Strategies for Compound Selection

Marius M. Olah, Cristian G. Bologa and Tudor I. Oprea*

Division of Biocomputing, University of New Mexico School of Medicine, BMSB61, MSC 08 4670, 1 University of New Mexico, Albuquerque, NM 87131-0001

Abstract: In-house pharmaceutical collections are no longer sufficient for sampling chemical spaces. As novel bioactive chemotypes are successfully identified by virtual and high -throughput screening, the ability to rapidly sift through large numbers of chemicals prior to acquisition or experiment is required. Strategies for compound selection include some of the following steps: 1.) database assembly ('in silico' inventory); 2a.) structural integrity verification (keep unique structures only); 2b.) limited exploration of alternative chemical representations for the uniques (stereoisomers, tautomers, ionization states); 3.) property and structural filtering (remove unwanted structures); 4.) 3D-structure generation (for virtual screening or 3D-based similarity); 5a.) clustering or statistical design for selection; 5b.) similarity-based selection (if bioactives are known); 5c.) receptor-based selection (if target binding site is known); 6.) add a random subset to the final list.

Key Words: Cheminformatics, drug discovery, high-throughput screening, leadlikeness, property filtering, unwanted structures, virtual screening.

1. INTRODUCTION

Efforts to be the "first in class" or the "best in class", with respect to the drug market, are putting a severe pressure on the pharmaceutical industry as a whole. There is a disconnection between cellular and animal models (in vitro and *in vivo*) and human disease [1], and a diminishing interest in good clinical research – an observation valid three decades ago [2]. Given current regulations, there is however no alternative to the study of small molecules in surrogate models, using a vast array of methods ranging from ligand-receptor in vitro assays to humanized (transgenic) animals, before a compound is progressed from preclinical status to Phase I. In the effort to reduce costs, screen more compounds and reduce the time from idea to drug, the industry embarked on a novel paradigm, best represented by high -throughput screening (HTS) and multiple parallel synthesis (MPS) and combinatorial chemistry. There were, however, no guarantees for success, and the unrestricted MPS-based exploration of chemical space during the early 1990s produced a wide array of rather large, hydrophobic and low-solubility compounds [3].

This practical observation was followed by the suggestion [3] to restrict small molecule synthesis in the property space defined by ClogP (octanol/water partition coefficient [4]), MW (molecular weight), HDO (number of hydrogen bond donors) and HAC (number of hydrogen bond acceptors). Lipinski *et al.* [3] analyzed 2245 compounds from the World Drug Index [5] that had reached phase II clinical trials or higher, and looked at the 90th percentile for the distribution of MW ($\leq$ 500), ClogP ($\leq$ 5), HDO ($\leq$ 5) and the sum of nitrogen and oxygen, accounting for hydrogen bond acceptors (HAC $\leq$ 10). Natural products and actively transported molecules were excluded from this analysis. They concluded that if any two of the above conditions are violated, the molecule is unlikely to be an orally active drug. The pharmaceutical industry became aware of that "rule-of-five" (RO5) compliance was needed, in particular for oral availability. This significantly changed the MPS library design criteria to include RO5 filters. Computer-based perception of "druglikeness" [6, 7] provided the ability to discriminate between "drugs" and "nondrugs" based on chemical fingerprints. It was by default assumed that RO5 criteria are also "druglike", but this is incorrect [8]. As HTS campaigns frequently yield micromolar hits that are not RO5 compliant, hence not easily amenable to medicinal chemistry optimization [9], it was clearly noted that the aim of MPS and HTS programs is to identify *chemical leads,* not drugs [10].

Revising the RO5 criteria for leadlikeness [11] leads to more restrictive values: MW $\leq$ 460, -4 $\leq$ ClogP $\leq$ 4.2, HDO $\leq$ 5 and HAC $\leq$ 9. Further filters include aqueous solubility (LogSw $\geq$ -5), flexibility (number of rotatable bonds, RTB $\leq$ 10), and complexity (number of rings, RNG $\leq$ 4). Surveying a decade of medicinal chemistry literature (1991-2000), captured in WOMBAT 2004.1 [12] shows that out of 68,543 unique molecules having over 143,000 reported activities, about two-fifths (23,107 molecules) of the structure meet the above leadlike criteria. High-activity lead-like molecules (activity $\geq$ 10 nM) represent 13.35% of this dataset, in contrast to 47% low-activity molecules (activity < 1 μM). The remainder (39.65%) are medium-activity (1 μM $\leq$ activity < 10 nM) molecules. The biological source for the targets, as indicated in Figure (**1**), shows that a significant proportion of the high-activity leadlike molecules are active on human receptors and enzymes. Further examining Figure (**1**), it becomes apparent that there is a shortage of leadlike

*Address correspondence to this author at the Division of Biocomputing, University of New Mexico School of Medicine, BMSB61, MSC 08 4670, 1 University of New Mexico, Albuquerque, NM 87131-0001. Tel.: +1 (505) 272-3694; Fax: +1 (505) 272-8738; E-mail: toprea@salud.unm.edu

