The Design and Docking of Virtual Compound Libraries to Structures of Drug Targets

Amy C. Anderson* and Dennis L. Wright

Department of Chemistry, Dartmouth College, Hanover, NH 03755, USA

Abstract: This review provides a detailed analysis of the use of virtual library screening (VLS) in the drug discovery process. The first part is intended as a larger overview of the integrated VLS process. Small molecule and target macromolecule considerations will be described separately and will be subsequently integrated in a discussion of docking, scoring and evaluation. The second half of the review will focus on recent case studies that use VLS as part of an integrated drug discovery program. The case studies will illustrate the range of possible targets in VLS, provide an account of inclusive methodology and reveal the expectations for realistic goals. Recent efforts provide compelling evidence that VLS is successful when practiced in an integrated fashion involving synthetic, structural and computational expertise.

Keywords: Virtual screening, structure-based drug design, library design, drug discovery, diversity, filtering, ligand binding, docking.

INTRODUCTION

It is widely appreciated that advances in the biological component of drug development have catalyzed a shift in the strategies and tactics that underlie the drug discovery process. New information has evolved to describe disease states at the molecular rather than organismic level, which in turn presents those involved in drug development with a large array of well-defined targets. Additionally, economic factors are driving the need for a shorter lead-to-drug development time.

A number of methodologies have evolved to integrate the higher degree of molecular information, number of new targets and need for efficiency. This integration has been most widely implemented in the coupling of high-throughput screening (HTS) with high-output chemical synthesis [1-3]. HTS relies on the development of efficient and reliable assays to permit the evaluation of a large number of compounds against a target in a rapid and often automated manner. The large volume of HTS data is modeled in order to assess structure-activity relationships, but problems arise when these models suffer from distortion by false positives. Combinatorial chemistry, the synthesis of a very large number of compounds using a single scaffold and a diverse array of reactants, has also attempted to address the need for a large number of new drug leads [4-6]. This methodology is severely limited by the labor-intensive and costly efforts required to prepare and purify such large numbers of compounds.

Virtual screening, using a computational approach to assess the interaction of an *in silico* library of small molecules and the structure of a target macromolecule, has arisen as an alternative method for the rapid identification of new drug leads. A great deal of effort has been extended to create reliable and efficient software that evaluates the highly complex nature, both enthalpic and entropic, of the interaction between small molecules and their macromolecular receptors. A typical virtual library screening (VLS) approach involves several stages, Fig. (**1**), including parallel efforts that involve small molecule and macromolecule preparation.

Various stages of the VLS methodology described in Figure **1** have been previously reviewed [7-11] and provide an excellent resource for detailed analyses of many of the components of this process. The first part of this review is intended as a larger overview of the integrated VLS process. Small molecule and target macromolecule considerations will be described separately and will be subsequently integrated in a discussion of docking, scoring and evaluation. The second half of the review will focus on recent case studies that use VLS as part of an integrated drug discovery program. The case studies will illustrate the range of possible targets in VLS, provide an account of inclusive methodology and reveal the expectations for realistic goals.

SMALL MOLECULE VIRTUAL LIBRARIES

Sources of New Chemical Entities

At the present time, the overwhelming majority of clinically used drugs are small, organic-based molecules that represent an amazing array of structural diversity [12, 13]. Although macromolecular agents such as proteins and nucleic acids are entering the clinical arena [14], small molecules are certain to play a major role in therapeutic development for decades to come. New chemical entities represent one of the key pillars of the modern drug discovery effort. It is critical that the identification of a biological target (usually protein) that mediates a disease state is followed by the identification of a small-molecule effector (ligand) that can interact with and alter the biological function of the target. In the past, these ligands have been identified through a screening process that involves the establishment of a reliable biological assay followed by testing of collections of compounds, usually proprietary legacy collections that have been established in-house by

*Address correspondence to this author at the Department of Chemistry, Dartmouth College, Hanover, NH 03755, USA; Tel.: (603) 646-1154; Fax: (603) 646-3946; Email: Amy.C.Anderson@Dartmouth.edu

1573-4099/05 $50.00+.00

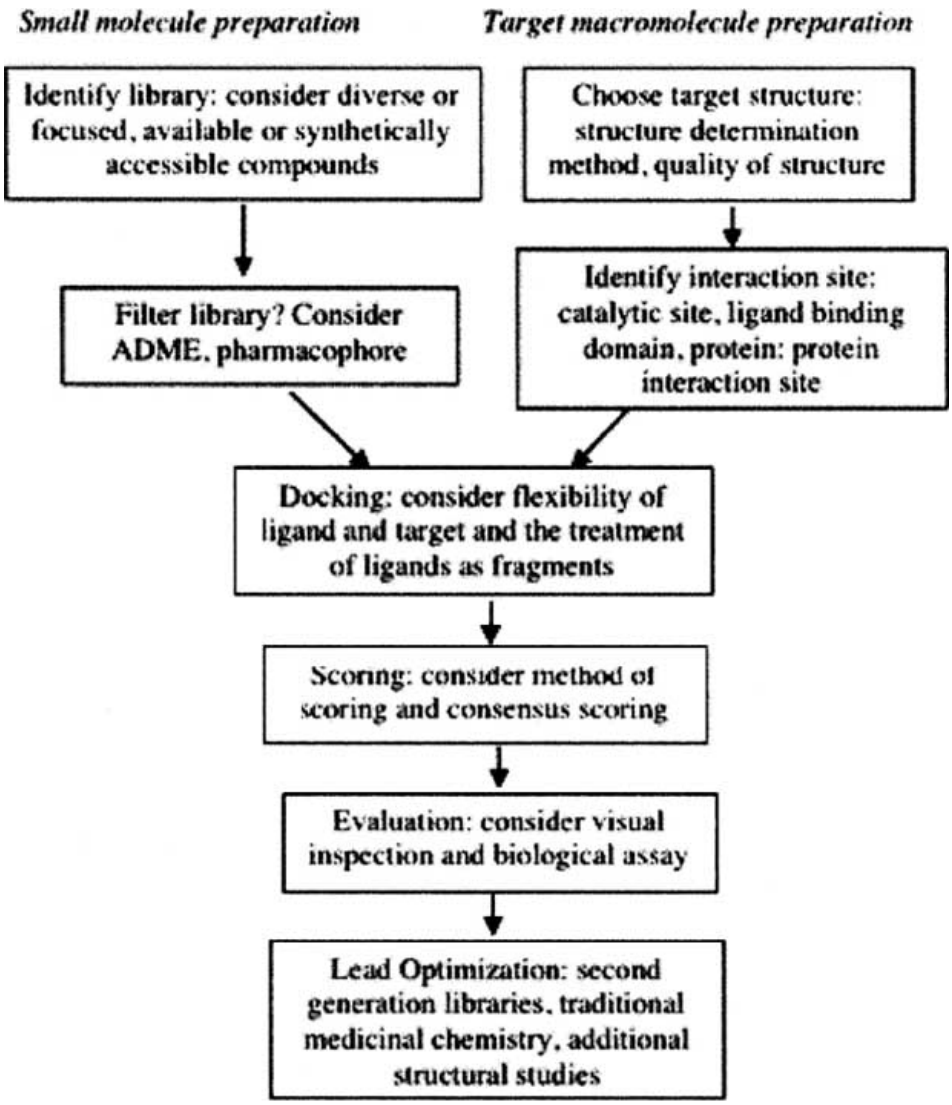


Fig. (1). A typical VLS scheme. Stages include both small molecule library preparation (choice of library, considerations for filtering) and target preparation (choice of structure of target and identification of binding site). In VLS, the library is docked into the target, scored and evaluated. Any possible leads are optimized in later stages.

large pharmaceutical companies. The assembly of these compound collections has been a highly variable and often random process that involved natural product isolates, synthetic intermediates and purchased samples from academic labs. Often, these compounds were byproducts of other medicinal chemistry efforts and may account for some of the redundancy in medicinal structural types. Despite the lack of an overriding rationale for the assembly of these collections, these screening libraries have produced many initial "hits" against the biological system of interest and have led to the development of new therapeutic agents. New strategies and tactics such as parallel synthesis and solid-phase organic synthesis have appeared in synthetic chemistry to assemble large groups of novel compounds for biological evaluation.

Synthetic Challenge in Library Synthesis

From the perspective of the synthetic chemist, the post-genomic age represents a wealth of new opportunities coupled with new challenges that are pushing the limits of contemporary synthetic methodology. Since the 1950s, synthetic chemistry has primarily been practiced as target-oriented synthesis (TOS) whereby single compounds were slated for preparation [15]. The translation of the designed route to practice is almost always complicated by difficulties with certain synthetic steps. These difficulties are often attributed to an imperfect understanding of reactivity and selectivity of complex molecules. Frequently, these problematic steps can be circumvented by careful reaction optimization or tactical modification of the first-generation route that can make synthesis a costly and time-consuming practice.

In order to satisfy the demand for novel compounds, a new strategy of diversity-oriented synthesis (DOS) has emerged [16, 17] to complement the older TOS strategy. In DOS, the synthetic chemistry is geared towards the preparation of large arrays of compounds that sample a vast swath of chemical space. In addition to changes in strategy,

changes in the practice of synthetic chemistry such as parallel synthesis [18] (manual or automated) have made it possible to prepare thousands of compounds in much shorter periods of time. However, these high speed methods of synthesis do not guarantee that each individual member of the library will actually be made or even ensure that the quality of the materials produced will be sufficient for biological assay. Another major limitation in parallel synthesis is the extreme cost associated with the preparation and purification of such a large number of chemical entities.

Again from the perspective of the synthetic chemist, the challenge of preparing these large arrays becomes daunting. It has been a commonly held belief that the larger a library is, the better the chance of locating a hit against a novel biological target. However, as the size of the library increases, the cost and difficulty of synthesis necessarily increases. Often, the types of compounds slated for library synthesis are given a high priority based on ease of preparation, not diversity or druggability considerations. The high cost of each additional step in a sequence frequently biases the synthetic chemist towards reactions with high step-economy whereby a single reaction leads to a large increase in molecular complexity [19]. Multi-component reactions, such as the Ugi, Biginelli or Hantzch condensations [20] and "click" reactions [21] are prized in library production because of their very high step-economy. These compounds tend to be overrepresented in libraries due solely to their ease of preparation.

Virtual Screening and the Synthetic Challenge

The application of virtual screening protocols to lead identification offers an opportunity to drastically reduce the time and cost associated with the production of libraries for screening. It is intuitive that the production of a library *in silico* is a significantly simpler exercise than the chemical synthesis of the same library. This is arguably the greatest impact virtual screening will have on the drug discovery process. If a reliable and efficient computational protocol can be established to rank each member of a library as to its ability to interact in a productive manner with a macromolecular receptor, then it becomes unnecessary to prepare, through synthesis, each member of the library. A much smaller group of target compounds can be slated for production. This drastic reduction in the demand placed on the synthetic component of the discovery team allows for a greater effort to be placed in a more directed TOS-like activity that can take advantage of more reliable methods of chemical synthesis.

The ideal screening library would contain every conceivable molecule that had good drug-like properties with the notion that a ligand could then be found for every receptor in the genome. However, in reality, even the most extensive screening library will contain only a very small fraction of all chemical space. Approximations of the total number of unique chemical entities have ranged as high as 10^{100}. This number, if multiple conformations of each library member is considered, dramatically increases. The enormous size of these numbers indicate that there are far too many possible structures to prepare through chemical synthesis or even generate *in silico*.

The realization that, even in virtual screening programs, only a subset of all possible compounds can be examined has prompted many investigators to consider methods to reduce the library to a manageable size [22, 23]. Care should be taken to ensure that those compounds remaining in the virtual library have desirable properties for potential drug leads [24-26].

Composition of a Virtual Library

It has become desirable to prepare descriptors for chemical libraries to evaluate how much of chemical space is sampled, often referred to as the diversity of a given library [27, 28]. A simplified version of the binding of large molecules by small molecules regards the ligand as a series of interacting groups (VDW, electrostatic, H-bond donor/acceptor) projected in three-dimensional space [29, 30]. When a molecule presents these groups in the correct orientation to make energetically favorable contacts with side-chain or backbone groups in the protein, binding can occur. In considering the diversity of a library, the types of groups and their relative orientation are critical, but not easily defined parameters. Despite the difficulty, a number of different computational descriptors such as Tanimoto coefficients have been formulated to define the degree of difference between members of a library [31, 32]. These types of dissimilarity analyses are often used to guide compound clustering whereby molecules that are structurally related are grouped together [33]. Frequently, this type of grouping exercise is used as a library filter and candidate compounds for screening are selected from each cluster.

In essence, there are different ways of viewing molecular diversity and the degree of diversity desired is often related to the particular screening task at hand. In the absence of any information regarding potential ligands that would bind to a defined site on a macromolecular target, it is desirable to have the maximum structural diversity in the virtual library such that the chance of locating hits increases. For *in silico* screening, recourse is typically made to large, public databases such as the Available Chemicals Directory (ACD) which contains ~250,000 compounds or the National Cancer Institutes (NCI) compound database with ~208,000 compounds. These libraries can be regarded as highly diverse since the members of the library bear little relationship (i.e. not prepared through parallel synthesis) and are assembled from random sources (chemical suppliers, academic labs, etc.). It is often stated that these are the ideal types of screening libraries since the high diversity is viewed as a benefit in locating hits. One important caveat with these libraries (ACD, NCI) is whether the compounds are actually available. Groups that have relied on these compound sources frequently report difficulties in actually obtaining all of the desired compounds that are identified though their screening efforts.

In contrast to these diverse libraries, focused or biased libraries can be used when some knowledge of an active ligand structure is available [34]. These are typically second-generation libraries constructed around a hit located from the larger, more diverse screening libraries. Libraries of this type tend to be much smaller (hundreds to thousands of compounds) and are built around a single, central scaffold. These libraries are lower in diversity with regard to all

chemical space but represent a greater depth of diversity in a local region of structure space. Focused libraries often take on the familiar attributes of traditional medicinal chemistry programs as a lead compound is heavily analoged by altering the appendages around a central core structure. There have also been a variety of attempts to generate smaller, focused libraries by matching the library members to a particular pharmacophore [35], produced either from a known ligand or from the mapping of a binding site.

In recent years the focus on drug-like or lead-like compounds has emerged as a primary concern in the early stages of drug discovery. It has been appreciated for some time that many drug candidates fail because of poor pharmacokinetic parameters. Several guidelines, most famously Lipinski rules [36], have been developed to relate structures of compounds to ADME (absorption, distribution, metabolism and excretion) properties [25]. Critical parameters include cLogP, molecular weight, rotatable bonds, H-bond donors/acceptors, Caco-2 permeability, etc., many of which are now evaluated computationally to filter virtual libraries to remove compounds with poor druggability from consideration very early in the drug discovery process [37].

TARGET SELECTION

The structure of the target macromolecule is usually obtained by one of three techniques: X-ray crystallography, nuclear magnetic resonance (NMR) or homology modeling from a previously determined structure. Crystal structures are the most common choice, but the method of structure determination is often a consequence of the ability of the protein to crystallize, the availability of instrumentation and the expertise of the investigator.

X-ray crystal structures are an excellent source for the target structure. A large number of crystal structures of different macromolecules are available from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Database. The majority of deposited coordinates originate from soluble proteins since these are most amenable to crystallization. The number of membrane protein structures is steadily increasing with improvements in membrane protein purification and crystallization techniques. One specific family of proteins, the G-protein coupled receptors (GPCRs) represent excellent drug targets, but are difficult to crystallize due to their membrane-spanning component. The structure of rhodopsin [38] was recently determined and is often used as a starting point for homology modeling of other GPCRs.

