus neural networks to group molecules according to the target families to which they bind (41). These methods allowed for greater than 80% success rate in the identification of the correct target family for a particular molecule. Methods such as these strengthen the central chemical genomics activity of grouping compounds into families according to their target binding and biological activity.

A visualization of biological activity that reveals the target family bias of a focused compound library is shown in Figure 14. The raw in-house data resulting from 65 HTS campaigns for a set of 673 compounds designed as kinase inhibitors reveal a strong signal for several kinase-family enzyme assays. Also noted are signs of affinity for a number of GPCR target family assays. Ideally, affinity comparison should be made with IC_{50} and K_i numbers, but it is possible to perceive initial signs of affinity from raw data. The chemogenomics approach of linking small molecule structural motifs to protein family targets requires algorithms to recognize initially weak signals of association before costly and time consuming secondary assays can be run on a more focused set of compounds and targets. This example illustrates the importance of medium- to large-sized libraries in associating gene family targets to chemical themes or motifs. If a particular design is represented by only a few compounds, then it is unlikely to provide a signal of sufficient strength for detection against the background of hundreds of thousands of data points. These plots exemplify the scale of a chemogenomics information integration effort; relating

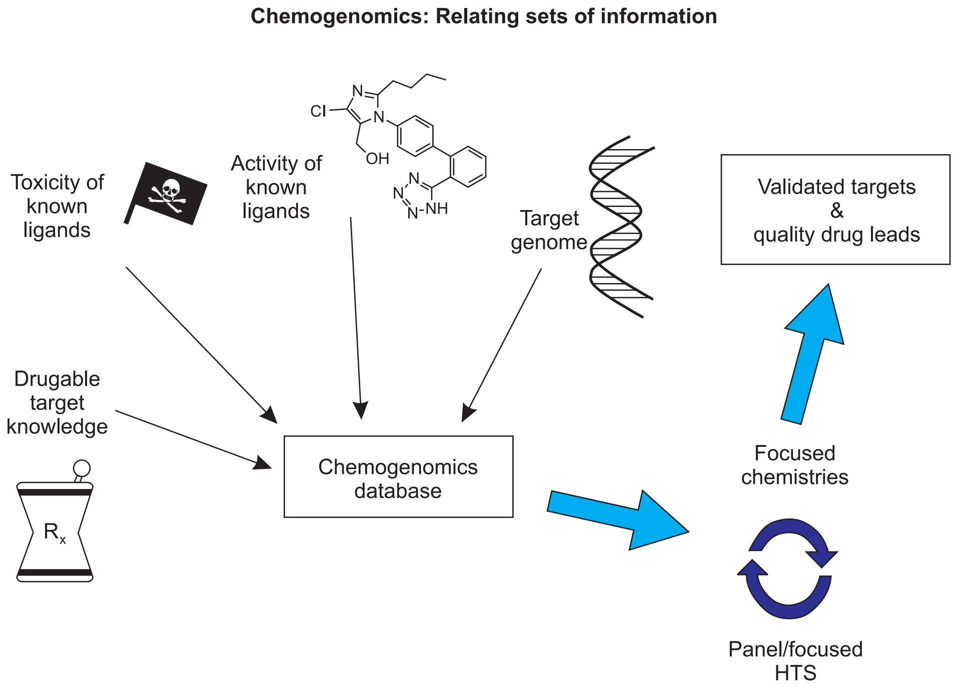

Chemogenomics: Relating sets of information
Toxicity of known ligands
Activity of known ligands
Cl
N
N
OH
N
N
N-NH
Target genome
Validated targets & quality drug leads
Drugable target knowledge
R_x
Chemogenomics database
Focused chemistries
Panel/focused HTS

Fig. 13. Central to the chemical genomics approach is the construction of a database of small molecule ligands which are annotated with respect to the relevance to a particular gene and protein target, relation to other known drugs, toxicity, *etc.* The usefulness of the chemogenomics database is particularly highlighted in the way it may be used to direct the synthesis and bioassay of focused libraries.