1570-1638/04 $45.00+.00

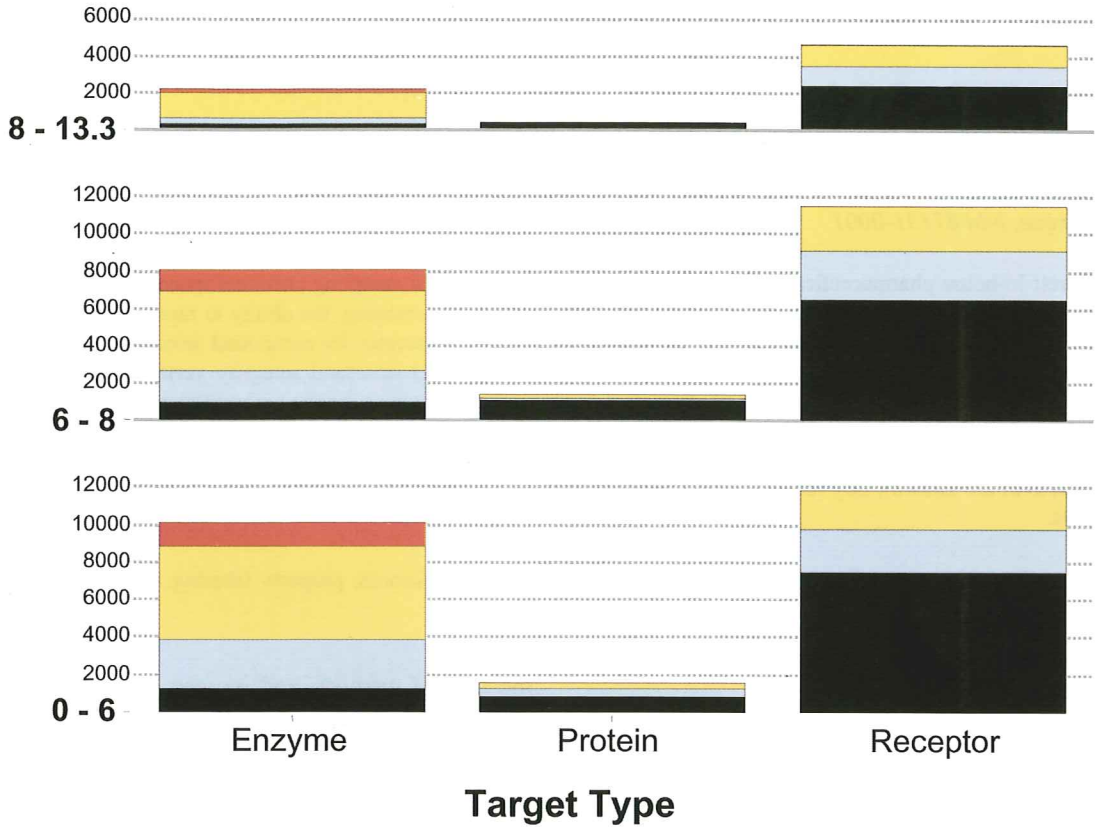


Fig. (1). Binned activities (on the $-\log_{10}$ scale), target type and target source breakdown for 51,083 biological activities reported for 23,107 unique leadlike-compliant molecules from WOMBAT [12] that meet the computational leadlike criteria. Colors indicate the biological source for the *in vitro* target: Bacterial, parasitic and viral targets in red, human targets in yellow, rat targets in black; all other mammalian targets (not rat, not human) are in magenta. Human recombinant targets expressed in other systems are also binned as 'human'.

molecules with high activity on bacterial, parasitic or viral targets (all of them being enzymes), and that a significant fraction (almost 40%) of the lead-like molecules are tested on rat receptors. The data captured in Figure (**1**) indicates that there are a significant number of chemical leads active on human targets that meet the computational leadlike criteria, justifying the introduction of leadlike compliance as one of the compound selection strategies discussed here.

This paper highlights cheminformatics strategies for compound acquisition, focusing on database assembly, "garbage" removal [13] and chemical structure integrity, property and structural filtering, followed by clustering and statistical molecular design during the selection step. Chemical similarity and virtual screening are also discussed, since they can significantly influence the selection strategy. Issues particular to compound selection prior to virtual screening [14], and a step-by-step procedure for compound acquisition [15] are addressed elsewhere.

2. RESULTS AND DISCUSSION: STRATEGIES FOR ACQUISITION

HTS can sift through a large amount of structures, often based on single-dose, single-experiment results. Its in silico equivalent [16, 17], virtual screening (VS), can also sift through a possibly even larger amount of structures, in order to rapidly prioritize structures of interest for bioscreening. Both procedures rely on our ability to process a large number of structures using cheminformatics [18]. However, post-HTS analyses [19] are often clouded by the presence of reactive species or optically interfering components (which can be the result of sample degradation) in biochemical assays [20], the tendency of chemicals to aggregate [21] or to turn up as frequent hitters [22]. Hence, a key step prior to compound acquisition is to generate computational filters geared to remove "unwanted" molecular species, as discussed below. This has to be performed in conjunction with the systematic treatment conjugated systems, ionization states, and tautomers.

As the property distribution of drugs in the RO5 space is well understood with respect to 'nondrugs' [8], suitable consideration needs to be given to those key properties that have, repeatedly over time, emerged as important in bioactivity. Hansch and co-workers observed [23] that more than 60% of the 8,500 biological QSAR (Quantitative Structure-Activity Relationships) equations stored in the C-QSAR database [24] contain either the ClogP [4] term (4,614 equations) or the π [25] constant (784 equations). Hence, hydrophobicity, size and other RO5–related (donors,

Table 1. On-Line Databases Containing Commercially Available Chemicals.

Company	Number of compounds	Brief description	Company website
4SC	4,600,000	virtual library; small molecule drug candidates	http://www.4sc.de/
Asinex	350,000	Platinum and Gold collections	http://www.asinex.com/prod/index.html
BioFocus plc	100,000	diverse primary screening compounds	http://www.biofocus.com/pages/drug__discovery.mhtml
CEREP	>350 chemical families	Odyssey II library: diverse and unique discovery library	http://www.cerep.fr/Cerep/Utilisateur/Download/frs_Download.asp
Chemical Diversity Labs, Inc.	>650,000	lead-like compounds for bioscreening	http://www.chemdiv.com/discovery/downloads/
Com Genex	260,000	'pharma relevant', discrete structures for multitarget screening purposes	http://www.comgenex.hu/cgi-bin/inside.php?in=products&l_id=compound
InterBio Screen	288,000	synthetic compounds	http://www.ibscreen.com/products.shtml
InterBio Screen	30,000	natural compounds	
Maybridge plc	60,000	organic drug-like compounds	http://www.maybridge.com/html/m_company.htm
Maybridge plc	13,000	building blocks	
Sigma-Aldrich	130,000	diverse library of small-molecule compounds, including plant extracts and microbial cultures	http://www.sigmaaldrich.com/Area_of_Interest/Drug_Discovery/High_Throughput_Screening/Compound_Libraries.html
Specs	230,000	diverse library	http://www.specs.net/
Specs	10,000	World Diversity Set: preplateled library	
Tripos	80,000	LeadQuest compound libraries	http://www.tripos.com/sciTech/researchCollab/chemCompLib/lqCompound/index.html

acceptors) or unrelated properties (e.g., flexibility [8], complexity [26]) require consideration.