A good crystal structure for structure-based drug design meets certain statistical metrics achieved during the model building and refinement process. Higher resolution data, that is, data with diffraction intensities extending to at least 2 Å (where lower numbers indicate higher resolution and the distance between sampling planes) yield electron density maps that are better defined and models that are often more accurate. The success of the model building and the refinement is primarily judged based on the agreement with the electron density map (measured by the R-factor and free R-factor), the standard deviations of the bond lengths and angles from known values and agreement with good biophysical sense. Temperature factors indicate the relative motion of the atom by a number in $Å^2$ of the transcribed circular space. Atoms with higher temperature factors have less precisely determined coordinates than atoms with lower temperature factors. Structures suitable for drug design usually have diffraction intensities that extend beyond 2.5 Å, allowing the atoms of the model to be accurately placed. The R-factor should be lower than 0.25 and the R-free should be lower than approximately 0.28. The standard deviations from known bond lengths and angles should be no more than 0.001 Å and 3 °, respectively. The temperature factors of atoms of interest, those in the binding site and any water molecules that will be maintained during the procedure, should be at the average or lower than the average temperature factor for the entire molecule.

There are often several crystal structures deposited in the database for a single protein. A comparison of the apo structure, without ligands bound, and a structure of the target with ligands bound or a comparison of multiple ligand-bound structures can reveal some of the conformational changes associated with ligand binding. Superimposing many structures of the target provides a more comprehensive view of the ensemble of possible conformations that the target can capably assume in solution. The ensemble, therefore, provides a more comprehensive view of the possible conformations that could be encountered in solution. It is possible to dock libraries of compounds against the entire ensemble and, in fact, allow the ensemble to simulate the flexibility of the protein [39].

Initial target structures can also be determined using NMR methods. A great advantage to using NMR is that the protein, although highly concentrated, is in the solution phase and is not subject to the forces of crystal packing. Usually multiple NMR spectra (^{1}H, ^{13}C, ^{15}N) are needed for structure determination and the acquisition of labeled protein can be very difficult and costly. A well-determined NMR structure has few violations (observed resonances that are unexplained in the model), low standard deviations with ideal bond lengths and angles (same tolerances as reported for crystal structures) and a high total number of NOE restraints per residue. An NMR structure with a low standard deviation between individual members of the ensemble and the average is often more precisely determined than one with larger standard deviations in the ensemble [40]. Since NMR structure determination results in multiple models, one question that arises is whether to use the average structure, for which no experimental information may exist, or to use one or all of the individual structures of the ensemble. Again, multiple structures can be used in docking and can simulate the range of flexibility of protein residues.

SAR by NMR [41] is an efficient technique to rapidly screen new ligands and examines the changes in resonances of residues known to interact with an initial ligand. Briefly, chemical shift data for a target protein are measured and a structure of the protein is determined. The protein is then incubated with a ligand at a particular concentration and new chemical shifts are measured. If the chemical shifts of residues in the ligand-binding pocket are perturbed from those measured with the protein alone, the ligand is assumed to have bound the site. If the chemical shifts are not

perturbed, then increasing concentrations of the ligand are added to determine the concentration at which the shifts are evident. This technique was very effectively used to screen ligands that bind to FKBP [41].

Homology modeling is another method that results in a structure of the target. This method is less frequently used, however, because of the lack of experimental information for the target. The quality of the homology model is judged based on adherence to ideal values for bond lengths and angles and the number of template structures, since a greater number of template structures encompass a greater number of possible structures of members of that family.

BINDING SITE IDENTIFICATION

A crucial step in preparing the target for virtual screening is the identification of the proper ligand binding site. Ideally, the ligand binding site is well-defined and capable of specifically binding a small molecule that will modulate its function. In many cases, such as enzymes, the targeted ligand binding site is well-known, in other cases, such as small molecules that disrupt protein:protein interactions, it is more obscure.

Enzymes represent a large percentage of the validated drug targets. Identifying the ligand binding site of an enzyme to be used for virtual screening is usually quite easy since the active site has already evolved to bind small molecule ligands. Often, the crystal structure of an enzyme bound to its substrate, cofactor, or both, is used as the initial starting point. The small molecule ligands are removed from the structure and several compounds are screened against the active site structure. The apo enzyme structure, if available, is also a good potential starting point since this structure may represent a more accurate picture of the state of the enzyme in the absence of ligands. However, several enzymes undergo conformational changes upon ligand binding [42, 43] and it is often difficult to reproduce these changes *in silico*. Allosteric sites, sites removed from the active site but capable of binding a small molecule that affects the conformation and activity of the enzyme, also represent good choices for target binding sites for virtual screening. In the case of p38 MAP kinase, a small molecule inhibitor binds an allosteric site with a subnanomolar K_i [44].

There are several excellent examples of virtual screening against enzymes, using the active site as the target ligand binding site. HIV protease represents one of the most well-known examples [45-49], there are also many examples involving dihydrofolate reductase (DHFR) [50, 51]. Thymidylate synthase (TS) is an interesting example since many of the efforts for virtual screening have targeted the cofactor, not substrate, binding site. The substrate of TS, deoxyuridine monophosphate (dUMP), is involved in many other cellular processes. Small molecules that target this site could potentially target several other sites on different macromolecules, creating toxicity problems. The cofactor, 5,10-methylene tetrahydrofolate, however, is not involved in other cellular processes and therefore represents a better choice for targeted ligand binding.

Receptor ligand binding sites are also reasonably well-defined. The nuclear hormone receptor ligand binding domains have been targeted in the search for therapeutics for several diseases including hyperthyroidism, diabetes and cancer. The structures of several ligand binding domains of nuclear hormone receptors bound to agonists and the structure of the estrogen receptor bound to tamoxifen, an antagonist [52], have been determined. These structures can be used as starting points to identify new agonists and antagonists.

IDENTIFYING PREVIOUSLY UNKNOWN LIGAND BINDING SITES

In some cases, the ligand binding site to be used for virtual screening against a target protein is not known *a priori*. For example, it is often difficult to characterize the site of interaction between two proteins and as such it can be very difficult to locate a site to disrupt the interaction. Three algorithms that identify ligand binding sites will be discussed.

GRID [53] is one of the first algorithms developed to identify ligand binding sites on proteins. GRID calculates the interaction energy, using a simple force field with terms for the Lennard-Jones potential, electrostatics and hydrogen bonds, between a probe molecule such as water, a methyl group, an amine nitrogen, carboxyl oxygen and hydroxyl and a protein. The calculated interaction energies are displayed as contour surfaces, allowing the user to adjust the contours to eliminate non-specific interactions. Later improvements in the algorithm [54] extended the ability of GRID to handle probes capable of three or four hydrogen bonds. This extension is important in the treatment of water molecules that are capable of donating two and accepting two hydrogen bonds.

GRID was especially useful in the development of neuraminidase inhibitors. Neuraminidase is a membrane-bound glycoprotein of the influenza virus and is responsible for destroying the hemagglutinin receptor. Neuraminidase inhibitors have been sought for the treatment of influenza since inhibition of the enzyme leads to slower rates of viral attachment to the cell. The structure of neuraminidase [55] bound to an initial inhibitor and transition state analog, Neu5Ac2en, Fig. (**2**), shows that modifications to Neu5Ac2en may improve the potency of the inhibitors. Using GRID, favorable binding sites for carboxylate and amino nitrogen probes were predicted. Substitution of a hydroxyl group at the 4-position of the pyranose ring of Neu5Ac2en by an amino or guanidinyl group was predicted to increase interactions with a nearby Glu 119, Fig. (**2**). In fact, 4-amino-Neu5Ac2en and 4-guanidino-Neu4Ac2en bound to neuraminidase with K_i 50 nM and 0.2 nM, respectively. X-ray crystal structures of the complexes showed that the amino and guanido groups bound to locations close to those predicted with only small changes in the interactions between the guanidinyl group and the glutamic acid residues [56].

The algorithm, Multiscale [57], addresses the question of where a drug candidate molecule will bind to a 3-dimensional structure of the protein. The brute force approach used by GRID, in which a small probe group is used to locate a binding site, is not practical for a larger ligand or the examination of an entire protein. Glick, *et al.*

Fig. (2). Neuraminidase bound to guanidinoNeu5Ac2en, the guanidino group was predicted by GRID.

estimate that, for a 35 atom ligand, a 3,500 atom protein and reasonable search parameters, 1.5 x 10^{14} nonbonding interactions would have to be calculated to scan the possible docking configurations. Instead, they support a method called the multiscale approach derived from signal and image processing.

The multiscale approach relies on a scaling operator that removes the fine levels of detail to emphasize larger features. A k-means clusters algorithm generates a series of ligand models, each one with an increasing level of detail. An example using nevirapine, the HIV reverse transcriptase inhibitor, shows that the first cluster is a single point at the mean position of the ligand. The second cluster contains two feature points, each removed to the furthest distance from the mean position. The final cluster contains nine feature points at different positions around the ligand.

The first cluster, consisting of the single feature point, is docked to the protein and low-energy configurations that have lower energy than a given threshold are retained. The second cluster, with two points at either end of the ligand, is now docked and low-energy configurations are retained. At this stage, any potential sites that are not large enough to accommodate a ligand with the axis defined by the two points are rejected. With increasing levels of filters generated by each cluster, the entire series of clusters for a particular ligand is docked against the target protein. Surviving configurations reveal the region where the ligand may dock successfully.

The multiscale algorithm was tested using several protein:ligand complex structures from the PDB. In all cases, except one, the distance between the centroid of the ligand in the crystal structure and the centroid of the ligand in the predicted configuration fell between 0.62 and 1.68 Å. The test cases demonstrated success under some very difficult circumstances: phosphocholine, a very small molecule and therefore potentially capable of binding several sites, was correctly predicted to bind McPC-603, the immunoglobulin Fab-phosphocholine complex, and three non-nucleoside inhibitors were correctly predicted to bind HIV reverse transcriptase, despite the disparity of size between the ligand and protein. Additionally, the algorithm can handle cases in which the protein is flexible since it was able to predict binding modes for ligands against structures of proteins even when they adopted different conformations.

Glick, *et al.* used the multiscale approach to predict a key binding site on the anthrax protective antigen heptamer. In order for anthrax toxin to be lethal, the protective antigen (PA) must undergo proteolytic cleavage, form a heptamer and bind the edema factor (EF) and lethal factor (LF). It was recently reported [58] that the YWWL tetrapeptide inhibited the binding of the PA to EF and LF. Identifying the tetrapeptide binding site on the PA heptamer, Fig. (**3**), allows the virtual screening of 3.5 billion small molecule compounds for candidates that will bind to the PA heptamer-binding site, using a distributed computer screen-saver project [59].

Small molecule compounds are often sought to disrupt key protein:protein interactions. The structure of the complex of the two proteins or the structure of a peptide from one protein bound to the other allows the identification of the protein:protein interaction site. Proteins often associate using “hot spots”, compact regions responsible for most of the affinity of the interaction. The region for targeted virtual screening can be the “hot spot”, identified by structures of the complex or by site-directed mutagenesis [60] or the region resulting after removal of one of the proteins or the peptide from the structure. This strategy has been used successfully in the identification of inhibitors that disrupt the association of Bcl-2 and other anti-apoptotic

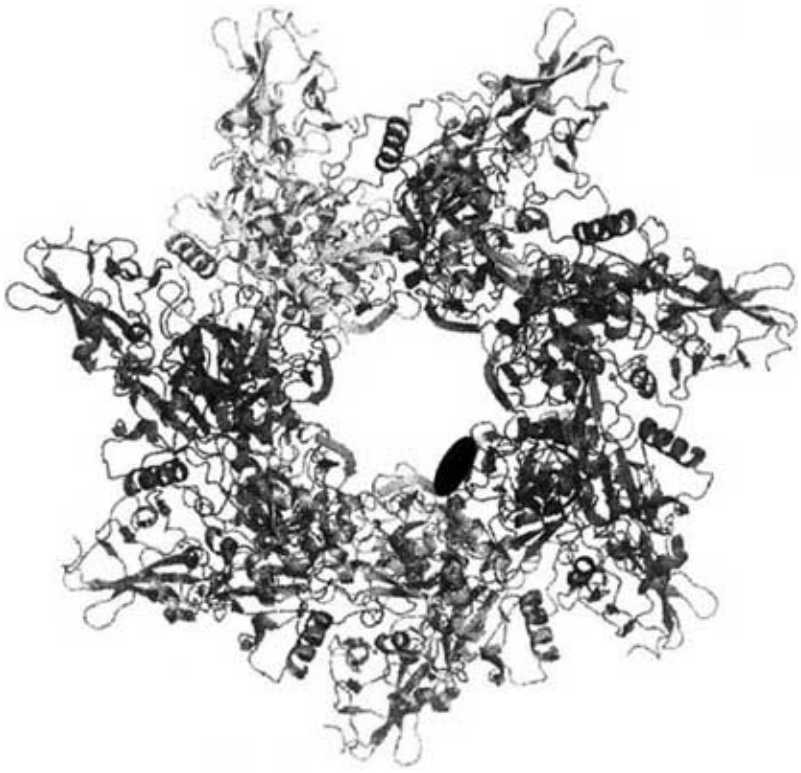


Fig. (3). Anthrax protective antigen heptamer. The binding site for a tetrapeptide is noted.

proteins [61] using a homology model of Bcl-2 bound to a peptide from the Bak protein. The Bcl-2 case study is discussed in greater detail in the Case Studies section of this review. In another example of this method, NMR revealed the binding site between IL-2 and the IL-2 receptor (IL-2Rα). Ro26-4550 was designed as a peptidomimetic of the IL-2 portion of this site. A crystal structure later revealed that Ro26-4550 bound to IL-2 itself, at the IL-2Rα binding site [62]. Since Ro26-4550 did not have any cell-based activity, it was abandoned, but more potent derivatives based on Ro26-4550 were designed using a fragment-based approach [63].

DOCKING

Once the virtual library is created and the target is prepared, including the specification of the ligand binding site, the library must be docked into the target site and evaluated for goodness-of-fit. The two stages represented in this step are 1) docking – the search for the conformation and configuration of the ligand in the binding site and 2) scoring – the evaluation of the interaction energy between the target and ligand. Many previous reviews have extensively covered the aspects of docking and scoring [10, 64-67] . This review will not serve to reiterate that material, but will cover some of the essential questions raised in considering a docking problem. Specifically, the flexibility of ligands during docking, the treatment of ligands as fragments and the flexibility of the target will be discussed.

Ligand Flexibility During Docking

Ligands may adopt different configurations with different proteins, therefore allowing ligand flexibility is important in the docking process. A library of pre-calculated ligand conformers may be generated and docked into the target, each conformer treated essentially as a new ligand. Any algorithm that reliably generates accurate ligand conformers could be used to generate this library. An advantage to this method is that the ligand conformers only have to be generated once. DOCK [68], EUDOC [69] and FLOG are examples of algorithms that can use a pre-generated ensemble of ligand conformers.

Alternatively, the ligand could be allowed to be flexible in the ligand binding site during the docking process and the energy of interaction with the target assessed on the fly. Several methods including incremental construction, Monte Carlo generation and genetic algorithms carry out this task.