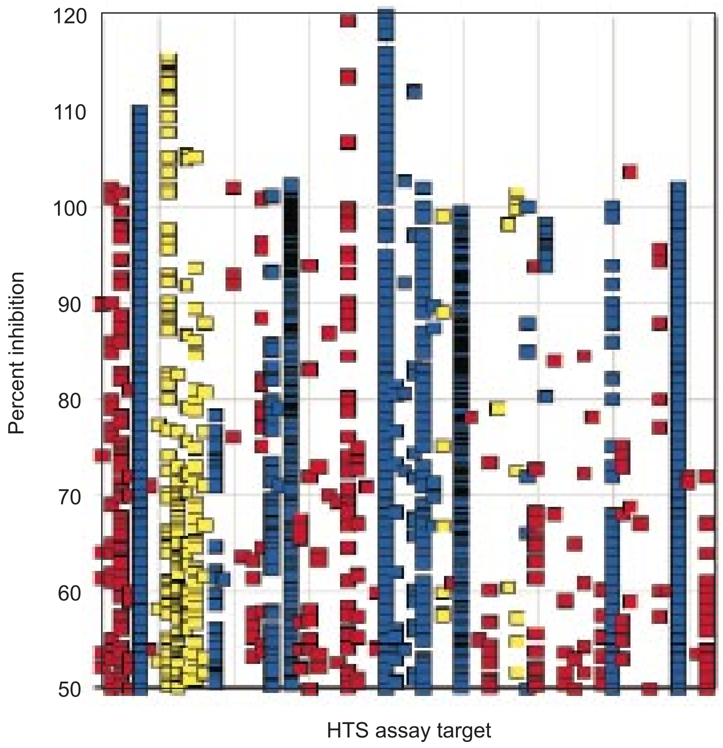


Fig. 14. Example of an affinity fingerprint of a small molecule library of 643 members designed for kinase inhibition assayed against 65 HTS assays at 10-20 µM. Data points represent percent inhibition for compounds for kinase-family targets (blue), GPCR-family targets (yellow) and other targets (red).

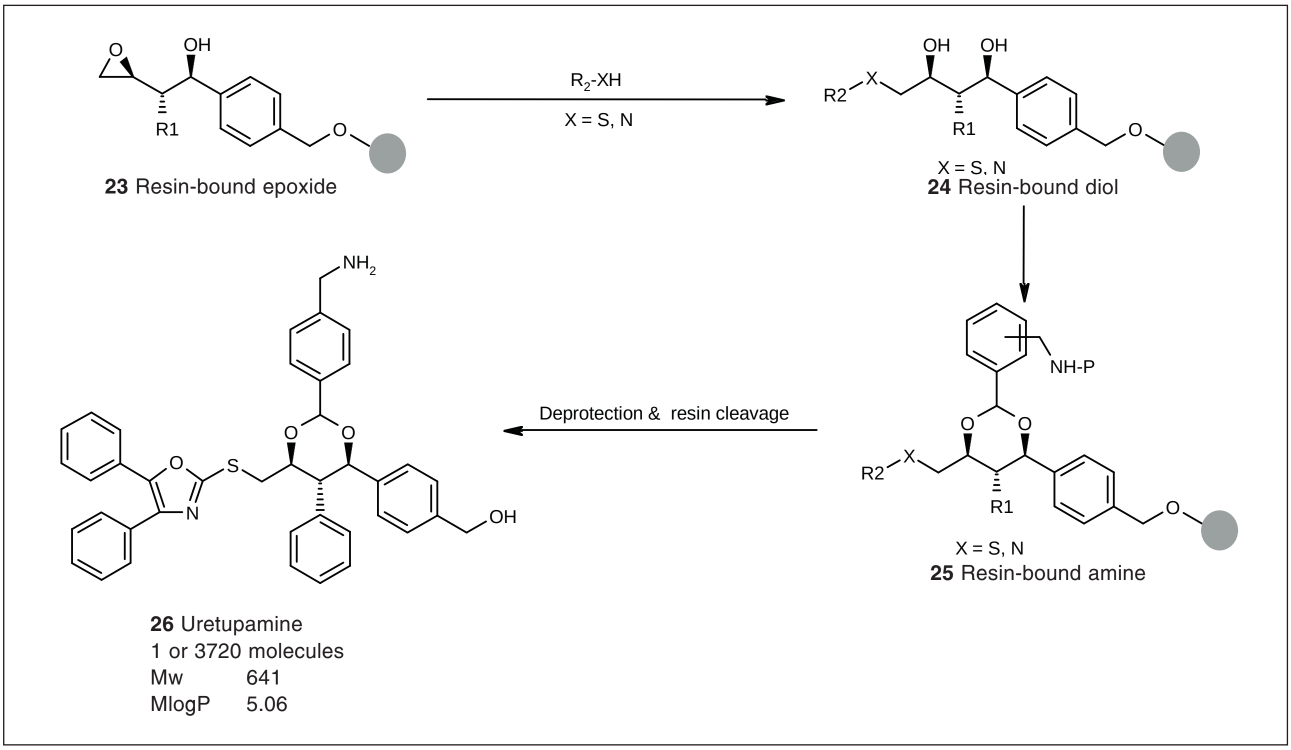


Fig. 15. Discovery of uretupamine, a modulator of the glucose-sensing Ure2p repressor.

such large data sets composed of diverse types of information requires new ontological strategies (42) to extract and leverage useful information to generate leads with a higher level of sophistication in terms of “drug-likeness”.