- **Database Assembly**

Pharmaceutical companies have large compound collections that contain a significant number of bioactive molecules, e.g., clinical candidates, launched drugs and other high-activity compounds. A valuable resource, these are routinely screened against novel targets. However, these collections tend to reflect the chemistry used to address current and past biological targets, while novel targets – the argument goes – may require novel chemistry. Appropriate cheminformatics tools are required to assemble compound databases in preparation for acquisition. The largest collection of accessible (existing or virtual) molecules is assembled by ChemNavigator's iResearch™ Library [27] (over 10 million unique chemicals) from over 148 chemical vendors. Others make their collections available on-line – see Table **1**.

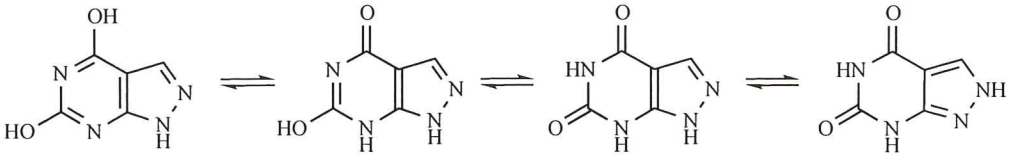
- **Database Clean-up and Structural Integrity**

Efforts to clean up the database at an early stage are warranted, since all subsequent steps rely on the quality of chemical structural representations. Pipeline Pilot [28], CCG [29] and FILTER [30] provide commercial solutions for this process. Canonical isomeric SMILES (e.g., Daylight's *cansmi*) should be generated using structural normalization (e.g., nitro, keto-enol, etc.), and whatever standard is utilized on in-house chemical databases should be imposed, in order to maintain chemical structure compatibility. Some user-defined steps may be required for this process, since chemical inventory systems require additional information, e.g., synthetic batch number, yield, purity and salt formulation, information that is not necessarily relevant from an acquisition strategy perspective. In fact, it is recommended that before any additional processing, chemical structures are normalized, the correct valency is checked, counter-ions are removed, and formal charges are standardized.

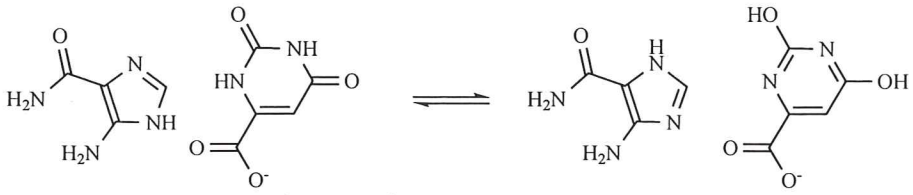
Particular care should be exercised to the 2D representation of chiral information, since the chemist-friendly representation with two up/down wedges at the same chiral center is not machine-readable and leads to 3D conversion errors [31]. This can become a significant source of errors in structure-activity relationship studies [32, 33]. When computational and software resources are not an issue, two separate runs, using independent software implementations, are recommended [14]. Duplicates should be removed and only unique structures should be processed further. At this stage, the introduction of practical aspects (e.g., price, purity, preferred chemical vendor, etc.) is left to the end-user. Typically, such topics are decisive once the final list of compounds is compiled, although the issue of availability is perhaps better addressed early, to reduce the computational burden.

- **Limited Exploration of Alternative Structures**

Chemical structures are, sometimes, amenable to alternative representations, e.g., different ionization states and tautomeric structures may be possible [34]. Alternate



Alternative tautomeric structures for Oxypurinol



Alternative tautomeric structures for Orezamide

Fig. (2). Examples of launched drug structures that are amenable to tautomeric representations for Oxypurinol (xanthine oxidase inhibitor) and Orezamide (choleretic). The keto and enol forms of pyrimidine-dione component are just the two illustrative examples. In fact, there are 6 possible tautomeric forms of the pyrimidine-dione compoment of Orezamide, hence 12 possible tautomeric structures can be enumerated for both components of this drug.

structures should be considered prior to database processing for, e.g., 3D-based similarity or diversity analyses. The results will influence the acquisition strategy and the results of virtual screening. Starting from canonical isomeric SMILES, alternate structures should be explored for:

- Tautomers, by shifting a hydrogen atom along a path of alternating single/double bonds involving nitrogens and oxygens. Figure (**2**) shows four alternatives for Oxypurinol, and two for the two-compound Orezamide (both structures allowing tautomeric representations);

- Acid/base equilibria, which explore different protonation states by assigning formal charges to those chemical moieties that are likely to be charged (e.g., phosphate or guanidine) and by assigning charges to some of those moieties that are likely to be charged under different micro-environmental conditions ("chargeable" moieties such as tetrazole and aliphatic amine);

- Unspecified chirality (e.g., using Daylight's *chiralify*), since 3D structure conversion from SMILES is frequently seeking a low-energy representation. For molecules having 3 or more chiral centers, 'limited exploration' indicates that contextual judgment needs to be applied on a case-by-case basis. Other examples include pseudo-chiral centers such as pyramidal nitrogen inversions that explore non-charged, non-aromatic, pseudo-chiral nitrogens (3 substituents), as they easily flip in three dimensions.

More accurate computations for tautomers and acid/base equilibria involve the correct partition of various ionized micro-species, that in turn involves pK_a estimation for various tautomers per compound. Computational tools that estimate pK_a values and the microenvironment dependent tautomeric ratio are expected to be less reliable. While speed remains the trade-off vs. accuracy of prediction, this is an advantage when evaluating large numbers of compounds.

- **"Unwanted" Structure and Property Filtering**

The list of "unwanted" chemical structures typically includes metals and isotopes, as well as other structures deemed to interfere with biochemical assays under HTS assay conditions [20]. Acyl-halides, sulfonyl-halides, Michael acceptors, etc. are typically removed using the SMARTS substructure language [35] – some such examples are given in Figure (**3**). "Unwanted" structures should always be judged in context: Some anti-neoplastic agents are alkyl-halides and some flavor compounds are monoaldehydes, both deemed 'reactive species' in Fig. (**3**). If seeking actives in oncology or flavor science is the goal, the "unwanted" species list has to be reconsidered. By the same token, steroids are "unwanted" in some companies that deliberately avoid them due to over-patenting and high cost of synthesis for large scale production, while other companies regard them as "privileged structures". In addition, one may remove known promiscuous binders [21] or frequent hitters [22], as well as compounds that contain a high percentage of fragments believed to be involved in cytotoxicity – see Figure (**3**).