During incremental construction, rigid portions of the ligand, or "anchors", are docked first, followed by the flexible portions. An early implementation of this method [70] broke the ligand into a small set of rigid fragments that were docked separately into the site, using DOCK, later fused and finally, energy minimized. DOCK 4.0 extends this method by docking the anchor and growing the flexible portions of the ligand [71]. Rarey, *et al.* [72] presents an incremental construction algorithm that samples the conformation space of the ligand and uses a hierarchical system for placing the flexible pieces of the ligand. The rigid portion of the ligand, or the base, is placed first, maximizing interactions between the fragment and the protein. Many alternatives for the placement of the flexible portions of the ligand, starting with those nearest the base, are considered and only those with favorable energies are kept for sequential rounds during which additional flexible portions are added. This incremental construction algorithm is implemented in FlexX [73].

In Monte Carlo implementations, the ligand undergoes random changes in translation, orientation and torsion angles. The structure is minimized and the energy is evaluated, forcing a decision to keep the new configuration or to reject it and start with a new random change. AutoDock [74], a method devised by Caflisch [75], MCDOCK [76] and Prodock [77] all use Monte Carlo methods during the docking procedure.

Genetic algorithms are another method to generate ligand conformers during docking. Genetic algorithms select the most "fit" conformers to continue to the next generation, allowing diversity by "mutations" during each generation. Jones, *et al.* [78] wrote the algorithm, GOLD, which allows full acyclic and partial cyclic flexibility of the ligand and partial flexibility of the protein receptor. The GOLD program was tested on 100 cases of protein:ligand complexes derived from the PDB and achieved a 71 % success rate in reproducing the experimental binding mode.

Treating Ligands as Fragments

Docking fragments, for example functional groups such as phenyl rings, hydroxyl, methyl or carboxyl groups, is another approach toward ligand flexibility. LUDI [79] exploits this advantage by using a library of small molecule fragments that bind to the target in an independent fashion and are subsequently joined to form a single entity. In a recent application of the LUDI approach, inhibitors of bacterial DNA gyrase were designed and developed [80]. DNA gyrase is a bacterial-specific topoisomerase involved in DNA replication, transcription and recombination. Using the structure of the ATP binding site on one of the two subunits of DNA gyrase (B) and a pharmacophore model that emphasized hydrogen bonding to two key groups and lipophilic interactions with other residues in the active site, Bohm *et al.* used LUDI to screen the ACD and a Roche compound inventory, a total of approximately 350,000 compounds, to find novel inhibitors of DNA gyrase [80] LUDI was used to dock small "needles", so-named because of their ability to reach deep into pockets of the target site. The many hits that resulted from the search were filtered according to molecular weight, to diversity properties according to Tanimoto coefficients and manual selection. Six hundred compounds from the virtual screening were tested in biological assays that examined ATPase activity. These assays included a number of control enzymes capable of ATP hydrolysis. Hits were subjected to validation tests including supercoiling assays, surface plasmon resonance assays, elaboration of the initial SAR, analytical ultracentrifugation experiments, heteronuclear $^{1}H/^{15}N$ correlation NMR spectroscopy and X-ray analysis. These validation experiments narrowed the list of potential candidates to seven structural classes. Primarily, the X-ray structures and secondarily, the preliminary SAR data of the validated needles and SAR data for previously known inhibitors were employed to guide the optimization of the needle hits and led to inhibitors that are potent and ten times more active than the best known inhibitor, novobiocin.

CombiSMoG is an algorithm that incorporates the principles of combinatorial synthesis, a knowledge potential and a Monte Carlo ligand growth algorithm [81]. The potential is based on interactions observed in a large (1,000) representative set of protein:ligand complexes in the PDB. Ligands are generated in the active site from 100 common functional groups in the program's virtual combinatorial library. Steps in the growth algorithm represent additions of random fragments from the library. The energy of the ligand after each incremental addition of a new fragment is assessed and evaluated according to a Boltzmann criteria, biasing the choice toward low-energy complexes.

In an elegant application of CombiSMoG, inhibitors were sought against human carbonic anhydrase, an enzyme targeted in the treatment of glaucoma. A starting fragment, *para*-substituted benzene sulfonamide, with a well-defined binding orientation, was chosen for design. Ligands were grown from one of the carboxamido hydrogens of the benzene sulfonamide. The five top-scoring ligands were energy minimized and the R and S isomers, both of which had low CombiSMoG scores, were synthesized. The R and S antipodes had 30 pM and 230 pM inhibition constants, respectively, making the R isomer the most potent human carbonic anhydrase inhibitor known. Crystal structures of the inhibitors with the enzyme verify the predicted binding modes.

Target Flexibility

Many proteins respond to the introduction of a ligand with structural changes. The computational time needed to calculate both protein and ligand flexibility, especially while docking a large library, can be quite high and therefore, many docking programs assume a rigid protein model. For some proteins that undergo significant conformational changes, this assumption can create hit lists that do not reflect the results of biological assays. The issues inherent in the "rigid receptor problem" have been extensively reviewed [43, 65, 82].

Multiple structures of a target can be compared in order to decide whether a particular protein needs to be treated as a flexible entity. Davis and Teague [43] review several cases, comparing different crystal structures of many targets, and show that significant conformational changes are often governed by the additive conformational changes of hydrophobic portions of a protein. Biochemical data may also indicate that conformational flexibility is an important component of ligand binding for a particular target.

Ensembles of target structures may be used to simulate protein flexibility. Methods using multiple crystal or NMR structures [39], molecular dynamics [83] or side-chain rotamers [42, 84] have been used to simulate the range of conformations an active site may adopt. Docking algorithms such as SLIDE [84], FlexE [85] and MCSA-PCR [86] use ensembles to allow some degree of protein flexibility during docking.

Scoring

The scoring process evaluates and ranks each ligand pose in the target site. Scoring can involve a simple forcefield that predicts the enthalpy of binding using van der Waals repulsive and dispersive functions (often the Lennard-Jones 6-12 potential), electrostatics and hydrogen bonds. More accurate scoring functions often include a correction for the

solvation of the ligand and target and the dielectric constant of the medium, either determined as a constant or determined continuously in a distance-dependent manner such as with the Poisson-Boltzmann or generalized Born equations. In the most rigorous form, scoring can calculate the free energy perturbation. Each step from a simple scoring function to a more complex scoring function involves a greater computational price. Often, when screening a large virtual library, a simple scoring function suffices for an early cutoff and more precise scoring functions are used at later stages.

There are several different scoring functions associated with different docking/scoring algorithms and these have been extensively reviewed [10, 87]. In general, many investigators have found that rescoring the top hits from docking using several different scoring functions, a process called consensus scoring [88], is valuable. Those hits appearing at the top of multiple lists are then selected for further investigation.

Charifson, *et al.* analyzed two different docking methods and thirteen different scoring functions for ligands targeting p38 MAP kinase, inosine monophosphate dehydrogenase and HIV protease. Consensus scoring provided a significant reduction in the number of false positives that arose during docking. Perola, *et al.* [89] used the same three systems (p38 MAP kinase, IMPDH and HIV protease) to compare three docking functions (Glide, GOLD and ICM) with respect to their ability to reproduce a crystallographically observed ligand orientation and three different scoring functions with respect to their ability to discriminate between actives and inactives. They found that energy minimization and reranking of the top poses overcomes some of the limitations of the individual docking programs. Their results confirm that the choice of the best scoring function is system-dependent but that consensus scoring reduces the number of false positives.

Evaluation

After docking a virtual library, scoring the ligands in that library and, possibly, rescoring and reranking those ligands, some decision must be made to prioritize the ligands for purchase or synthesis and later, biological testing. Many groups use a means of visual evaluation at this stage. Criteria often used in visual inspection are: formation of key hydrogen bonds or electrostatic interactions, surface complementarity and the stability of the configuration of the ligand in the target site compared to conformational preferences of the free ligand. Other criteria may be imposed at the evaluation stage as well, including the ability to synthesize or purchase the intended ligand or the ability to synthesize derivatives and second generations of the hit. Additional filters may be imposed for the drug-like properties of the hit. Teague, *et al.* [90] suggest that it is easier to optimize small (MW 100-350) and less lipophilic compounds to increase potency than to optimize larger compounds to increase pharmacokinetic and metabolic properties. These authors examined historical data for the development of drug molecules from leads and found that the most effective strategy for improving the potency of a lead was the addition of hydrophobic moieties on the

Table 1. Summary of Case Study Features

Case study	Structure Source	Library Source	Pharm. constraint?	Library filtering?	Hit-to-Lead?	Docking Program	Ref.
CDK	X-ray	Maybridge	N	N	Y	LIDAEUS	91
Carbonic anhydrase	X-ray	Maybridge and Leadquest	Y	Y	N	FlexX	93
TS	X-ray	ACD	N	Y	Y	DOCK	96
Cathepsin D	X-ray	Focused	N	N	Y	DOCK	98
Aldose reductase	X-ray	ACD	Y	Y	N	FlexX	99
Chk-1	X-ray	AstraZeneca	Y	Y	N	FlexX	100
Rm1C	X-ray	Focused	Y	N	N	FlexX	101
Bcr-Abl	X-ray	Commercial	N	Y	N	DOCK	103
Thyroid receptor	Homology Model	ACD	N	Y	Y	ICM	104
Retinoic acid receptor	Homology Model	ACD	N	N	N	ICM	105
NK-1	Homology Model	Various	Y	Y	N	FlexX	107
K^+ channel	Homology Model	China Natural Product Database	N	N	N	DOCK	108
Bcl-2	Homology Model	NCI	N	Y	N	DOCK	61
Rac-1	X-ray	NCI	N	N	N	FlexX	110
p56 Lck	X-ray	Various	N	Y	N	DOCK	111

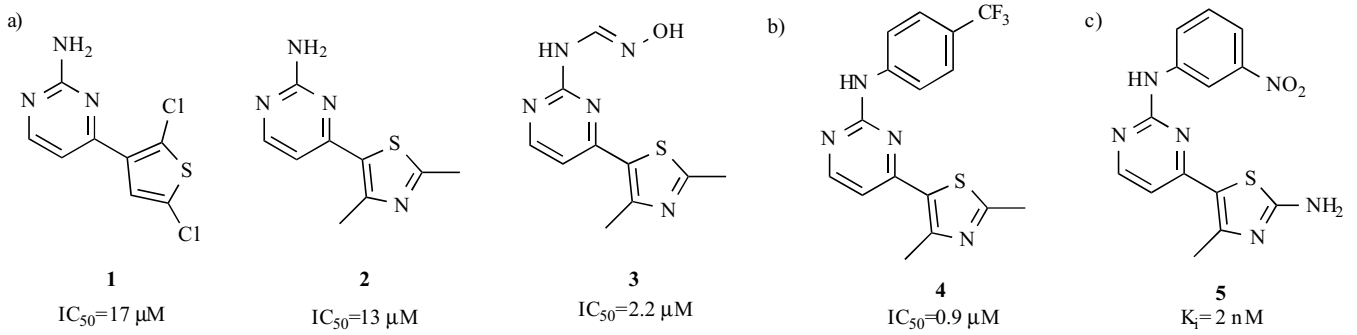


Fig. (4). Structure of CDK inhibitors. (a) initial hits from VLS (b) initial optimized lead structure (c) lead structure found by medicinal chemistry optimization. All values are measured against CDK2.

starting scaffold. By necessity, this process increases the molecular weight and clogP value of the final optimized structure, therefore the smaller and less lipophilic compounds should be chosen from the hit list in order to arrive at optimized structures that still maintain physical properties consistent with drug-like candidates.

RECENT CASE STUDIES

In the second half of the review, fifteen case studies involving VLS in lead discovery will be examined in detail. These case studies have been grouped into three broad classes: catalytic site inhibitors, ligand binding domains and inhibitors of protein:protein interactions. The discussion of each case study will encompass the methodology applied, any filtering applied to the VLS library, the results of the VLS study and any further lead optimization. Table (**1**), offers a summary of the key features of each case study.

Catalytic Site Inhibitors

CDK

The cell cycle contains various checkpoints to ensure that the integrity of the daughter cells is maintained throughout the mitotic process. Some transformed cells override these checkpoints in order to maintain a high rate of proliferation. Cyclin-dependent kinases (CDKs) are one of the key players in checkpoint regulation and it has been shown that inhibition of these enzymes can have anti-cancer effects. A group in the United Kingdom recently used a VLS approach to the discover novel inhibitors of CDK [91]. These workers utilized LIDAEUS, an in-house docking program that generates a cubic grid to define the binding pockets. The candidate compounds were docked into the active site of CDK2 generated from the X-ray crystal structure of CDK2 complexed with staurosporine. A 3D virtual library was generated from ~50,000 commercially available compounds from the Maybridge database. Screening of the database to yield the top 28 compounds gave a 29% hit rate *in vitro* as opposed to a 7% hit rate with 28 compounds chosen randomly from the same database. One of the families of compounds found in the screening was the 2-amino-4-heteroaryl-pyrimidines (**1-3**), which showed no cross-reactivity against other kinases such as PKCα or ERK-2, Fig. (**4**).

These three compounds found in the screen were utilized as a starting point for the design of a superior inhibitor. Modeling of these compounds in the crystal structure suggested that an aromatic ring bearing an electron donating substituent linked to the pyrimidyl amino group would generate additional interactions in the binding pocket not utilized with these three inhibitors. Compound (**4**) was designed to include this group and showed an improved activity of 900 nM against CDK2. X-ray structures of each of the four ligands (**1-4**) bound to CDK2 were determined and revealed that each of these compounds bind the ATP binding site in a similar orientation as that predicted in the docking Fig. (**5**). In a follow up study [92], these compounds discovered through VLS were used to initiate a traditional medicinal chemistry effort to find CDK inhibitors with higher levels of potency. This led to the preparation of inhibitor (**5**) with a K_i value of 2 nM. This compound stopped progression through the cell cycle in A549 cells.