A variation of the chemical genomics approach has been reported by Schreiber who have coined the term “chemical genetics” to described the use of small molecules to directly modulate the activity of certain gene products, affecting a specific biological pathway or process (43). Schreiber *et al.* published an example of chemical genetics describing the use of solid phase combinatorial synthesis of structurally diverse molecules found to be useful for modulating the protein Ure2p, a repressor of several transcription factors (Fig. 15) (44). Structures of this type acted with relative selectively on Ure2p and modulated a subset of the activities of that protein, thus allowing for a dissection of Ure2p’s signaling role. With uretupamine, it was possible to define a mechanistic relation between repressor Ure2p, transcription factor Nil1p and glucose levels.

Acknowledgements

Drs. Paul Gillespie, Sung-Sau So, Jefferson Tilley and Stephen Ferguson are acknowledged for their helpful discussion, critical reading of this manuscript and collaborative efforts in lead generation. Dr. John Proudfoot of Boehringer Ingelheim (CT) is gratefully acknowledged for making available the sd file noted in reference 7.

References

1. Dimasi, J. (Tufts University). Unpublished communication.

2. Harris, G. Study Increase in Costs of Discovering, Developing Drug. The Wall Street Journal Europe 2001, December 4, 11.

3. Bailey, D., Brown, D. *High-throughput chemistry and structure-based design: Survival of the smartest.* Drug Discov Today 2001, 6: 57-9.

4. Venter, J. Craig, Adans, M.D., Myers, E.W. et al. *The sequence of the human genome.* Science 2001, 291: 1304-51.

5. Sadee, W., Hoeg, E., Lucas, J., Wang, D. *Genetic variations in human G protein-coupled receptors: Implications for drug therapy.* AAPS Pharm Sci 2001: 3: E22.

6. Hertzberg, R.P., Pope, A.J. *High-throughput screening: New technology for the 21st century.* Curr Opin Biol 2000, 4: 445-51.

7. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J. *Experimental and computational approaches to estimate solubility and permability in drug discovery and development settings.* Adv Drug Deliv Rev 1997, 23: 3-25.

8. Lipinski, C.A. *Drug-like properties and the causes of poor solubility and poor permeability.* J Pharmacol Toxicol Methods 2000, 44: 235-49.

9. Simulations Plus - QMPR Plus software was used for property prediction. The drug data set included 29 recently launched drugs noted by John Proudfoot in Proudfoot, J.R. *Drugs, leads and drug-likeness: An analysis of some recently launched drugs.* Bioorg Med Chem Lett 2002, 12: 1647-50.

10. Walters, W.P., Murcko, M.A. *Prediction of ‘drug-likeness’*. Adv Drug Deliv Rev 2002, 54: 255-71.

11. Veber, D.F., Johnson, S.R., Cheng, H.-Y., Smith, B.R., Ward, K.W., Kopple, K.D. *Molecular properties that influence the oral bioavailability of drug candidates.* J Med Chem 2002, 45: 2615-23.

12. Matter, H., Baringhaus, K.-H., Naumann, T., Klabunde, T., Pirard, B. *Computational approaches towards the rational design of drug-like compound libraries.* Comb Chem High Throughput Screen 2001, 4: 453-75.

13. Teague, S.J., Davis, A.M., Leeson, P.D., Oprea, T. *The design of leadlike combinational libraries*. Angew Chem Int Ed 1999, 38: 3743-8.

14. Oprea, T.I., Davis, A.M., Teague, S.J., Leeson, P.D. *Is there a difference between leads and Drugs? A historical perspective.* J Chem Info Comput Sci 2001, 41: 1308-15.

15. Fejzo, J., Lepre, C.A., Peng, J.W. et al. *The SHAPES strategy: An NMR-based approach for lead generation in drug discovery.* Chem Biol 1999, 6: 755-69.

16. Hann, M.M., Leach, A.R., Harper, G. *Molecular complexity and its impact on the probability of finding leads for drug discovery.* J Chem Info Comput Sci 2001, 41: 856-64.

17. Kennedy, T. *Managing the drug discovery/development interface.* Drug Discov Today 1997, 2: 436-44.

18. Jorgensen, W.L., Duffy, E.M. *Prediction of drug solubility from structure.* Adv Drug Deliv Rev 2002, 54: 355-66.

19. Rishton, G.M. *Nonleadlikeness and leadlikeness in biochemical screening.* Drug Discov Today 2003, 8: 86-96.

20. Roche, O., Schneider, P., Zuegge, J. et al. *Development of a virtual screening method for identification of frequent hitters' in compound libraries.* J Med Chem 2002, 45: 137-42.