Additional suggestions for compound exclusion include:

- More than 3 fused aromatic rings, since most polyaromatic rings are processed by cytochrome P450 enzymes into epoxides and quinones that can cause direct DNA damage;

- More than 4 halogens (but up to 2 $-CF_3$ groups) in order to avoid "pesticides". In crop protection research, the database processing step is very likely to differ in the "unwanted" list and property profiling, although the rest of the strategy is probably the same;

- More than eight $-CH_2$ groups in an unbranched (highly flexible) chain;

- More than 22 atoms in a macrocyclic ring, or more than 10-14 flexible bonds in a macrocycle (if *not* screening natural products);

- Compounds having too few (e.g., C, H, X and Si) or too many O and N atoms (including hydrazines, azides, diazonium salts, and pyrillium salts);

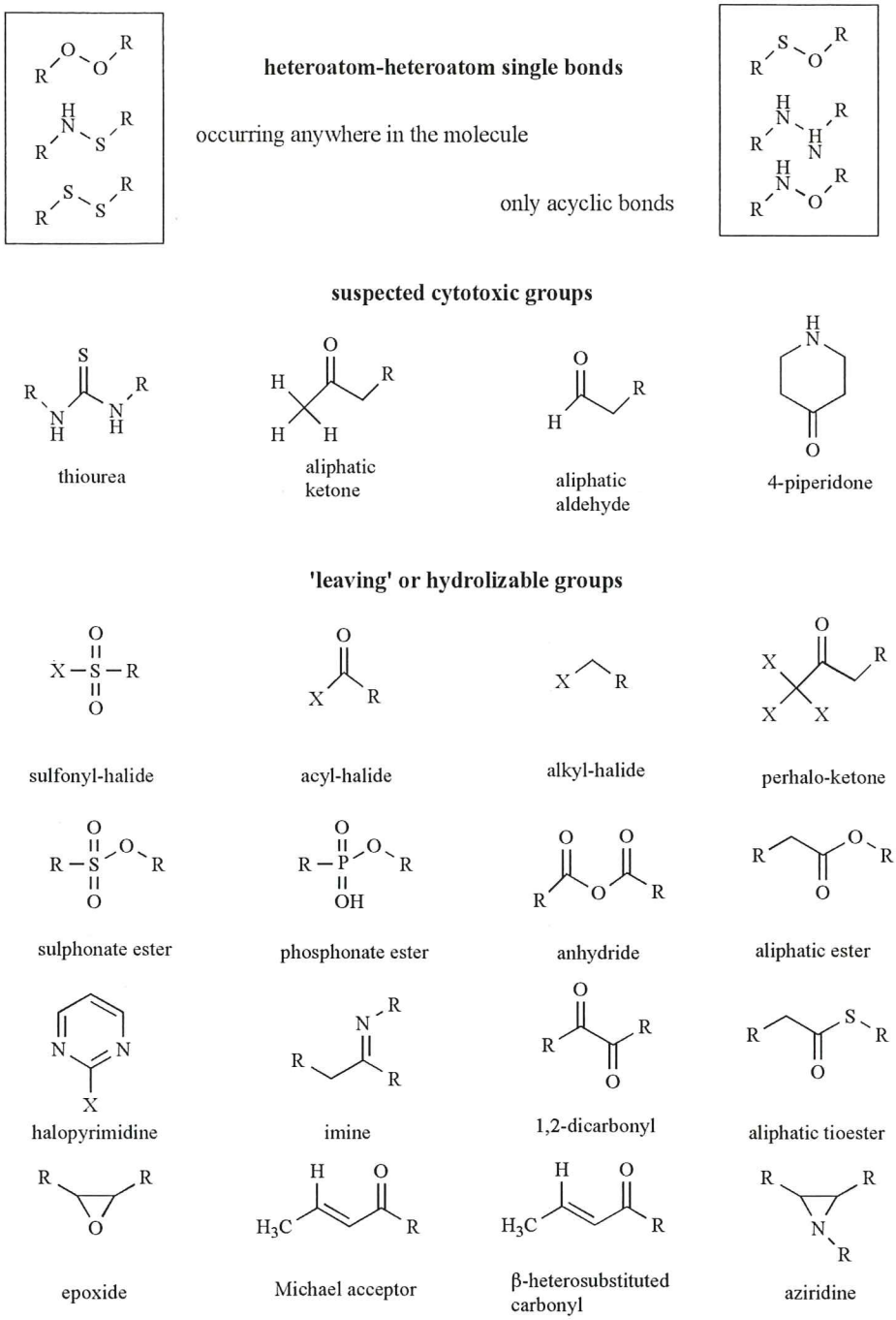


Fig. (3). Examples of chemical substructures that can cause interference with biochemical assays [20]. Modified from Ref. [8] with permission.

• More than 5 analogs per series (e.g., methyl, ethyl, propyl) – which is usually addressed via clustering and diversity-based selection.

Property filtering is, likewise, a target-, process- and company-dependent procedure. We routinely remove a significant percentage of the compounds that do not meet the computational leadlike criteria [11] discussed above. One may regard these property filters as excessive, as it is exemplified in Figure (**1**), where 4289 unique SMILES are high-activity, leadlike molecules. On the other hand, since 47% of the leadlike molecules have low activity, the remaining 53% (14,931 unique SMILES) have sub-micromolar activity in at least one instance. Compounds that pass the leadlike property filters – between 30% to 90%, depending on the context – are prioritized for acquisition.

One drawback when imposing filtering criteria for an entire database is that, whatever the criteria are, they influence the entire HTS / VS effort for that collection, so there has to be a convergence of goals regarding the targets and mode of administration for that collection. The

collection may, at times, require subsetting with different criteria: For inhalation therapy, e.g., targets located in the respiratory system, the pharmacokinetic profile will differ from that of orally administered drugs, which in turn will differ from topically applied dermatological agents. Such biases need to be introduced as much as possible at the property filtering stage, because they reduce the size of the chemical space that needs to be sampled.