Human Carbonic Anhydrase

Using a multiply-filtered database and a staged docking approach, Gruneberg, *et al.* [93] discovered novel inhibitors of human carbonic anhydrase (hCAII), an enzyme target in the treatment of glaucoma. High-resolution crystal structures of hCAII bound to several ligands have been determined and a comparison of 24 of these structures reveals that the binding site is relatively constant. The initial database for docking contained 90,000 compounds from the Maybridge and LeadQuest libraries that satisfy the Lipinski rules, including 35 known inhibitors of hCAII for calibration. Several filters were applied to reduce the number of compounds in the virtual library. The first filter selected ligands that bear functional groups that could bind a zinc atom. A second stage of the first filter used “hot spots” of binding derived from LUDI, GRID, SuperStar and DrugScore [53, 94, 95] - these were translated into a protein-derived pharmacophore model, leaving 3,300 compounds. The second filter excluded compounds that were not similar to known reference ligands. Finally, 100 compounds were docked using the program FlexX [73], and visually inspected. Four water molecules, found in every crystal structure of the enzyme bound to inhibitors, were kept in place during the docking procedure. Based on the visual inspection, which included assessing the degree of occupancy of the amphiphilic binding pocket next to the zinc atom, the number of rotatable bonds in the ligand, quality of the overall binding conformation and formation of hydrogen bonds to two key residues in the active site, thirteen compounds were selected for testing. Three of the 13 compounds are subnanomolar inhibitors, one is a nanomolar

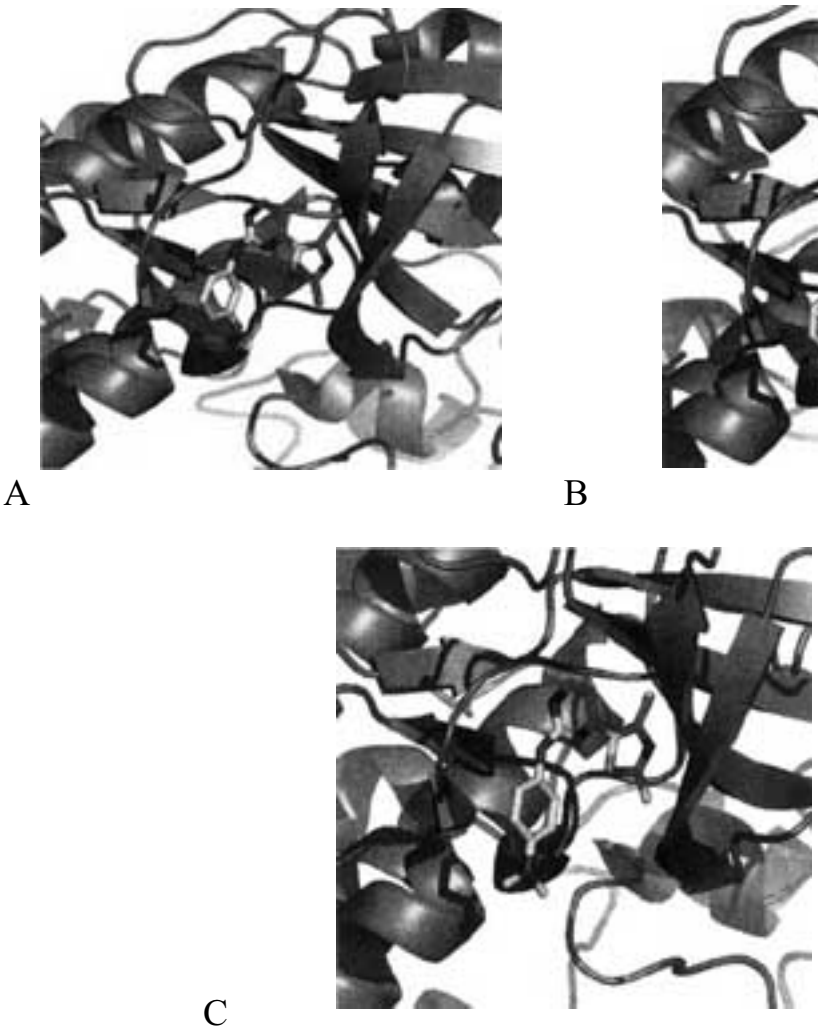
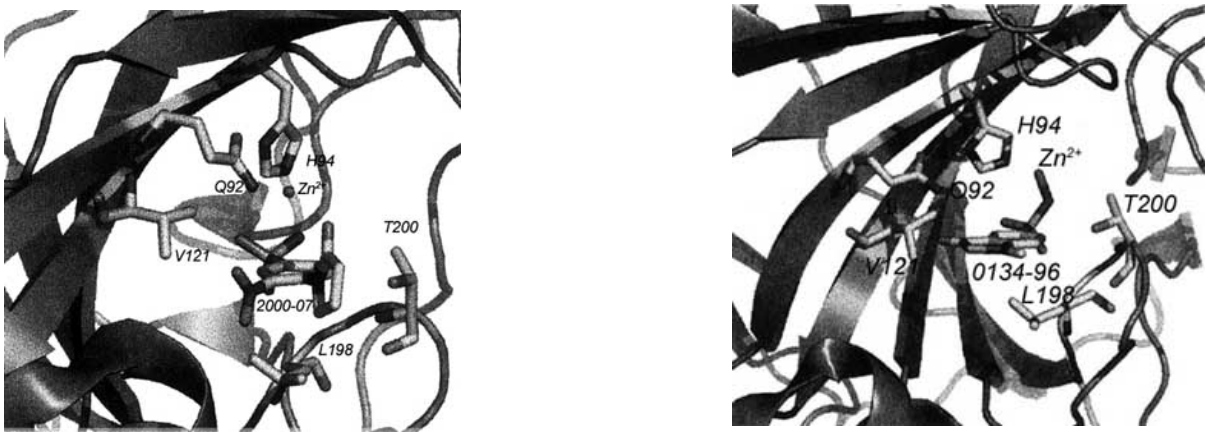


Fig. (5). Bound structures of inhibitors **1, 2** and **4** with CDK.

inhibitor and seven are micromolar inhibitors. Crystal structures of two discovered inhibitors bound to with the enzyme verify the predicted docking poses, Fig. (**6**).

Thymidylate Synthase

Thymidylate synthase (TS) plays a critical role in the folate biosynthetic pathway, responsible for the production of deoxythymidine monophosphate (dTMP). As such, its inhibition leads to the cessation of production of one of the key nucleotides for DNA synthesis, making TS a validated anticancer and antipathogenic drug target. In efforts to discover novel scaffolds for TS inhibitors that are compatible with solid-phase in-parallel derivatization and would bind the folate site [96], compounds from the ACD (153,516 total) were docked into the active site of *L. casei* TS with the program, DOCK. A list of the top 400



A B

Fig. (6). Inhibitors of human carbonic anhydrase bound to the enzyme.

compounds, scored for van der Waals interactions and electrostatic interaction energy and corrected for ligand desolvation, was generated. Five compounds that show a number of polar interactions and the opportunity for later parallel synthesis were chosen for *in vitro* enzyme assays. Dansyl hydrazine (**6**), Fig. (**7a**), was shown to inhibit TS competitively with an IC_{50} of 439 μM.

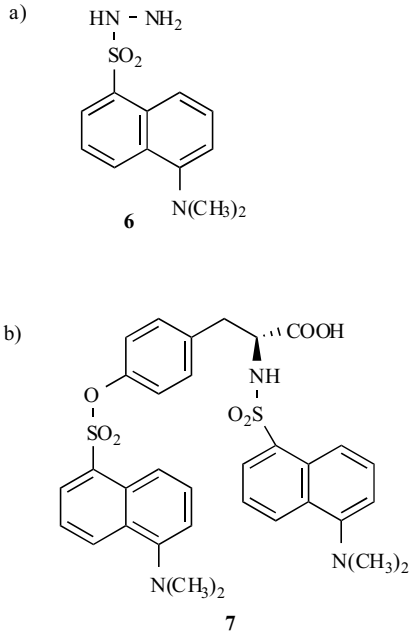


Fig. (7). a) Dansyl hydrazine, discovered from the original DOCK run and b) didansyltyrosine, the optimized lead compound.

Seven dansyl amino acid analogs were tested, resulting in the selection of O-dansyl-L-tyrosine with an IC_{50} of 163 μM and K_i of 65 μM. A small library of 33 amino derivatives of O-dansyl-L-tyrosine was synthesized, modeled and scored in the active site and a new derivative, didansyl tyrosine (DDT, **7**), Fig. (**7b**), was selected for testing. DDT has a K_i of 1.3 μM and is 30-fold more active against *L. casei* TS than human TS.

A crystal structure of DDT and dUMP, the natural substrate, with *E. coli* TS reveals an unexpected binding mode for the inhibitor [97]. DOCK correctly predicted the binding pose for the O-dansyl and phenyl rings of DDT, but the crystal structure shows that DDT binds more deeply in the active site than DOCK predicted. Much of the new binding pose can be ascribed to the fact that TS undergoes significant protein rearrangements on binding DDT. The interactions seen in the crystal structure explain the binding affinity of DDT and the specificity is explained by the interaction with bacterial-specific residues over residues in human TS.

The selection of dansyl hydrazine from the ACD, using DOCK, and the further elaboration to create DDT used an elegant combination of computer-based scaffold selection and in-parallel synthetic derivatization. The improved binding affinity of dansyl derivatives, as well as improved specificity against the human form of the enzyme, validates this methodology. The crystal structure of the inhibitor:enzyme complex revealed that, without accounting for protein flexibility, proper docking poses may not be discovered.

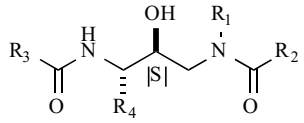
Cathepsin D

Combinatorial chemistry and structure-based drug design are two modern methods for drug discovery. Kick, *et al.* [98] demonstrated a procedure to combine the two methods and applied their procedure to discover non-peptidic

Fig. (8). Crystal structure of didansyl tyrosine in *E. coli* TS.

inhibitors of cathepsin D, an aspartyl protease implicated in cancer and Alzheimer's disease.

A crucial step in library design is the selection of compounds to synthesize. Kick, *et al.* tested two theories for reducing the synthetic effort: 1) diversity-based approaches that attempt to maximize the sampling of chemical and biological properties and 2) directed approaches that select compounds that are predicted to have favorable binding affinity to the target. Two reduced libraries were constructed from a large virtual library of potential inhibitors based on the (hydroxyethyl)amine isostere (**8**).



8

Fig. (9). The (hydroxyethyl)amine isostere (S epimer) as a scaffold for the design of cathepsin D inhibitors. Diversity was incorporated at positions R_1, R_2, R_3 and R_4.

The directed library was constructed using a structure-based screening process and the program, CombiBuild, as well as the structure of cathepsin D. The diverse library was designed to maximize the variety of functional groups and structural motifs. The directed and diverse libraries, each containing 1,000 compounds, were synthesized and assayed. The hit rate for activity (inhibiting the enzyme to ≥50 %) at 1 μM was 67 compounds for the directed library and 26 for the diverse library. The hit rate for activity at 330 nM was 23 compounds for the directed library and three for the diverse library. Seven compounds from the directed library and one compound from the diverse library were active at 100 nM. In general, the diverse library gave hits that were 3-4 times less potent than those from the directed library. In order for the diverse library to achieve the same results as the directed library, approximately ten times the number of compounds would have to be synthesized.

Aldose Reductase

Kraemer and co-workers [99] recently disclosed a virtual screening strategy for the identification of inhibitors of human aldose reductase. This enzyme is responsible for the reduction of aldo-sugars to their corresponding alcohols. Inhibition of this enzyme is viewed as a therapeutic strategy to alleviate some of the symptoms associated with chronic diabetes. An ultrahigh resolution structure of this enzyme (0.66Å) in complex with IDD594, a potent inhibitor, was recently disclosed and provided an excellent starting point for the discovery of novel classes of inhibitors. The catalytic site of the enzyme is located at the center of a $(\beta/\alpha)_8$ TIM-barrel and contains a nicotinamide co-factor. The ultra high-resolution structure permits a detailed examination of the binding mode of the inhibitor including the protonation state of atoms. The inhibitor contains an ionized carboxylic acid which makes an electrostatic interaction with C4 of NADP+ and functions as an H-bond acceptor from three different residues: Tyr48, His110 and Trp111. Kraemer *et al.* utilized this structural information to search for new inhibitors of the enzyme. Only a single conformation of the protein and a non-flexible binding pocket was used. This

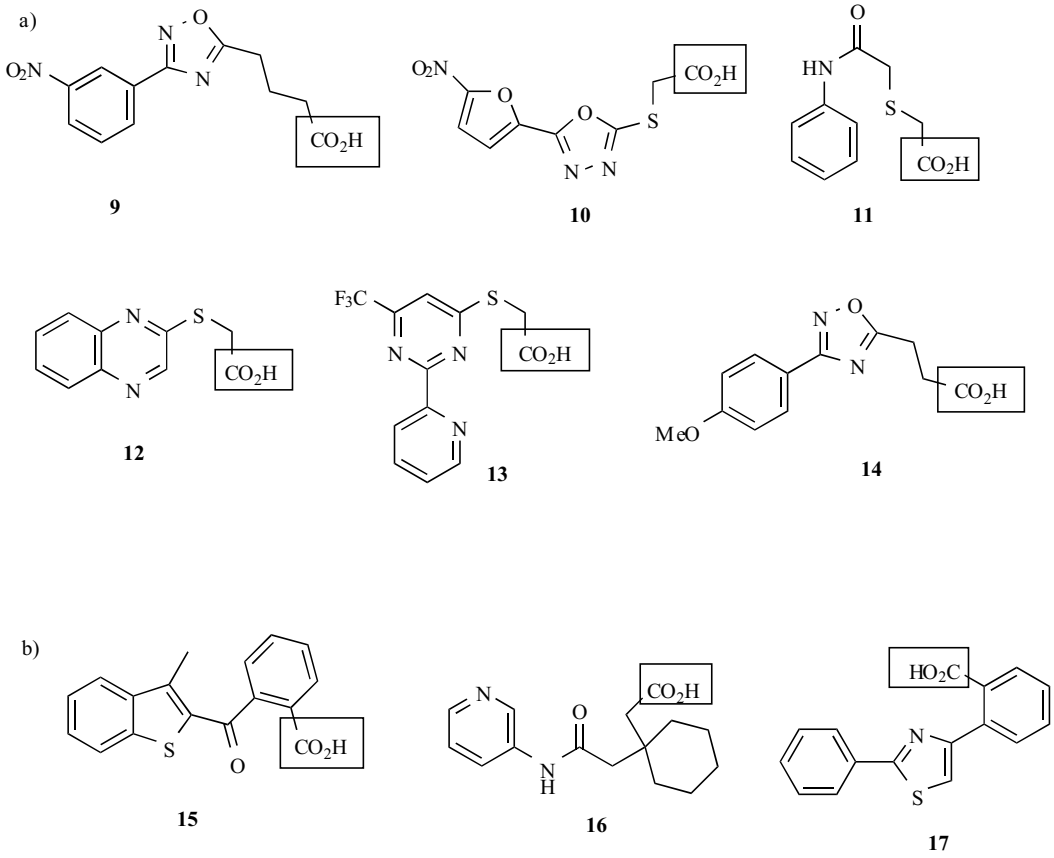


Fig. (10). Nine top-scoring compounds from VLS. a) actives b) inactives.

allows for the screening to be executed in a fast and efficient manner although it inherently limits the scope of the search. The ACD was used as the starting virtual library and was passed though a series of hierarchical filters, described below, to sequentially narrow the library to good lead candidates.

A 2D filter guaranteed the presence of a functional group to anchor the anion binding pocket. This anchor group could be an acid as is found in the bound inhibitor or a carboxylic acid surrogate such as a hydroxamic acid, tetrazole, phosphonic acid or sulfonic acid. A second 2D filter removed compounds in violation of Lipinski rules, with a MW greater than 350 Da or more than eight rotatable bonds. This effectively limited the library from 260,000 initial structures to ~12,500 compounds. The next round of filtering subjected a 3D version of the dataset to a pharmacophore match with the UNITY program. The pharmacophore model used a structure of the bound inhibitor to aldose reductase and a computational analysis of the binding pocket that located groups predicted to have favorable interactions with the putative ligands. Screening of the ~12,500 remaining entries against this pharmacophore model further reduced the screening set to 1261 compounds, 97% of which possessed a carboxylic acid. These remaining candidate molecules were docked into the crystal structure of human aldose reductase using FlexX and the quality of the docked structures was evaluated with DrugScore. Only the best scoring pose for each compound was considered. The top scoring 120 compounds were grouped into 40 clusters with SYBYL and the best candidates from each cluster were evaluated visually. Nine compounds that appeared to interact favorably with the binding pocket were selected and examined for biological activity. These nine compounds, Fig. (**10**), show a diverse range of functionality although each retains a carboxylic acid, a result of the initial bias in the first filter.