21. Patchett, A.A., Nargund, R.P. *Privileged structures.* Annu Rep Med Chem 2000, 35: 289-98.

22. Honma, T., Hayashi, K., Aoyama, T. et al. *Structure-based generation of a new class of potent Cdk4 inhibitors: New de novo design strategy and library design.* J Med Chem 2001, 44: 4615-27.

23. Boehm, H-J., Boehringer, M., Bur, D. et al. *Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening.* J Med Chem 2000, 43: 2664-74.

24. Boehm, H.J. *On the use of LUDI to search the fine chemicals directory for ligands of proteins of known three-dimensional structure.* J Comput-Aided Mol Des 1994, 8: 243-56.

25. Sprague, P.W. *Automated chemical hypothesis generation and database searching with catalyst.* Perspect Drug Discov Design 1995, 3: 1-20.

26. Schneider, G., Neidhart, W., Giller, T., Schmid, G. *"Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening.* Angew Chem Int Ed 1999, 38: 2894-6.

27. Naerum, L., Norskov-Lauritsen, L., Olesen, P.H. *Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors.* Bioorg Med Chem Lett 2002, 12: 1525-8.

28. Bemis, G.W., Murcko, M.A. *The properties of known drugs. 1. Molecular frameworks.* J Med Chem 1996, 39: 2887-93.

29. Johnson, B.M., Babcock, W.C., West, J.B., Friesen, D.T. *Preparation of bioactive compounds by plasma synthesis.* US (1998), 22 pp. CODEN: USXXAM US 5772855 A 19980630 CAN 129: 53942 AN 1998: 430001.

30. Terstappen, G.C., Reggiani, A. *In silico research in drug discovery trends in pharmacol.* Sciences 2001, 22: 23-6.

31. Dolle, R.E. *Comprehensive survey of combinatorial library synthesis 2000*. J Comb Chem 2001, 3: 477-517.

32. Dolle, R.R. *Comprehensive survey of combinatorial library synthesis: 1999.* J Comb Chem 2000, 2: 383-433.

33. Dolle, R.E., Nelson, K.H. Jr. *Comprehensive survey of combinatorial library synthesis: 1998.* J Comb Chem 1991, 1: 235-82.

34. Adang, A.E.P., Hermkens, P.H.H. *The contribution of combinatorial chemistry to lead generation: An interim analysis.* Curr Med Chem 2001, 8: 985-98.

35. Maloney, P.R., Parks, D.J., Haffner, C.D. et al. *Identification of a chemical tool for the orphan nuclear receptor FXR.* J Med Chem 2000, 43: 2971-4.

36. Rohrer, S.P., Birzin, E.T., Mosley, R.T. et al. *Rapid identification of subtype-selective agonists of the somatostatin receptor through combinatorial chemistry.* Science 1998, 282: 737-40.

37. Szardenings, A.K., Antonenko, V., Campbell, D.A. et al. *Identification of highly selective inhibitors of collagenase-1 from combinatorial libraries of diketopiperazines.* J Med Chem 1999, 42: 1348-57.

38. van Niel, M.B., Beer, M.S., Castro, J.L. et al. *Parallel synthesis of 3-aryloxy-2-propanolamines and evaluation as dual affinity $5\text{-}HT_{1A}$ and 5-HT re-uptake ligands.* Bioorg Med Chem Lett 1999, 9: 3243-8.

39. Bleicker, K. *Chemogenomics: Bridging a drug discovery gap.* Curr Med Chem 2002, 9: 2077-84.

40. Jacoby, E. *A novel chemogenomics knowledge-based ligand design strategy - Application to G protein-coupled receptors.* Quant Struct-Act Relat 2001, 20: 115-23.

41. Manallack, D.T., Pitt, W.R., Gancia, E. et al. *Selecting screening candidates for kinase and G protein-coupled receptors using neural networks.* J Chem Inf Comput Sci 2002, 42: 1256-62.

42. Schuffenhauer, A., Zimmerman, J., Stoop, R. et al. *An ontology for pharmaceutical ligands and its application for silico screening and library.* Design J Chem Inf Comput Sci 2002, 42: 947-55.

43. Schreiber, S.L. *Chemical genetics resulting from a passion for synthetic organic chemistry.* Bioorg Med Chem 1998, 6: 1127-52.

44. Kuruvilla, F.G-, Shamji, A.F., Sternson, S.M. et al. *Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microaarays.* Nature 2002, 416: 653-7.