- **3D Structure Generation**

For 3D-based chemical similarity selection and for virtual screening pre-processing, one or more conformers need to be investigated per molecule. Some docking software, e.g., FRED [36], can generate 3D structures from SMILES prior to actual docking. Other virtual screening processors require a separate 3D conversion step [37], e.g., using CONCORD [38], CORINA [39] or RUBICON [40]. Multiple conformations can be explored using Catalyst [41], Confort [42] and OMEGA [43]. Improper chirality information can be addressed with StereoPlex [42], a software that makes "educated guesses" about the chiral centers that require systematic 3D exploration. However, limited exploration of multiple stereoisomers is advised, as discussed above.

- **Compound Clustering and Statistical Molecular Design**

Often too large to be evaluated in detail, compound databases require a chemical fingerprints approach for clustering or statistical molecular design (SMD). Clustering methods group molecules into "families" (clusters) of related structures, which are perceived – at a given resolution – to be different from other chemical families. With clustering, the end-user has the ability to select one or more representatives from each family. SMD methods aim at sampling various areas of chemical space and selecting representatives from each area. Some software is designed to select compounds from multi-dimensional spaces, e.g., LibraryExplorer [44], which is based on the BCUT metric [45, 46].

Clustering algorithms can be classified using several criteria and implemented in several ways – a short list of such software is given in the Experimental Section. Due to its computational simplicity, hierarchical clustering has been used to a greater extent, but nonhierarchical methods are increasingly used more for chemical structure classification. Implementation choices (input descriptors, similarity measure, clustering algorithm) will influence the outcome, and more often than not CPU-time, hardware capability and software availability will be the deciding factors. Mesa's [47] implementation of the Taylor [48] - Butina [49] algorithm, which can perform asymmetric clustering and deal with 'false singletons' (borderline compounds often assigned to one of at least two equally-distant chemical families) has become our method of choice.

Another rational method for compound selection, SMD can be applied to select chemical representatives, e.g., for building block selection in MPS planning [50]. Part of the larger category of experimental design methods [51], e.g., fractional factorial or composite designs, SMD is typically preferred when selecting a small screening deck (for either HTS or VS) out of the larger solution space.

- **Ligand-Based and Receptor-Based Selection**

Selecting compounds similar to known bioactives has always been considered as a good strategy for finding many actives per screening deck. However, this strategy only works for targets where active ligands are available, and it is not guaranteed to lead to different chemotypes. The situation is more difficult when the known actives are patent-protected, or if the key chemotype is toxic. Whenever known actives are available, they should serve as positive controls in the screening deck even though they might have been removed in prior steps. These molecules should be recovered at the end of the virtual or high -throughput screen as "hits".

The chemical similarity principle states that ligands with similar features are expected to have similar biologic activity [52]. It is expressed as a number that quantifies the "distance" between a pair of compounds (dissimilarity, or 1 minus similarity), or how related the two compounds are (similarity). By definition, similarity needs a reference: That of a descriptor system (a metric by which similarity is judged), as well as that of an object, or class of objects – we need a reference point to which objects can be compared with [53]. Similarity depends on the choice of molecular descriptors [54], the choice of the weighting scheme(s) and the similarity coefficient itself. The coefficient is typically based on Tanimoto's symmetric distance-between-patterns [55], and on Tversky's asymmetric contrast model [56]. Multiple types of methods are available for chemical similarity evaluation [52, 57-60].

Since multiple strategies to select chemicals exist, we provide for illustrative purposes the highest similarity scores, given the sub-structure Tversky similarity (alpha 0.1) and super-structure Tversky similarity (alpha 0.9) against estradiol, using 320-bit MDL keys [61] – see Figure (**4**). The highlighted substructure (blue dot) is a potent (11 nM) antagonist on the estrogen receptor [62], whereas the most similar superstructure (red dot) is a 0.3 μM inhibitor of carbonic-anhydrase II [63]. This illustrates the conundrum of chemical similarity: The phenol moiety, a known 'signature' for many estrogen-receptor binding ligands, explains the presence of ~ 100 blue dots (tested and active on the estrogen receptor) in the upper right quadrant of Figure (**4**). These molecules are, however, not 'similar' according to the symmetric Tanimoto similarity index (all Tanimoto values were below 0.7, data not shown), nor are they the highest ranking molecules according to the superstructural Tversky similarity. While the remaining datasets are understandably in the lower quadrants in Figure (**4**) – Thrombin inhibitors in dark green, HIV-1 protease inhibitors in yellow, and intercellular adhesion molecule 1, ICAM-1, inhibitors in black, – there is a red cluster with high superstructural similarity to estradiol. These 30 molecules are bile acid derivatives [63] that contain the steroid skeleton – explaining their ranking. However, these 30 molecules also have the para-phenyl-sulfonamide 'signature', typical for carbonic anhydrase inhibitors. It is recommended to combine