Six of nine of the compounds (**9-14**) were shown to be active (66% hit rate) in the micromolar range, the remaining three (**15-17**) were non-binders. The authors speculate that these false positives may be inactive because they are less flexible than the six active compounds.

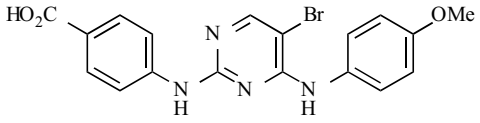
Chk-1

Recent work from AstraZeneca [100] nicely illustrates the use of knowledge-based library filtering during a discovery effort to locate new inhibitors of checkpoint kinase-1 (Chk-1 kinase), a key regulatory enzyme which prevents cells with DNA damage from progressing through the G2/M checkpoint in the cell cycle. Inhibition of this kinase has been shown to sensitize cancer cells to cytotoxins and have potential for use in combination with other antineoplastic agents.

The virtual library was generated from the AstraZeneca compound collection containing ~560,000 chemical entities. An initial round of prescreening removed compounds that were unlikely to be good drugs, in this case, compounds with a molecular weight greater than 600 Da or compounds containing more than ten rotatable bonds. A 3D conformational profile of each library member was generated as a prelude to pharmacophore matching. A two-point pharmacophore model, defined as a hydrogen bond donor and acceptor pair separated by a distance of 1.35-2.40 Å, was generated from knowledge of various kinase inhibitors bound to the adenine binding pocket. An in-house computational filter was employed to match the conformational library to this pharmacophore model, leaving ~200,000 compounds (approximately a 50% reduction in the size of the virtual library) to be examined though docking to the X-ray crystal structure of Chk-1 kinase.

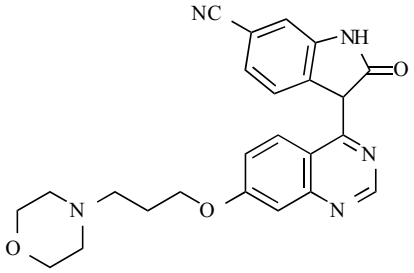
Utilizing the extensive structural database of complexes of kinases and inhibitors that bind to the highly-conserved ATP binding site, a knowledge-based approach to docking could be employed to profile compounds that were likely to have kinase binding motifs within their structures. This guided docking protocol was performed with FlexX-Pharm which docks compounds under a defined pharmacophoric constraint. Since it was known that most ATP-competitive inhibitors mimic the purine ring of ATP, an interaction with the backbone NH of Cys87 and the amide carbonyl of Glu85 was set as a requirement during the docking protocol along with other conserved interactions.

A maximum of 100 poses was saved for each docking event and the compounds rescored by consensus scoring methods (PMF and FlexX) which produced ~250 compounds that reached the cutoff set by the investigators. A final visual inspection of these candidates to remove compounds displaying unfavorable interactions reduced the number to 103 compounds to be evaluated through kinase assays. Thirty-six of the 103 compounds (~35% hit rate) showed activity against Chk-1 kinase (ATP competitive) with IC_{50} values ranging from 110 nM to 68 μM. These compounds represent four distinct chemical classes with a 0.35 Tanimoto similarity value. Two active compounds are shown below in Fig. (**11**).



18

IC_{50}=450 nM



19

IC_{50}=4 mM

Fig. (11). Structure of two VLS hits for Chk-1 kinase.

This example nicely shows how judicial use of information about related biological receptors can be used to build knowledge-based filters to eliminate compounds from

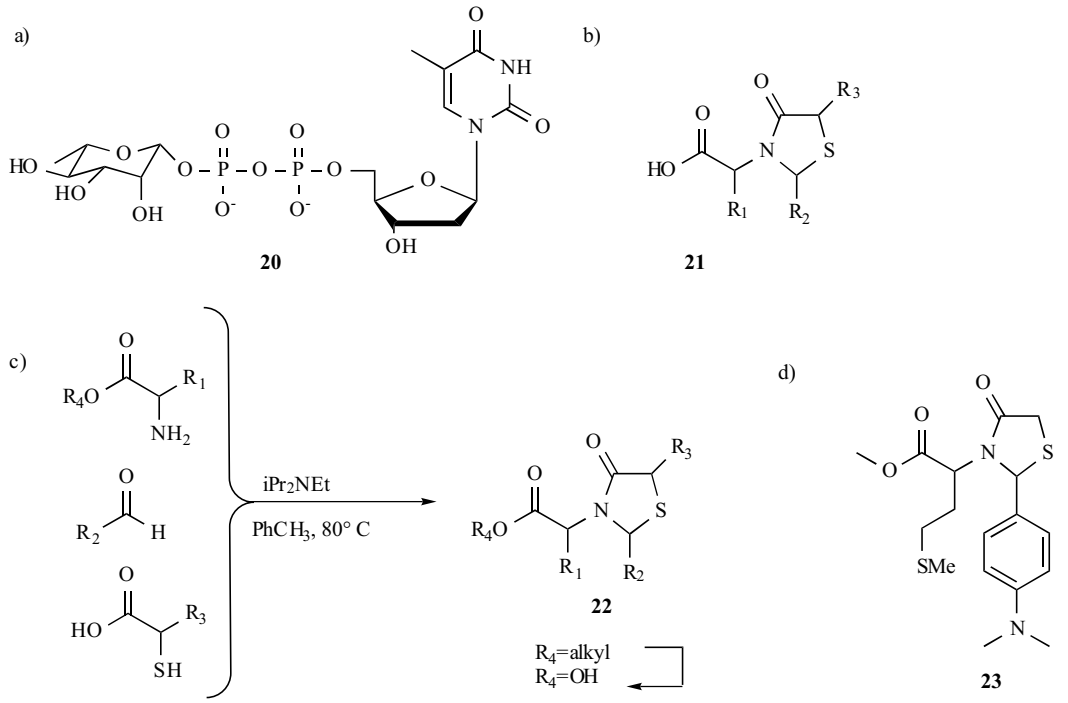


Fig. (12). (a) structure of TDP-rhamnose, (b) general library design with three points of variation, (c) synthetic sequence for library preparation and (d) structure of the most potent analog.

the virtual library prior to the more computationally demanding docking and scoring functions.

Rm1C

The application of a structure-guided library protocol to the discovery of inhibitors against a novel target in *Mycobacterium tuberculosis* has recently been reported [101] and features a virtual library of thiazolidinones as rhamnose mimetics, Fig (**12**). The recent emergence of tuberculosis strains that show multi-drug resistance against a range of clinically used agents has prompted several investigations to identify novel targets in the mycobacterium for further drug development. Babaoglu and co-workers targeted the machinery involved in the biosynthesis of rhamnose, a bacterial-specific component of the peptidoglycan layer of the cell wall. Specifically, they targeted deoxy-D-xylo-4-hexulose 3,5-epimerase (Rm1C), an enzyme that inverts the configuration of the 3' and 5' hydroxyl groups of the pyranose ring.

In the design of Rm1C inhibitors, a structure-guided library was constructed *in silico* based on an inhibitor of a related enzyme, MurB. Prior work had shown that 2,3,5-trisubstituted-4-thiazolidinones were effective inhibitors of MurB and function as diphosphate mimetics at the nucleotide sugar binding site. Based on this knowledge, a generalized thiazolidinone library (**21**) was formulated as a mimetic of TDP-rhamnose (**20**) which allowed for the introduction of variable groups at three distinct positions about the heterocycle.

The thiazolidinone is highly versatile and can be easily prepared through an efficient three-component coupling process which involves the cyclocondensation of various amino esters, aldehydes and α-thioacids, Fig. (**12c**). The initial products are formed as the esters but can be easily converted to the corresponding acids though hydrolysis.

A virtual library of 3,888 compounds based on this thiazolidinone scaffold was generated using CombiLibMaker. All possible combinations of the products resulting from the condensation of 24 aminoacids, 27 aldehydes and only two thioacids (R_3=H or CH_3), including diastereomers created at positions R_2 and R_3, were generated *in silico*. Once the new compounds were prepared, 3D coordinates were generated and the structures minimized. The macromolecular target was derived from the X-ray crystal structure of Rm1C bound to a substrate analog and the active site was defined as any residue containing an atom within 6.5 Å of the bound substrate. The library of 3,519 thiazolidinones was successfully docked into the model and consensus scoring was carried out using Cscore. The top scoring 144 compounds (5%) were identified and slated for chemical synthesis.

The 144 compound sublibrary was prepared in a parallel fashion using the condensation reaction depicted in Fig. (**12c**). This reaction sequence successfully generated the compounds with 47 of the 144 compounds synthesized both in the ester (R_4=alkyl) and the free acid form (R_4=OH). Utilizing a coupled assay system, 30 compounds were identified that inhibited the enzyme at a concentration of 20 μM. One derivative (**23**) completely inhibited the enzyme at this concentration. However, only seven of these 30 compounds show measurable MIC (minimum inhibitory concentration) values against *M. tuberculosis*. The authors suggest that the low MIC values may relate to poor cell permeation by these compounds.

This study illustrates how the diversity space covered by the initial library can be dramatically reduced using a focused library based on the structure of an inhibitor of a related enzyme. This required far fewer compounds in the initial library relative to an unbiased screening effort. If, as the authors suggest, that the low MIC values against the

Fig. (13). Crystal structure of Bcr-Abl bound to STI571.

organism are an artifact of poor penetration it may prove valuable to conduct a preliminary screen on the virtual library to eliminate compound candidates with poor pharmacological profiles.

Bcr-Abl Tyrosine Kinase

Inhibitors of Bcr-Abl are essential for the treatment of chronic myelogenous leukemia. Gleevec (Novartis Pharmaceuticals), also called STI571, has proven effective in this regard, but resistance to STI571 has arisen recently. The amino acid substitution of Thr 315 with Ile [102] and the loss of a hydrogen bond to Thr 315 is assumed to cause STI571 resistance. Peng, *et al.* [103] conducted a virtual screening effort directed against the catalytic domain of the Abl tyrosine kinase to discover inhibitors that may overcome the resistance phenomenon. The crystal structure of Abl tyrosine kinase bound to a variant of STI571 was used as the target structure. A database containing 200,000 commercially available compounds was converted into 3D coordinates. The STI571 binding site without the bound ligand or bound water molecules was chosen for docking studies, Fig. (**13**).

DOCK4.0.1 was used to dock and score the 200,000 compounds in the target site. The top 1,000 compounds were selected for further analysis. These compounds were clustered into structurally diverse sets and member compounds were chosen from individual groups. Fifteen compounds that obey Lipinski's rules were selected for biological assay. Two of these fifteen inhibited the growth of the Philadelphia-positive K562 cell line in a dose-dependent fashion with IC_{50} values of 24 and 29 μM. One of these novel compounds does not appear to require a hydrogen bond with Thr 315, suggesting that it may inhibit STI571-resistant human leukemia cell lines.

Ligand Binding Domains

Thyroid Receptor

Utilizing a combination of homology modeling, virtual screening, chemical synthesis and focused library design, Schapira and co-workers were recently able to generate novel antagonists of the thyroid hormone receptor (TR), a member of the nuclear hormone receptor superfamily [104]. Antagonism of this receptor is viewed as a new therapeutic strategy to treat hyperthyroidism which is currently treated with radiation, invasive surgery or abrogation of thyroid hormone biosynthesis.

These investigators utilized the crystal structure of other members of the nuclear hormone receptor superfamily, both free and bound to various agonists and antagonists, to generate a homology model of the ligand-binding domain (LBD) of TR. From these structures, it was observed that the binding modes of agonists and antagonists differed in the orientation of a C-terminal helix (H12) that folds onto the agonist thereby creating a hydrophobic binding pocket. When an antagonist is bound, the H12 helix is prevented from folding into the binding cavity. The crystal structure of the estrogen receptor-α complexed to raloxifene (an ER antagonist) and the structure of the agonist-bound TR LBD were taken as departure points for the generation of a model of the TR LBD bound to an antagonist. The TR-antagonist model has the H12 helix folded out of the binding pocket and retains the structure of the N-terminal domain since this domain does not differ significantly between the agonist and antagonist bound state.

The ACD collection, filtered to remove those compounds with problematic pharmacological properties in accordance with Lipinski rules, was docked into the TR model using Molsoft's ICM virtual library screening module. The

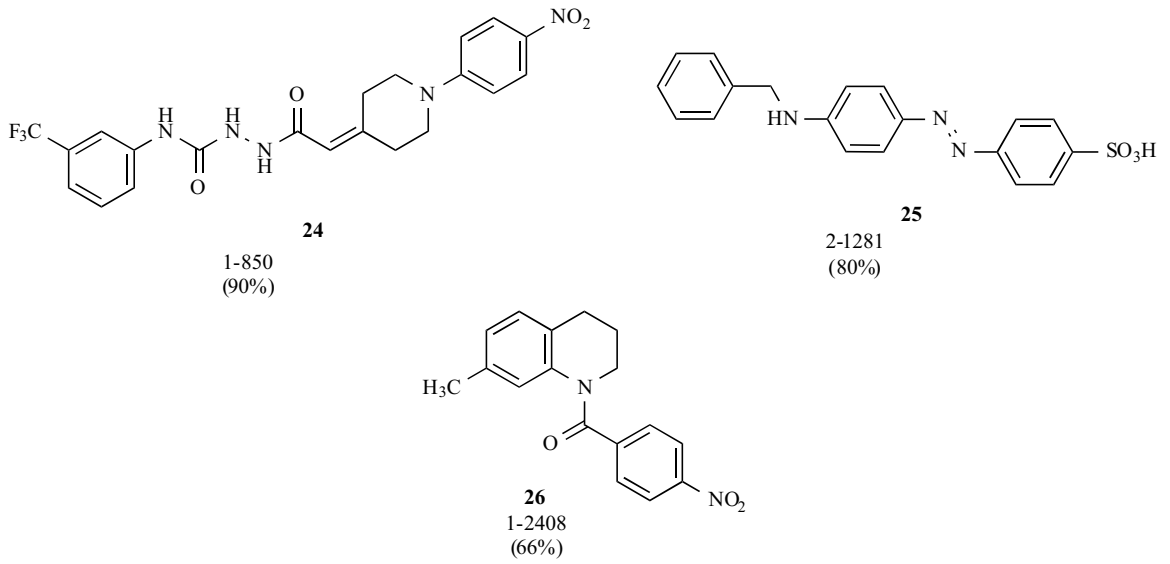


Fig. (14). Structure of the top three TR antagonists. Percent inhibition is at 20 μm inhibitor concentration.

Lipinski filter effectively reduced the size of the ACD library from 250,000 to 190,000 compounds. Each of the protein-ligand complexes was scored based on a grid energy that included terms for electrostatics and entropy. A second filter was designed based on the knowledge that nuclear hormone receptor ligands possessing a sterically demanding group projecting out of the binding cavity act purely as antagonists by preventing the H12 loop from folding onto the binding cavity. This filter eliminated those compounds that, although scoring within the desired threshold, did not have such a group. The top 1,000 candidates that passed this filter were refined in a model that permitted both ligand and side-chain flexibility. Visual inspection of the top scoring 300 compounds for shape complementarity, flexibility and hydrogen bonding reduced the target library to 100 compounds to be examined *in vitro* for TR-receptor antagonism.

Only 75 of the 100 top-scoring compounds were currently available from commercial sources. Of the 75 compounds screened, 14 showed antagonist activity against the TR receptor, Fig. (**14**).