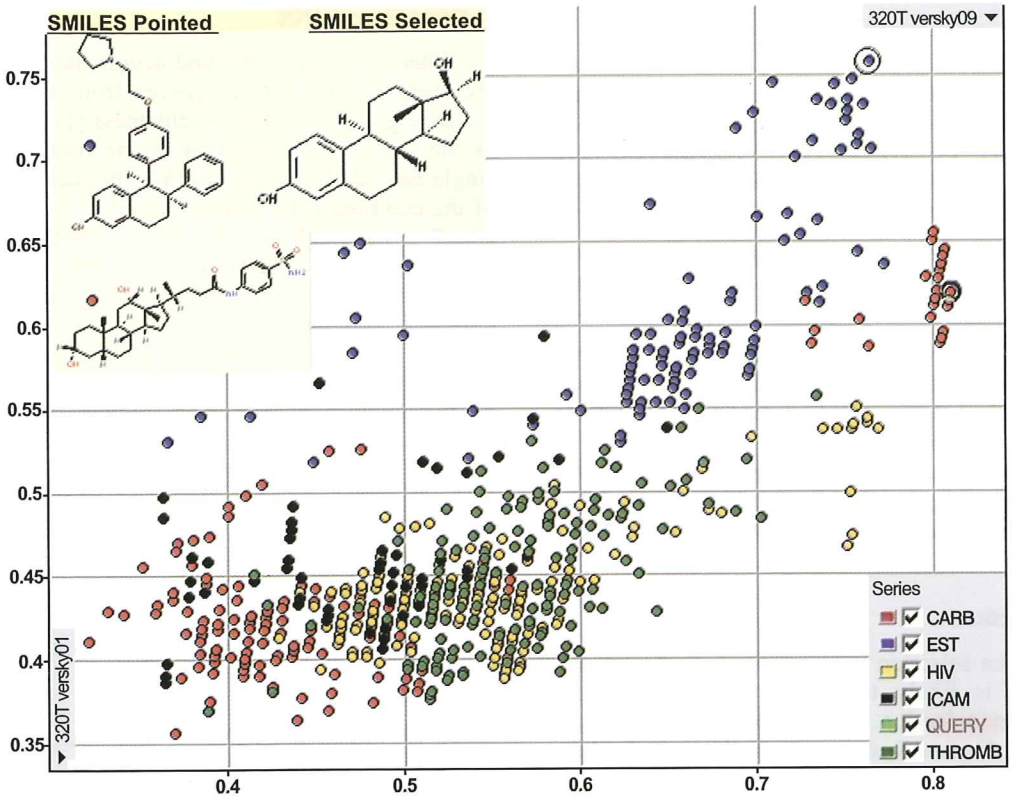


Fig. (4). Tversky similarity scores based on the 320 MDL keys [61] on an artificial dataset from WOMBAT [12]. For the query estradiol *(SMILES selected, above),* the most similar substructure (highest Tversky 0.1, blue highlight) and the most similar superstructure (highest Tversky 0.9, red highlight) are depicted in the yellow inset. See text for details.

similarity measures during selection and to include 'signature' moieties in the similarity search, whenever available.

For 3D-similarity, multiple conformers per molecule are recommended. ROCS (Rapid Overlay of Chemical Structures [64]) is a Gaussian-shape volume overlap filter that can rapidly identify shapes that match the query molecule. If the shape-selection algorithm is computationally efficient, the entire set of conformations should be explored for each query. This could be complemented by pharmacophore perception methods, preferably on those molecules that have significant shape similarity with the query molecule. For example, ROCS can be complemented by ALMOND [65], which encodes pharmacophoric elements (hydrophobic, H-bond donor and H-bond acceptor interaction fields computed with GRID [66]) into a reduced set of variables. ALMOND can 'fine-tune' the 3D-based similarity selection by adding pharmacophore-based similarity to an already matching shape. ROCS/ALMOND selected subset should be prioritized over other methods – with the possible exception of molecules selected using known ligand-receptor structures.

Whenever the experimentally-determined structure of the biological target is available (preferably co-crystallized with a bioactive ligand), this strategy can be applied to select additional molecules for acquisition. Some of the most popular docking routines are DOCK [67], GOLD [68], FlexX [69], AutoDock [70] and FRED [36]. The limiting step in this strategy is accurate scoring. There are four categories of scoring functions:

(i) Knowledge-based methods [71-73] based on Boltzmann-weighted Potentials of Mean Force (PMF) derived from statistical analyses of ligand-receptor inter-atomic contacts from PDB (Protein Data Bank [74, 75]) complexes. SMoG [76-78], Muegge [79, 80], and DrugScore [81,82] are examples of such scoring functions.

(ii) "Master equation" approaches [83] estimate the energetic contributions of various interaction types in

a semi-quantitative manner, e.g., Williams [84,85] and Rose [86,87]. Williams [84] targets peptide-peptide interactions [88], while Rose devised a scoring function intended for high-throughput virtual screening.

(iii) Regression-based methods [89,90] train on bioactives with known binding mode using sets of ligand-receptor complexes from the Protein Data Bank. See SCORE2 [91], VALIDATE [92], VALIDATEII [93], Pro_Score [94,95], Jain [96], Horvath [97] and SCORE [98,99], for example.

(iv) ZAP [100] is a Poisson-Boltzmann equation solver that scores ligands while incorporating solvent effects.

Since none of the above scoring schemes has been shown to be general, consensus scoring [101] can replace individual schemes. A novel steroidal scaffold binding to β-tubulin, an anti-cancer target previously known to bind taxol and its derivatives [102] was identified after virtually screening the NCI (National Cancer Institute) open database [103]. Other confirmed hits from virtual screening are reviewed elsewhere [104,105].

- **Random Selection**

Regardless of the selection strategy, the output from any selection method is likely to be complemented by an additional, random selection. Molecules that violate the exclusion criteria discussed above, including non-leadlike molecules can, and should, be entered in the final list of compounds (up to 30%). The procedure is recommended when little or no information is known about the target and its bioactives, since its random selection can perform equally well in such cases [106]. Random selection is not more successful when the strategy aims at focused selection. Serendipity has always played a role in drug discovery [107], therefore a certain degree of randomness in the final selection is advised.

3. EXPERIMENTAL SECTION

Compound acquisition is typically performed using software from Daylight [40], Tripos [38], Optive Research [42], Accelrys [41], SciTegic [28] and Chemical Computing Group [29]. Useful routines are also available from OpenEye [108] and Mesa [47]. Chemical structures need to be converted from various file formats (e.g., sdf, car, mol, etc.) into canonical isomeric SMILES [109,110] or equivalent linear notations [111,112]. Chemical fingerprints can then be generated from SMILES using the following formats: Daylight [113], Unity [114], Mesa [115], BCI [116], MDL [61] or CCG [117]. If RO5 compliance is required, then software that computes chemical properties from structures is also needed. For example, LogP, the octanol/water partition coefficient [4] can be estimated using ClogP [118], ACDLabs [119] or the freely available KowWIN [120] and ALogPS [121], among many other LogP predictors. 3D-structure generation is implemented in rule-based systems, e.g., CONCORD [38], CORINA [39], RUBICON [40] or OMEGA [43], whereas chemical clustering requires specialized routines [116,122-124]. Statistical molecular design, e.g., D-optimal, is performed in multiple dimensions, e.g., as implemented in MODDE [125].

4. CONCLUSIONS

When defining compound acquisition strategies, most of the steps require active choices from the user. Literature trends, e.g., RO5 [3] and leadlikeness [11], are useful as long as the aim is compatible with the 'orally available drug, single dose daily' paradigm [53]. In such cases, some or all of the computational leadlike criteria, i.e., MW $\leq$ 460, -4 $\leq$ ClogP $\leq$ 4.2, LogSw $\geq$ -5, RTB $\leq$ 10, RNG $\leq$ 4, HDO $\leq$ 5, and HAC $\leq$ 9, should be observed. The strategies for compound acquisition can be summarized as follows: 1. database assembly ('in silico' inventory); 2a. structural integrity verification (keep unique structures only); 2b. limited exploration of alternative chemical representations for the uniques (stereoisomers, tautomers, ionization states); 3. removal of "unwanted" structures and chemical property filtering (remove unwanted structures); 4. 3D-structure generation (for virtual screening or 3D-based similarity); 5a. clustering or SMD for selection; 5b. 2D- and 3D-similarity selection (when bioactives are known); 5c. receptor-based selection (if the ligand binding mode at the target is known); 6. random selection. The list should once again be verified for duplicate removal, and then forwarded for compound acquisition.

ACKNOWLEDGEMENT

These studies were supported by New Mexico Tobacco Settlement funds.

REFERENCES

[1] Horrobin, D.F. *Nature Rev. Drug Discov.*, **2003**, *2*, 151.
[2] DeFelice, S.L. *Drug Discovery – The Pending Crisis*, MEDCOM: New York, **1972.**
[3] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. *Adv. Drug Deliv. Reviews*, **1997**, *23*, 3.
[4] Leo, A. *Chem. Rev.*, **1993**, *5*, 1281.
[5] World Drug Index database is available from Derwent Publications Ltd., http://www.derwent.com/products/lr/wdi/ and Daylight Chemical Information Systems, http://www.daylight.com/products/databases/WDI.html.
[6] Ajay; Walters, W.P.; Murcko, M.A. *J. Med. Chem.*, **1998**, *41*, 3314.
[7] Sadowski, J.; Kubinyi, H. *J. Med. Chem.*, **1998**, *41*, 3325.
[8] Oprea, T.I. *J. Comput.-Aided Mol. Design*, **2000**, *14*, 251.
[9] Teague, S.J.; Davis, A.M.; Leeson, P.D.; Oprea, T.I. *Angew. Chem. Int. Ed.*, **1999**, *38*, 3743; German version: *Angew. Chem.*, **1999**, *111*, 3962.
[10] Oprea, T.I.; Davis, A.M.; Teague, S.J.; Leeson, P.D. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1308.
[11] Hann, M.M.; Oprea, T.I. *Curr. Opin. Chem. Biol.*, **2004**, *8*, 255.
[12] WOMBAT database is available from Sunset Molecular Discovery LLC., Santa Fe, NM; http://www.sunsetmolecular.com/.
[13] Walters, W. P.; Stahl, M.T.; Murcko, M.A. *Drug Discov. Today*, **1998**, *3*, 160.
[14] Oprea, T.I.; Bologa, C.G.; Olah, M. In *Virtual Screening in Drug Discovery*; Alvarez, J.; Shoichet, B., Eds.; CRC Press Inc.: Boca Raton, **2004**, in press.

[15] Bologa, C.G.; Olah, M; Oprea, T.I. In *Bioinformatics in Drug Discovery*; Larson, R.S., Ed.; Humana Press: New York, **2005**, in press.

[16] Mestres, J. *Biochem. Soc. Trans.*, **2002**, *30*, 797.

[17] Oprea, T.I. *Molecules*, **2002**, *7*, 51.

[18] Oprea, T.I.; Li, J.; Muresan, S.; Mattes, K.C. In *EuroQSAR 2002 - Designing drugs and crop protectants: Processes, problems and solutions*; Ford, M.; Livingstone, D.; Dearden, J.; Van de Waterbeemd, H., Eds.; Blackwell Publishing: New York, **2003**, pp. 40-47.

[19] Oprea, T.I.; Bologa, C.G.; Edwards, B.S.; Prossnitz, E.A.; Sklar, L.A. *J. Biomol. Screen.*, **2004**, *9*, in press.

[20] Rishton, G. *Drug Discov. Today*, **2003**, *8*, 86.

[21] McGovern, S.L.; Caselli, E.; Grigorieff, N.; Shoichet, B.K. *J. Med. Chem.*, **2002**, *45*, 1712.

[22] Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E. M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjogren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harrism W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. *J. Med. Chem.*, **2002**, *45*, 137.

[23] Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C.D. *Chem. Rev.*, **2002**, *102*, 783.

[24] Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C.D. *C-QSAR database*. Available from the BioByte Corporation, Claremont, CA; http://www.biobyte.com/.

[25] Hansch, C.; Fujita, T. *J. Am. Chem. Soc.*, **1964**, *86*, 1616.

[26] Allu, T.K.; Oprea, T.I. *J. Chem. Inf. Model.*, **2004**, submitted.

[27] *iResearch™ Library*, ChemNavigator, Inc., San Diego, CA; http://www.chemnavigator.com/cnc/chemInfo/iRL.asp.

[28] See the Scitegic website, http://scitegic.com/.

[29] See the Chemical Computing Group website, http://www.chemcomp.com/.

[30] *FILTER 2.1*, OpenEye Scientific Software Inc., Santa Fe, N.M; http://www.eyesopen.com/products/applications/filter.html.

[31] Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T.I. In *Chemoinformatics in Drug Discovery*; Oprea, T.I., Ed.; Wiley-VCH: New York, **2004**, pp. 223-239.

[32] Coats, E.A. In *3D QSAR in Drug Design*; Kubinyi, H.; Folkers, G.; Martin, Y.C., Eds.; Kluwer/ESCOM: Dordrecht, **1998**, Vol. *3*, pp. 199-213.

[33] Oprea, T.I.; Olah, M.; Ostopovici, L.; Rad, R.; Mracec, M. In *EuroQSAR 2002 - Designing drugs and crop protectants: Processes, problems and solutions*; Ford, M.; Livingstone, D.; Dearden, J.; Van de Waterbeemd, H., Eds.; Blackwell Publishing: New York, **2003**, pp. 314-315.

[34] Kenny, P.W.; Sadowski, J. In *Chemoinformatics in Drug Discovery*; Oprea, T.I., Ed.; Wiley-VCH: New York, NY, **2004**, pp. 271-285.

[35] Available from Daylight Chemical Information Systems; http://www.daylight.com/.