The best three antagonists are shown above (**24-26**) and are illustrative of the wide structural diversity that is possible in small molecules that bind to the same biological receptor. The acyl hydrazide (**24**) showed 90% inhibition of TR activity at 20 μM concentration and was used as the starting point for the preparation of a virtual, focused library of potential inhibitors.

The design of the new library was based on a combination of structural and synthetic considerations. Examination of the docked structure of (**24**) indicated a key hydrogen bond between the hydrazide oxygen and His-435 and an important interaction between the nitro group and an arginine sidechain and as such, these two moieties were conserved in the sublibrary design. Synthetically, a coupling reaction between a hydrazide and an isocyanate appeared as a facile process for the parallel synthesis of this library, Fig. (**15**).

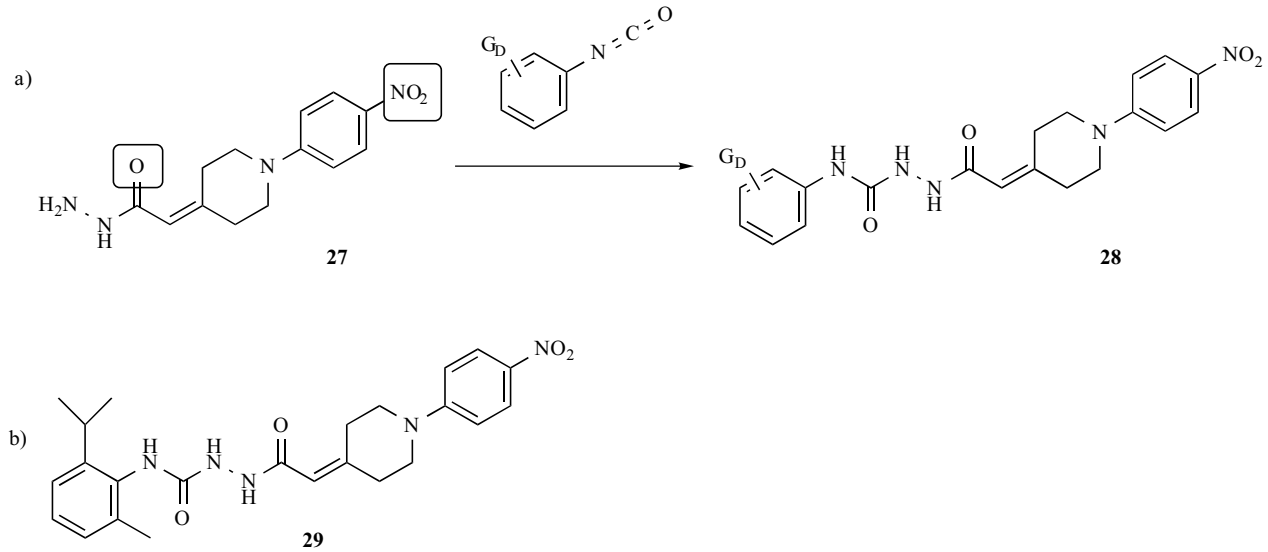
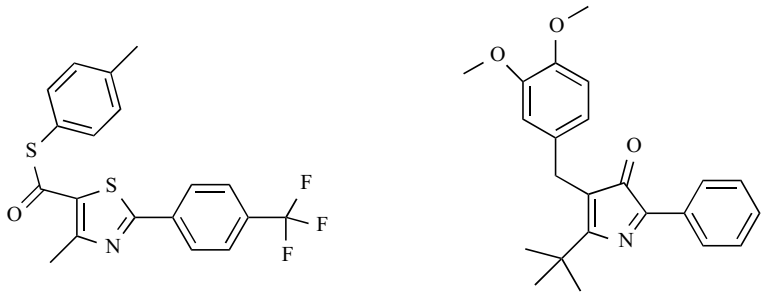


Fig. (15). a) Diversification step to generate focused libraries of 1-850 analogs. Conserved moieties are shown in boxes and G_D refers to diverse groups b) Structure of the most active analog.



30 **31**

Fig. (16). Two novel antagonists for RAR discovered through virtual screening.

Based on commercially available isocyanate building blocks, a virtual library of 101 derivatives of 1-850 were prepared, docked and scored as previously described. Eight of the top 57 compounds were prepared and tested. The new analogs showed inhibition between 10% and 84% at a concentration of 5 μM with the most active compound (**29**) showing a sub-micromolar IC_{50} value (750 nM) in a dose-response assay.

Using a combination of homology modeling, docking and chemical synthesis, a novel antagonist of the TR receptor with a nanomolar IC_{50} value was identified. A nice feature of this work is the use of focused library synthesis to greatly increase the activity of an initial hit from a large, diverse screening library.

Retinoic Acid Receptor

(RAR) agonists and antagonists block the growth of several neoplastic cells including breast tumor cells. Schapira *et al.* [105] developed a homology model of RAR-γ bound to an antagonist, based on the crystal structure of RAR-γ bound to an agonist and the structure of estrogen receptor bound to an antagonist. From there, a model of RAR-α bound to an antagonist was derived. Docking, using flexible ligands, flexible receptor side chains and a full atom representation of the receptor, was carried out using ICM. The ACD was docked into the site and scored based on electrostatic, hydrophobicity and entropy parameters. The top 500 compounds were visually evaluated and 32 of those were selected for biological assay. The visual selection criteria for the 32 included those with the best van der Waals fit or hydrogen bonding. Of the 32 compounds selected for biological testing, two candidates selectively inhibited activity by 55% and 33% at 20 μM, Fig. (**16**).

The models of these candidates with the receptor suggest that they present an additional arm that protrudes from the pocket and prevents the active conformation. Possibilities for increasing the potency of these compounds were evident from the models of the compounds in the receptor structure.

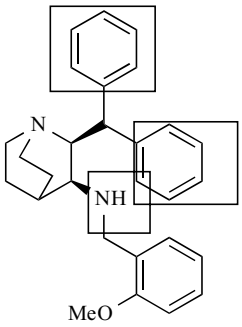
During this screen, all pharmacophore restraints were avoided in order to discover novel ligands. The database was filtered only by a good fit to the receptor and reasonable bioavailability. The authors acknowledged a compromise between the time allocated for each ligand and the reliability of the sampling of conformational space.

NK-1

As previously discussed, G-protein coupled receptors (GPCRs) are excellent drug targets but have not been commonly utilized in VLS because of the difficulties associated with obtaining high-resolution structural information. A recent high resolution X-ray structure of bovine rhodopsin encouraged Bissantz and co-workers [106] to examine the reliability of homology models constructed from this structure to function in VLS efforts to discover new modulators of GPCR activity. Six homology models of human GPCRs were constructed with SYBYL. Models of the dopamine D3 receptor, acetylcholine muscarinic M1 receptor and the vasopressin V1a receptor were used for antagonist screening while models of the dopamine D3, β2-adrenergic and δ-opiod receptors were employed for agonist screening. The homology structures were refined in the bound conformation with known agonists or antagonists. Finally, 990 compounds were randomly chosen from the ACD that were similar in molecular weight to known ligands. Ten known antagonists or agonists for each receptor were added to the virtual library to give six libraries of 1000 compounds each. Flexible docking of the test libraries to the modeled structures using DOCK, FlexX and GOLD with seven different scoring functions was studied. These test cases revealed that the hit rates for locating antagonists for the D3 and V1a receptors were 20 to 40 fold higher than random screening. Agonists, on the other hand, were much more difficult to locate, perhaps reflecting that the models were derived from the inactive state of bovine rhodopsin, which is closer to the antagonist rather than agonist bound state.

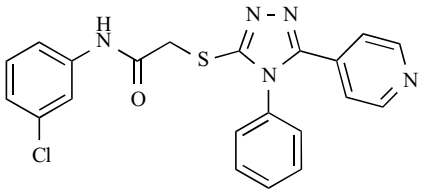
In a subsequent study by Evers *et al.* [107], a successful model for virtual screening was developed for another important GPCR, the NK-1 receptor. This study was conducted using their MOBILE approach which relies on the refinement of homology models with a known ligand docked as a further constraint. Again, using the structure of rhodopsin as a starting point (21% homology with NK1), an ensemble of 100 preliminary homology models was prepared and the well-studied NK1 antagonist CP-96345, Fig. (**17**) was docked (AutoDock) into the binding site which had been located by mutational studies. Those complexes that reproduced key interactions of the inhibitor were used to generate new homology models. Other known NK1 antagonists with diverse structures were examined in this model to help guide the creation of a pharmacophore hypothesis. Using a 2D filter that required the presence of two phenyl rings and one H-bond acceptor, lead-like compounds (less than 8 rotatable bonds and MW<450Da) were selected from 800,000 compounds accumulated from seven databases. A subsequent 3D filter in the UNITY program was used to constrain the spatial arrangement

a)



32

b)



33

K_i=251 nM

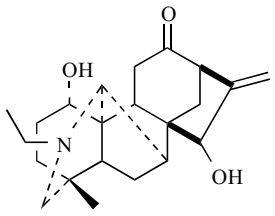
Fig. (17). a) Structure of CP-96345. Moieties used for pharmacophore constraint are shown in boxes. b) Structure of the most potent compound found by VLS.

between the aromatics and H-bond acceptor to produce 11,109 screening candidates. The FlexX-Pharm program was used to flexibly dock these compounds and the resulting complexes were scored with DrugScore. Visual inspection of the best hits to remove compounds with unsatisfied H-bond donors/acceptors or voids at the binding interface left seven compounds for biological evaluation. One of these compounds (**33**) produced a submicromolar K_i value (251 nM).

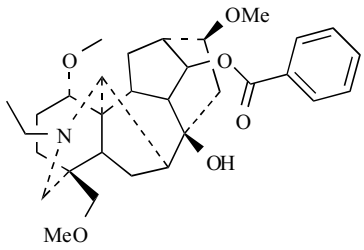
K^+ Channel

The majority of screening libraries used in VLS-based discovery efforts are composed either of commercially available compounds or in-house, proprietary compound collections. However, natural products have long served as a rich and diverse source of compounds for new pharmaceutical leads. A recent disclosure [108] from workers in China demonstrated that virtual libraries derived from the structure of natural products can also be valuable databases for lead identification.

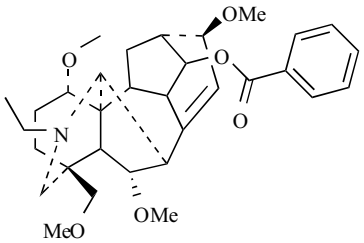
In this study, the screening efforts were directed towards identifying compounds that could inhibit the function of the potassium channel. These are pore forming, voltage-gated ion channels that are responsible for controlling a variety of cellular processes in both excitatory and non-excitatory cells. In 1998, the first report of a crystal structure of a potassium channel, the KcsA channel from *Streptomyces lividans* was described at 3.2 Å resolution [109]. This structure provided the basis for the generation of a homology model of the eukaryotic *Shaker* potassium channel and the homology structure was refined using SYBYL. The China Natural Product Database, containing structural information on approximately 50,000 natural products, was used as the screening library against the homology model of the potassium channel. Docking was performed against the pore forming site using DOCK in an attempt to locate inhibitors that would act by blocking the entry of potassium ion into the pore formed from the tetrameric aggregate. Complexes of the 200 top-scoring compounds were minimized and the compounds inspected visually to locate good potential ligands. Fourteen natural products were selected for biological assay, however, only four were ultimately available for examination, Fig. (**18**).



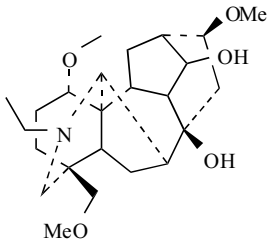
34



35



36



37

Fig. (18). Structures of natural product inhibitors of the potassium channel. Note the common structural features in the molecules.

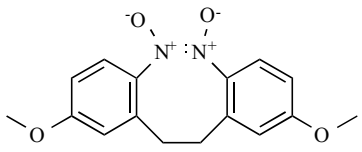
An examination of the four available compounds (**34-37**) using whole-cell voltage-clamp recording in rat brain showed that all four compounds exerted effects on potassium channels with each of these showing different preferences and potencies. Three of the four were determined to be selective blockers of I_K (potassium current) at the 1 mM level. Further electrophysiological experiments indicated that the compounds do bind to the extracellular site and block the movement of potassium ion current. This is one of the few examples that used natural product libraries in VLS. The level of success is quite surprising in light of the fact that a homology model based on a structure of relatively low resolution was used for docking.

Inhibitors of Protein:Protein Interactions

Bcl-2

Bcl-2 is a member of a family of proteins that regulates programmed cell death (apoptosis). Bcl-2 is an attractive anticancer target since overexpression of the protein has been seen in many types of cancer, including breast, prostate and lymphoma. The formation of heterodimers of Bcl-2 or Bcl-X_L and a number of other proteins, including Bak, Bad and Bax is believed to play a role in the prevention of apoptosis. Therefore, inhibitors targeting the interaction of Bcl-2 and Bcl-X_L, specifically the Bak BH3 peptide binding site, were sought using VLS [61]. The authors first used homology modeling to generate ten models of the structure of Bcl-2 based on the experimental NMR structure of Bcl-X_L in complex with the Bak BH3 peptide. The average model was then refined with extensive molecular dynamics simulations.

Fig. (19). NMR structure of Bcl-X_L bound to the Bak peptide.



38

Fig. (20). An inhibitor of Bcl-2 protein:protein interactions.

The target binding site was identified as all residues within 8 Å of the Bak BH3 peptide binding site since this interaction provided a reasonable model for the interaction of Bcl-2 with other proteins. The National Cancer Institute database of compounds with structures and biological data that are accessible to the public, comprising 206, 876 "open" compounds, was screened against the modeled structure of Bcl-2 using the program, DOCK. The top 500 compounds with the best scores, based on a simple enthalpic energy scoring function in DOCK, were considered and further filtered based on those with nonpeptide character. Thirty-five compounds were obtained from the NCI and tested in biological assays. A fluorescence-based polarization binding assay established an IC_{50} value of 0.3 μM for the interaction of the Bak BH3 peptide and Bcl-2. Seven of the 35 compounds from the NCI showed dose-dependent, competitive inhibition at levels greater than 100 μM.

Compound (**38**) potently inhibits cell viability and growth and represents a novel class of Bcl-2 inhibitors. Compound (**38**) was synthesized and shown to induce apoptosis in a dose-dependent manner in cell lines overexpressing Bcl-2. Compound (**38**) was also shown to bind Bcl-X_L specifically with a binding constant of 7 μM. NMR spectra of (**38**) in complex with Bcl-X_L, chosen because of its greater solubility over Bcl-2 and therefore its amenability to structure determination, were measured and peak shifts were evident for those residues in the region of the BH3 binding pocket of Bcl-X_L.

Rac 1

Structure guided identification of a novel binding site formed between two proteins was recently used by Gao [110] and co-workers to identify an inhibitor of a Rho GTPase, a key signal transduction mediator implicated in a variety of disease states including cancer. Activation of the Rho GTPase is controlled by the Dbl family of guanine nucleotide exchange factors (GEFs). Structures of several GEFs bound to Rac-1 revealed a key groove on Rac-1 that was responsible for the specificity of binding by GEFs such as Trio and Tiam1. Moreover, a specific residue, Trp 56, was identified as crucial for this interaction as a Rac1 mutant, W56F, lost the ability to interact with Trio and Tiam1. Model peptides containing the Trp 56 residue were also shown to inhibit the ability of GEF to activate Rac-1. Based on these findings, a study using VLS was undertaken to identify potential inhibitors of Rac-1 activation by locating compounds that would bind into this groove and make specific contacts with Trp 56.

With the structure of the Rac-1/Tiam complex as a departure point, a putative binding pocket was constructed from those residues surrounding Trp 56 (within 6.5 Å) which included Lys 5, Val 7 and Ser 71, Fig. (**21**).

The NCI database, which contains coordinate files for ~140,000 small organic molecules was used as virtual library. The UNITY program was used to screen the virtual compound collection under conditions which permitted flexibility in the ligands and the compounds were docked into the putative binding site using FlexX. Consensus scoring was used to determine and rank the top 100 candidate compounds. The investigators used a visual inspection process to remove compounds that did not appear

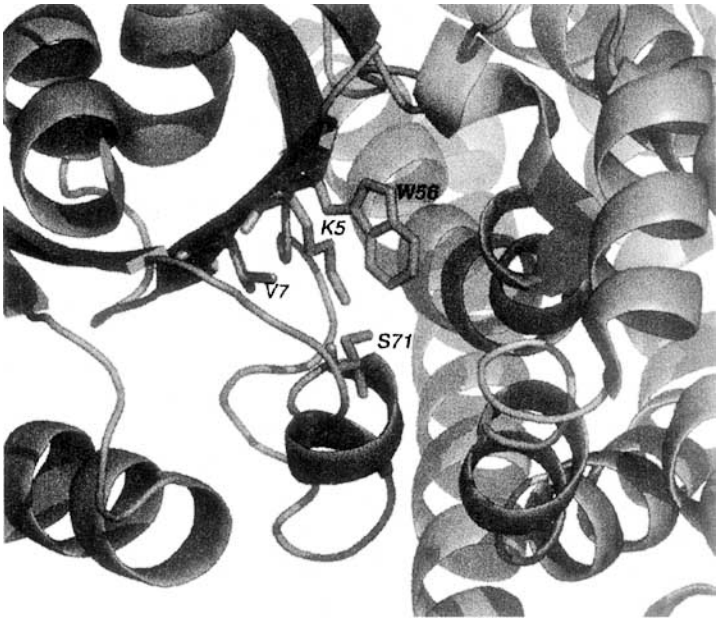
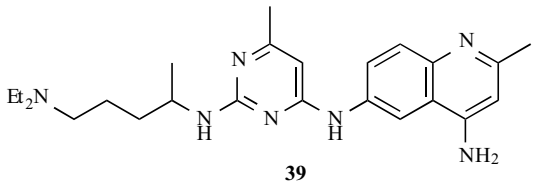


Fig. (21). Structure of Rac-1 (light gray) in complex with Tiam (dark gray).

to have an interaction with the crucial Trp 56 residue, a process that reduced the top ranking compounds to 58. Because of poor solubility or unavailability of some compounds, only 15 of these compounds could be examined in biological assay. Examination of the 15 candidates revealed a single compound that was able to inhibit the Rac-1/GEF I interaction *in vitro*, Fig. (**22**), NSC23766 (**39**) inhibited Rac1-TrioN interaction in a dose-dependent manner with an approximate IC_{50} value of 50 μM. The model of the interaction between NSC23766 with Rac-1 indicated that the compound binds in the desired cleft primarily through hydrophobic interactions with a stacking interaction between the central pyrimidine ring of the inhibitor and the indole sidechain of Trp 56. No attempts to improve the potency of (**39**) by further chemical modification were reported.



39

NCS23766

Fig. (22). Single VLS hit for Rac-1/Tiam inhibitors.

p56 Lck

Huang *et al.* [111] utilized the high–resolution structure of the SH2 domain of p56 Lck bound to a phosphopeptide in combination with VLS to identify novel, non-peptidic inhibitors of this Src family kinase. p56 Lck is predominantly involved in the regulation of the immune system and may be a valuable target for lymphomas and rheumatoid arthritis. The initial inhibitor (pY-E-E-I) was located during a screen of a phosphopeptide library against the SH2 domain of Lck, a region that interacts with phosphotyrosine residues in protein:protein interactions. However the phosphate group is unstable *in vitro* and the discovery of more druggable ligands against this site was desired.

The X-ray structure showed that the phosphotyrosine and isoleucine residues of the tetrapeptide each bound into two separate cavities, referred to as the pY and pY+3 cavities. These investigators chose to target the cavity that comprises the pY+3 site for non-peptide inhibitors of p56 Lck. Phenolphosphate (a model of phosphotyrosine) was maintained in the pY site to prevent the docking of ligands to this cavity. A 3D-database of 2,000,000 commercially available compounds was prepared from a 2D building program and the structures optimized with SYBYL6.4. The structure of the Lck SH2 domain bound to the phosphopeptide inhibitor was used for the screen. Docking was carried out with DOCK and the energies of the ligand complexes were evaluated. A series of further filters developed by these investigators were utilized to compensate for biasing in the DOCK protocol towards high molecular weight compounds. 25,000 compounds (mean MW=345 Da as compared to 475 Da) were selected after application of a van der Waals attractive energy normalizing function and taken through a second, more rigorous round of docking that included energy minimization. From this second screen, two sets of 1,000 compounds were selected for further analysis.

In the following round of selection, specific attention was paid to the diversity of the compounds. Tanimoto similarity indices generated with MOE were utilized to generate clusters of dissimilar compounds. Each cluster of compounds was analyzed for good candidates for biological evaluation by applying Lipinski-type filters to the compounds and removing those compounds that would be likely to have poor ADME properties. Approximately 100 different clusters were generated and two candidates from each cluster were chosen for biological assay. Final selection revealed 288 unique compounds between the two sets, of which 196 were still available from commercial sources.

Biological assay of the 196 compound sublibrary at a concentration of 100 μM against p56 Lck produced a hit rate of 17% which corresponded to 34 active lead compounds. Flourescence titration experiments subsequently showed that four of these compounds, Fig. (**23**), directly bound to the

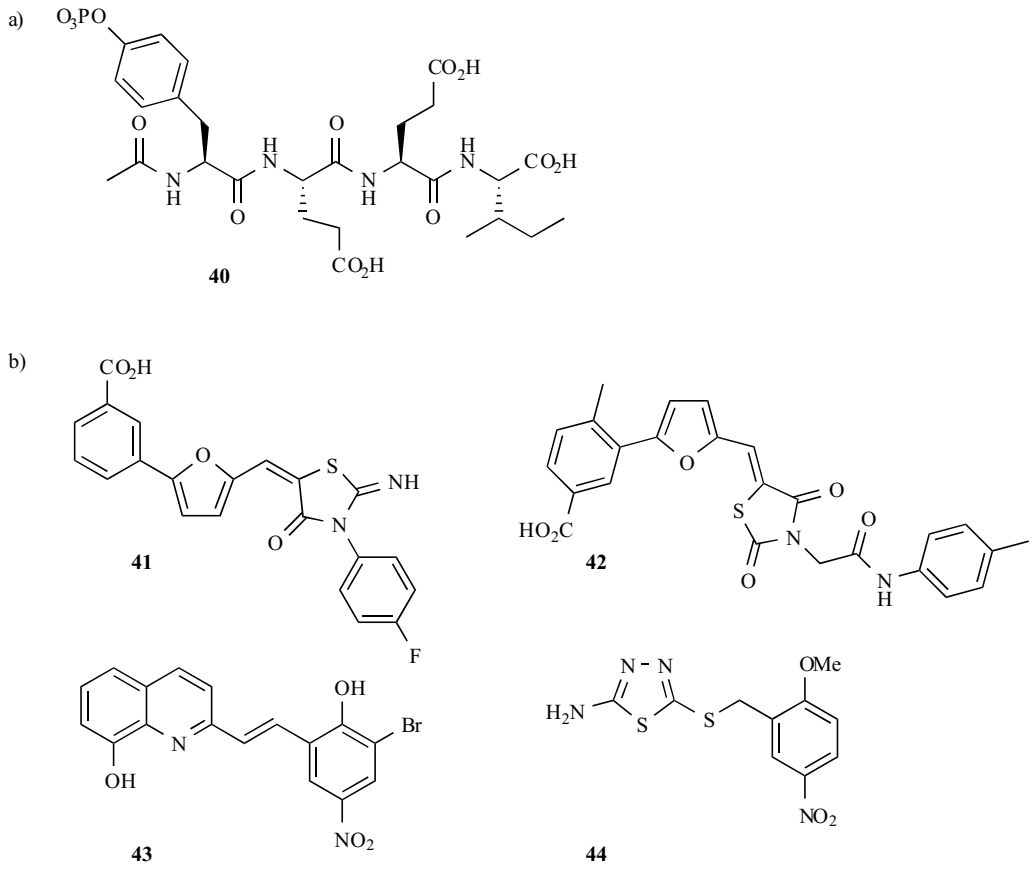


Fig. (23). a) Phosphopeptide lead for VLS b) Binders of the p56 Lck SH2 domain.

SH2 domain of Lck with K_D values in the low micromolar range.

This study nicely illustrates a family-based approach to maximizing diversity. Rather than looking only at the top-scoring compounds, clustering of a wide range of potential actives allowed a sampling of a much greater chemical diversity (**41-44**). This approach could be highly useful in a medicinal chemistry program as it is advantageous to pursue multiple lead scaffolds in parallel to provide adequate backup compounds against late-stage failure of clinical candidates.

SUMMARY

These case studies, described above, provide compelling evidence that VLS can identify lead compounds in a cost efficient and rapid manner. Recent work suggests that lead compounds can be found for a wide array of biological targets such as enzyme catalytic sites, ligand-binding domains and even protein:protein interactions.

It is evident that a range of challenges remain in the field and that progress towards removing these challenges will serve to increase the rate at which leads are identified, the initial potency of the leads and the quality of those leads as true drug candidates. Three challenges that emerge are: (1) the need for more purposefully designed virtual libraries, (2) the need to consider ligand and target flexibility in docking, and (3) the need to develop accurate scoring methods compatible with a reasonable computational time. The need for more purposefully designed libraries is highlighted in the fact that the majority of recent studies have relied on virtual libraries derived from existing compound collections such as ACD, NCI or historical collections. These libraries are attractive because the identified leads are seemingly available through purchasing rather than synthesis. However, many of these cases reveal that it can be difficult to actually obtain the identified hits. Clearly, the purposeful design of new virtual libraries that take into account diversity and lead-likeness, as well as synthetic accessibility and amenability to further elaboration will have a major impact on the use of VLS in drug discovery. The second challenge, considering the flexibility of the ligand and the target without generating a combinatorial explosion, will be necessary to produce a more accurate representation of the bound structure for subsequent scoring. The case study describing the interaction between didansyltyrosine and thymidylate synthase provides a striking example of the need to consider flexibility in docking. Finally, although several accurate scoring algorithms have been developed to fully evaluate the interaction between a ligand and its receptor, these algorithms are too computationally demanding for large virtual libraries containing hundreds of thousands of compounds. Further development to increase the accuracy and speed of minimal scoring algorithms will be needed to better correlate the computational and biological ranking.

Despite these important challenges, there have been many successful VLS studies. One of the important features revealed in these VLS studies is that a diversity of hits can be found for a single target. From a medicinal chemistry perspective, the ability to develop multiple families of leads

in parallel provides a margin of safety against late stage failure in the drug development process. The p56 Lck case study nicely illustrates how the leads were filtered to bias for a diverse collection. It is also evident from an analysis of the case studies that micromolar inhibitors can be identified in the virtual library and that it is rare to identify submicromolar hits. However, lead optimization either through traditional medicinal chemistry (for example the CDK inhibitors) or using focused libraries of an identified scaffold (for example cathepsin D inhibitors) can translate these micromolar hits to low nanomolar leads. Teague [90] recommends a strategy of selecting low molecular weight hits (MW≤350 Da) such that high affinity leads derived by elaboration of the hit maintain good drug-like properties. In terms of the macromolecule in VLS, a number of case studies including NK-1, TR and RAR, have shown that targets derived from homology models, the structural method that relies on the least amount of experimental data, functions well in VLS studies.

One of the most important conclusions from these studies is that the use of focused libraries that incorporate information about an initial scaffold yield improved hit rates relative to random screening. The study describing cathepsin D highlights that if focused libraries are used, significantly smaller libraries will be needed for hit identification and that the potency of the hits will be 3-4 times greater than those from a diverse library. The information about an initial scaffold can be as simple as a single functional (anchor) group. This strategy was used to discover potent inhibitors of aldose reductase and carbonic anhydrase. In another application of focused libraries, knowledge of inhibitors of related kinases was nicely used to filter virtual libraries of potential Chk-1 inhibitors. Focused libraries have also proven highly effective in lead optimization. This is nicely illustrated in the development of nanomolar inhibitors of CDK and TR. In summary the application of structure-based library design with a team encompassing expertise in synthesis, structure and computation will evolve VLS into a premier strategy in drug discovery.

ABBREVIATIONS

VLS = Virtual library screening

HTS = High-throughput screening

TOS = Target-oriented synthesis

DOS = Diversity-oriented synthesis

ACD = Available chemicals directory

NCI = National Cancer Institute

ADME = Absorption, distribution, metabolism and excretion

NMR = Nuclear magnetic resonance

RCSB = Research collaboratory for structural bioinformatics

GPCR = G-protein coupled receptor

SAR = Structure activity relationships

DHFR = Dihydrofolate reductase

TS = Thymidylate synthase

dUMP = Deoxyuridine monophosphate

PA = Protective antigen

EF = Edema factor

LF = Lethal factor

CDK = Cyclin dependent kinase

hCAII = Human carbonic anhydrase

dTMP = Deoxythymidine monophosphate

DDT = Didansyl tyrosine

MIC = Minimum inhibitory concentration

TR = Thyroid receptor

LBD = Ligand-binding domain

RAR = Retinoic acid receptor

GEF = Guaninenucleotide exchange factor

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the National Institutes of Health and the National Science Foundation.

REFERENCES

[1] Hoever, M.; Zbinden, P. *Drug Disc. Today* **2004**, *9*, 358-365.
[2] Goldsmith, Z. G.; Dhanasekaran, N. *Int. J. Mol. Med.* **2004**, *13*, 483-495.
[3] Deprez-Poulain, R.; Deprez, B. *Curr. Top. Med. Chem.* **2004**, *4*, 569-580.
[4] Lavastre, O.; Bonnette, F.; Gallard, L. *Curr. Op. Chem. Biol.* **2004**, *8*, 311-318.
[5] Lee, A.; Breitenbucher, J. G. *Curr. Op. Drug Disc. Devel.* **2003**, *6*, 494-508.
[6] Sanchez-Martin, R. M.; Mittoo, S.; Bradley, M. *Curr. Top. Med. Chem.* **2004**, *4*, 653-669.
[7] Lyne, P. *Drug Disc. Today* **2002**, *7*, 1047-1055.
[8] Oprea, T.; Matter, H. *Curr. Op. Chem. Biol.* **2004**, *8*, 349-358.
[9] Walters, W.; Stahl, M.; Murcko, M. *Drug Disc. Today* **1998**, *3*, 160-178.
[10] Shoichet, B.; McGovern, S.; Wei, B.; Irwin, J. *Curr. Op. Chem. Biol.* **2002**, *6*, 439-446.
[11] Gane, P.; Dean, P. *Curr. Op. Struc. Biol.* **2000**, *10*, 401-404.
[12] Erhardt, P. W. *Pure Appl. Chem.* **2002**, *74*, 703-785.
[13] Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001**, *3*, 157-166.
[14] Greish, K.; Fang, J.; Inutsuka, T.; Nagamitsu, A.; Maeda, H. *Clini. Pharmacokinetics* **2003**, *42*, 1089-1105.
[15] Nicolaou, K. C.; Vourloumis, D.; Winssinger, N.; Baran, P. S. *Angew. Chem. Int. Ed.* **2000**, *39*, 44-122.
[16] Schreiber, S. L. *Science* **2000**, *287*, 1964-1969.
[17] Burke, M. D.; Schreiber, S. L. *Angew. Chem. Int. Ed.* **2004**, *43*, 46-58.
[18] Gooding, O. W. *Curr. Op. Chem. Biol.* **2004**, *8*, 297-304.
[19] Wender, P. A.; Handy, S. T.; Wright, D. L. *Chem. Ind.* **1997**, 765.
[20] Weber, L. *Curr. Med. Chem.* **2002**, *9*, 2085-2093.
[21] Kolb, H. C.; Sharpless, K. B. *Drug Disc. Today* **2003**, *8*, 1128-1137.
[22] Jamois, E. A.; Lin, C. T.; Waldman, M. *J. Mol. Graph. Model.* **2003**, *22*, 141-149.
[23] Jamois, E. A. *Curr. Op. Chem. Biol.* **2003**, *7*, 326-330.
[24] Kerns, E. H.; Di, L. *Drug Disc. Today* **2003**, *8*, 316-323.
[25] Di, L.; Kerns, E. H. *Curr. Op. Chem. Biol.* **2003**, *7*, 402-408.
[26] Egan, W. J.; Walters, W. P.; Murcko, M. A. *Curr. Op. Drug Disc. Dev.* **2002**, *5*, 540-549.
[27] Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. *Nat. Rev. Drug Disc.* **2002**, *1*, 337-346.

[28] Mason, J. S.; Hermsmeier, N. A. *Curr. Op. Chem. Biol.* **1999**, *3*, 342-349.
[29] Sotriffer, C.; Klebe, G. *Farmaco* **2002**, *57*, 243-251.
[30] Gohlke, H.; Klebe, G. *Angew. Chem. Int. Ed.* **2002**, *41*, 2645-2676.
[31] Matter, H. *J. Med. Chem.* **1997**, *40*, 1219-1229.
[32] Xu, L.; Yang, J. A. *Comp. Chem.* **1998**, *22*, 393-398.
[33] Willett, P. *Biochem. Soc. Trans.* **2003**, *31*, 603-606.
[34] Beno, B. R.; Mason, J. S. *Drug Disc. Today* **2001**, *6*, 251-258.
[35] Langer, T.; Krovat, E. M. *Curr. Op. Drug Disc. Devel.* **2003**, *6*, 370-376.
[36] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Del. Rev.* **2001**, *46*, 3-26.
[37] Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. Z. *Mol. Divers.* **2000**, *5*, 255-275.
[38] Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C.; Motoshima, H.; Fox, B.; LeTrong, I.; Teller, D.; Okada, T.; Stenkamp, R.; Yamamoto, M.; Miyano, M. *Science* **2000**, *289*, 739-745.
[39] Knegtel, R.; Kuntz, I.; Oshiro, C. *J. Mol. Biol.* **1997**, *266*, 424-440.
[40] Doreleijers, J.; Rullman, J.; Kaptein, R. *J. Mol. Biol.* **1998**, *281*, 149-164.
[41] Shuker, S.; Hajduk, P.; Meadows, R.; Fesik, S. *Science* **1996**, *274*, 1531-1534.
[42] Anderson, A.; O'Neil, R.; Surti, T.; Stroud, R. *Chem. Biol.* **2001**, *8*, 445-457.
[43] Davis, A.; Teague, S. *Angew. Chem. Int. Ed.* **1999**, *38*, 736-749.
[44] Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P.; Gilmore, T.; Graham, A.; Grob, P.; Hickey, E.; Moss, N.; Pav, S.; Regan, J. *Nat. Struc. Biol.* **2002**, *9*, 268-272.
[45] Dorsey, B.; Levin, R.; McDaniel, S.; K., V.; Guare, J.; Darke, P.; Zugay, J.; Emini, E.; Schleif, W.; Quintero, L.; Lin, J.; Chen, I.-W.; Holloway, M.; Fitzgerald, P.; Axel, M.; Ostovi, D.; Anderson, P.; Huff, J. *J. Med. Chem.* **1994**, *37*, 3443-3451.
[46] Erickson, J.; Neidhart, D.; VanDrie, J.; Kempf, D.; Wang, X.; Norbeck, D.; Plattner, J.; Rittenhouse, J.; Turon, M.; Wideburg, N.; Kohlbrenner, W.; Simmer, R.; Helfrich, R.; Paul, D.; Knigge, M. *Science* **1990**, *249*, 527-533.
[47] Kaldor, S.; Kalish, V.; Davies, J.; Shetty, B.; Fritz, J.; Appelt, K.; Burgess, J.; Campanale, K.; Chirgadze, N.; Clawson, D.; Dressman, B.; Hatch, S.; Khalil, D.; Kosa, M.; Lubbehusen, P.; Muesing, M.; Patick, A.; Reich, S.; Su, K.; Tatlock, J. *J. Med. Chem.* **1997**, *40*, 3979-3985.
[48] Lam, P.; Jadhav, P.; Eyerman, C.; Hodge, C.; Ru, Y.; Bacheler, L.; Meek, J.; Otto, M.; Rayner, M.; Wong, Y. *Science* **1994**, *263*, 380-384.
[49] Roberts, N.; Martin, J.; Kinchington, D.; Broadhurst, A.; Craig, J.; Duncan, I.; Galpin, S.; Handa, B.; Kay, J.; Krohn, A.; Lambert, R.; Merrett, J.; Mills, J.; Parkes, K.; Redshaw, S.; Ritchie, A.; Taylor, D.; Thomas, G.; Machin, P. *Science* **1990**, *248*, 358-361.
[50] Chan, D.; Laughton, C.; Queener, S.; Stevens, M. *J. Med. Chem.* **2001**, *44*, 2555-2564.
[51] Whitlow, M.; Howard, A.; Stewart, D.; Hardman, K.; Kuyper, L.; Baccanari, D.; Fling, M.; Tansik, R. *J. Biol. Chem.* **1997**, *272*, 30289-30298.
[52] Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. *Cell* **1998**, *95*, 927-937.
[53] Goodford, P. *J. Med. Chem.* **1985**, *28*, 849-857.
[54] Wade, R.; Goodford, P. *J. Med. Chem.* **1993**, *36*, 148-156.
[55] Varghese, J. *Drug Dev. Res.* **1999**, *46*, 176-196.
[56] Varghese, J.; Epa, V.; Colman, P. *Protein Sci.* **1995**, *4*, 1081-1087.
[57] Glick, M.; Robinson, D.; Grant, G.; Richards, W. G. *J. Am. Chem. Soc.* **2002**, *124*, 2337-2344.
[58] Mourez, M.; Kane, R.; Mogridge, J.; Metallo, S.; Deschatelets, P.; Sellman, B.; Whitesides, G.; Collier, R. *Nat. Biotechnol.* **2001**, *19*, 958-961.
[59] Richards, W. G. *Nat. Rev. Drug Disc.* **2002**, *1*, 551-555.
[60] Arkin, M.; Wells, J. *Nat. Rev. Drug Discov.* **2004**, *3*, 301-317.
[61] Enyedy, I.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X.; Cao, Y.; Guo, R.; Li, B.; Zhu, X.; Huang, Y.; Long, Y.; Roller, P.; Yang, D.; Wang, S. *J. Med. Chem.* **2001**, *44*, 4313-4324.
[62] Arkin, M.; Randal, M.; Delano, W.; Hyde, J.; Luong, T.; Oslob, J.; Raphael, D.; Taylor, L.; Wang, J.; McDowell, R.; Wells, J.; Braisted, A. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1603-1608.
[63] Braisted, A.; Oslob, J.; Delano, W.; Hyde, J.; McDowell, R.; Waal, N.; Yu, C.; Arkin, M.; Raimundo, B. *J. Am. Chem. Soc.* **2003**, *125*, 3714-3715.
[64] Taylor, R.; Jewsbury, P.; Essex, J. *J. Comput. Aided Mol. Des.* **2002**, *16*, 151-166.
[65] Carlson, H. *Curr. Op. Chem. Biol.* **2002**, *6*, 447-452.
[66] Bohacek, R.; McMartin, C. *Curr. Op. Chem. Biol.* **1997**, *1*, 157-161.
[67] Joseph-McCarthy, D. *Pharm. Ther.* **1999**, *84*, 179-191.
[68] Shoichet, B.; Leach, A.; Kuntz, I. *Proteins* **1999**, *34*, 4-16.
[69] Pang, Y.-P.; Perola, E.; Xu, K.; Prendergast, F. *J. Comp. Chem.* **2001**, *22*, 1750-1771.
[70] DesJarlais, R.; Sheridan, R.; Dixon, J.; Kuntz, I.; Venkataraghavan, R. *J. Med. Chem.* **1986**, *29*, 2149-2153.
[71] Ewing, T.; Makino, S.; Skillman, G.; Kuntz, I. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411-428.
[72] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996**, *261*, 470-489.
[73] Kramer, B.; Metz, G.; Rarey, M.; Lengauer, T. *Med. Chem. Res.* **1999**, *9*, 463-478.
[74] Goodsell, D.; Morris, G.; Olson, A. *J. Mol. Recognit.* **1996**, *9*, 1-5.
[75] Caflisch, A.; Niederer, P.; Anliker, M. *Proteins* **1992**, *13*, 223-230.
[76] Liu, M.; Wang, S. *J. Comput. Aided Mol. Des.* **1999**, *13*, 435-451.
[77] Trosset, J.; Scheraga, H. *J. Comput. Chem.* **1999**, *20*, 412-427.
[78] Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727-748.
[79] Boehm, H. *J. Comput. Aided Mol. Des.* **1992**, *6*, 61-78.
[80] Bohm, H.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. *J. Med. Chem.* **2000**, *43*, 2664-2674.
[81] Gryzbowski, B.; Ishchenko, A.; Kim, C.; Topalov, G.; Chapman, R.; Christianson, D.; Whitesides, G.; Shakhnovich, E. *Proc. Natl. Acad. Sci.* **2002**, *99*, 1270-1273.
[82] Anderson, A. *Chem. Biol.* **2003**, *10*, 787-797.
[83] Carlson, H.; Masukawa, K.; Jorgensen, W.; Lins, R.; Briggs, J.; McCammon, J. *J. Med. Chem.* **2000**, *43*, 2100-2114.
[84] Schnecke, V.; Swanson, C.; Getzoff, E.; Tainer, J.; Kuhn, L. *Proteins* **1998**, *33*, 74-87.
[85] Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. *J. Mol. Biol.* **2001**, *308*, 377-395.
[86] Ota, N.; Agard, D. *J. Mol. Biol.* **2001**, *314*, 607-617.
[87] Brooijmans, N.; Kuntz, I. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335-373.
[88] Charifson, P.; Corkery, J.; Murcko, M.; Walters, W. *J. Med. Chem.* **1999**, *52*, 5100-5109.
[89] Perola, E.; Walters, W.; Charifson, P. *Proteins* **2004**, *56*, 235-249.
[90] Teague, S.; Davis, A.; Leeson, P.; Oprea, T. *Angew. Chem. Int. Ed.* **1999**, *38*, 3743-3747.
[91] Wu, S. Y.; McNae, I.; Kontopidis, G.; McClue, S. J.; McInnes, C.; Stewart, K. J.; Wang, S. D.; Zheleva, D. I.; Marriage, H.; Lane, D. P.; Taylor, P.; Fischer, P. M.; Walkinshaw, M. D. *Structure* **2003**, *11*, 399-410.
[92] Wang, S. D.; Meades, C.; Wood, G.; Osnowski, A.; Anderson, S.; Yuill, R.; Thomas, M.; Mezna, M.; Jackson, W.; Midgley, C.; Griffiths, G.; Fleming, I.; Green, S.; McNae, I.; Wu, S. Y.; McInnes, C.; Zheleva, D.; Walkinshaw, M. D.; Fischer, P. M. *J. Med. Chem.* **2004**, *47*, 1662-1675.
[93] Gruneberg, S.; Stubbs, M.; Klebe, G. *J. Med. Chem.* **2002**, *45*, 3588-3602.
[94] Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337-356.
[95] Verdonk, M.; Cole, J.; Taylor, R. *J. Mol. Biol.* **1999**, *289*, 1093-1108.
[96] Tondi, D.; Slomczynska, U.; Costi, M.; Watterson, D. M.; Ghelli, S.; Shoichet, B. *Chem. Biol.* **1999**, *6*, 319-331.
[97] Fritz, T.; Tondi, D.; Finer-Moore, J.; Costi, M.; Stroud, R. *Chem. Biol.* **2001**, *8*, 981-995.
[98] Kick, E.; Roe, D.; Skillman, G.; Liu, G.; Ewing, T.; Sun, Y.; Kuntz, I.; Ellman, J. *Chem. Biol.* **1997**, *4*, 297-307.
[99] Kraemer, O.; Hazemann, I.; Podjarny, A. D.; Klebe, G. *Proteins* **2004**, *55*, 814-823.
[100] Lyne, P. D.; Kenny, P. W.; Cosgrove, D. A.; Deng, C.; Zabludoff, S.; Wendoloski, J. J.; Ashwell, S. *J. Med. Chem.* **2004**, *47*, 1962-1968.
[101] Babaoglu, K.; Page, M. A.; Jones, V. C.; McNeil, M. R.; Dong, C. J.; Naismith, J. H.; Lee, R. E. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3227-3230.
[102] Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, W.; Clarkson, B.; Kuriyan, J. *Science* **2000**, *289*, 1938-1942.

[103] Peng, H.; Huang, N.; Qi, J.; Xie, P.; Xu, C.; Wang, J.; Yang, C. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3693-3699.
[104] Schapira, M.; Raaka, B. M.; Das, S.; Fan, L.; Totrov, M.; Zhou, Z. U.; Wilson, S. R.; Abagyan, R.; Samuels, H. H. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 7354-7359.
[105] Schapira, M.; Raaka, B.; Samuels, H.; Abagyan, R. *BMC Struct. Biol.* **2001**, *1*, 1.
[106] Bissantz, C.; Bernard, P.; Hibert, M.; Rognan, D. *Proteins* **2003**, *50*, 5-25.
[107] Evers, A.; Klebe, G. *Angew. Chem. Int. Ed.* **2004**, *43*, 248-251.
[108] Liu, H.; Li, Y.; Song, M. K.; Tan, X. J.; Cheng, F.; Zheng, S. X.; Shen, J. H.; Luo, X. M.; Ji, R. Y.; Yue, J. M.; Hu, G. Y.; Jiang, H. L.; Chen, K. X. *Chem. Biol.* **2003**, *10*, 1103-1113.
[109] Doyle, D. A.; Cabral, J. M.; Pfuetzner, R. A.; Kuo, A. L.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. *Science* **1998**, *280*, 69-77.
[110] Gao, Y.; Dickerson, J. B.; Guo, F.; Zheng, J.; Zheng, Y. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7618-7623.
[111] Huang, N.; Nagarsekar, A.; Xia, G. J.; Hayashi, J.; MacKerell, A. D. *J. Med. Chem.* **2004**, *47*, 3502-3511.

Received: September 16, 2004 **Accepted: December 1, 2004**