[36] *FRED*, OpenEye Scientific Software Inc., Santa Fe, NM; http://www.eyesopen.com/products/applications/fred.html.

[37] Sadowski, J.; Gasteiger, J. *Chem. Rev.*, **1993**, *93*, 2567.

[38] Tripos, Inc. St. Louis, Missouri; http://www.tripos.com/.

[39] *CORINA* is available from Molecular Networks GmbH, Erlangen, Germany; http://www.mol-net.de/.

[40] See the Daylight Chemical Information Systems, Inc., Santa Fe, NM, website, http://www.daylight.com/.

[41] Accelrys Inc., San Diego, California; http://www.accelrys.com/.

[42] Optive Research Inc., Austin, TX, http://www.optive.com/; Optive products are also available from Tripos Inc., http://www.tripos.com/.

[43] *OMEGA*, OpenEye Scientific Software Inc., Santa Fe, NM; http://www.eyesopen.com/products/applications/fred.html.

[44] *LibraryExplorer*, Optive Research Inc.; http://www.optive.com/productDetailLibraryExplorer.do.

[45] Pearlman, R.S. Novel Software Tools for Addressing Chemical Diversity. Network Science; http://netsci.org/Science/Combichem/feature08.html.

[46] Pearlman, R.S.; Smith, K.M. *Perspect. Drug Discov.*, **1998**, *9-11*, 339.

[47] See the Mesa Analytics & Computing, Santa Fe, NM website, http://www.mesaac.com/.

[48] Taylor, R. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 59.

[49] Butina, D. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 747.

[50] Linusson, A.; Gottfries, J.; Lindgren, F.; Wold, S. *J. Med. Chem.*, **2000**, *43*, 1320.

[51] Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wikström, C.; Wold, S. *Design of Experiments: Principles and Applications*, Umetrics Academy: Umeå, **2000**.

[52] Johnson, M.A.; Maggiora, G.M. *Concepts and Applications of Molecular Similarity*, Wiley-VCH: New York, **1990**.

[53] Oprea, T.I. *Curr. Opin. Chem. Biol.*, **2002**, *6*, 384.

[54] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, **2002**.

[55] Tanimoto, T.T. *Trans. N.Y. Acad. Sci. Ser 2*, **1961**, *23*, 576.

[56] Tversky, A. *Psychol. Rev.*, **1977**, *84*, 327.

[57] Willett, P. *Similarity and Clustering Techniques in Chemical Information Systems*, Research Studies Press: Letchworth, **1987**.

[58] Willett, P. *Curr. Opin. Biotech.*, **2000**, *11*, 85.

[59] Lewis, R.A.; Pickett, S.D.; Clark, D.E. *Rev. Comput. Chem.*, **2000**, *16*, 1.

[60] Martin, Y.C. *J. Comb. Chem.*, **2001**, *3*, 231.

[61] Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273; see also MDL website http://www.mdli.com/.

[62] Rosati, R.L.; Da Silva Jardine, P.; Cameron, K.O.; Thompson, D.D.; Ke, H.Z.; Toler, S.M.; Brown, T.A.; Pan, L.C.; Ebbinghaus, C.F.; Reinhold, A.R.; Elliott, N.C.; Newhouse, B.N.; Tjoa, C.M.; Sweetnam, P.M.; Cole, M.J.; Arriola, M.W.; Gauthier, J.W.; Crawford, D.T.; Nickerson, D.F.; Pirie, C.M.; Qi, H.; Simmons, H.A.; Tkalcevic, G.T. *J. Med. Chem.*, **1998**, *41*, 2928.

[63] Scozzafava, A.; Supuran, C.T. *Bioorg. Med. Chem. Lett.*, **2002**, *12*, 1551.

[64] Grant, J.A.; Pickup, B.T.; Nicholls, A. *J. Comput. Chem.*, **2001**, *22*, 608.

[65] Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. *J. Med. Chem.*, **2000**, *43*, 3233.

[66] Goodford, P.J. *J. Med. Chem.*, **1985**, *28*, 849.

[67] Ewing, T.J.A.; Makino, S.; Skillman, A.G.; Kuntz, I.D. *J. Comput.-Aided Mol. Design*, **2001**, *15*, 411.

[68] Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. *J. Mol. Biol.*, **1997**, *267*, 727.

[69] Kramer, B.; Metz, G.; Rarey, M.; Lengauer, T. *Med. Chem. Res.*, **1999**, *9*, 463.

[70] Osterberg, F.; Morris, G.M.; Sanner, M.F.; Olson, A.J.; Goodsell, D.S. *Proteins: Struct., Funct., Genet.*, **2001**, *46*, 34.

[71] Sippl, M.J. *J. Comput.-Aided Mol. Design*, **1993**, *7*, 473.

[72] Sippl, M.J. *Curr. Opin. Struct. Biol.*, **1995**, *5*, 229.

[73] Domingues, F.S.; Koppensteiner, W.A.; Jaritz, M.; Prlic, A.; Weichenberger, C.; Wiederstein, M.; Floeckner, H.; Lackner, P.; Sippl, M.J. *Proteins: Struct., Funct., Genet.*, **1999**, *Suppl. 3*, 112.

[74] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. *Nucl. Acids Res.*, **2000**, *28*, 235.

[75] Protein Data Bank is accesible online at http://www.rcsb.org/.

[76] DeWitte, R.S.; Shakhnovich, E.I. *J. Am. Chem. Soc.*, **1996**, *118*, 11733.

[77] DeWitte, R.S.; Ishchenko, A.V.; Shakhnovich, E.I. *J. Am. Chem. Soc.*, **1997**, *119*, 4608.
[78] Ishchenko, A.V.; Shakhnovich, E.I. *J. Med. Chem.*, **2002**, *45*, 2770.
[79] Muegge, I.; Martin, Y.C. *J. Med. Chem.*, **1999**, *42*, 791.
[80] Muegge, I.; Martin, Y.C.; Hajduk, P.J.; Fesik, S.W. *J. Med. Chem.*, **1999**, *42*, 2498.
[81] Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.*, **2000**, *295*, 337.
[82] Sotriffer, C.A.; Gohlke, H.; Klebe, G. *J. Med. Chem.*, **2002**, *45*, 1967.
[83] Ajay; Murcko, M. *J. Med. Chem.*, **1995**, *38*, 4953.
[84] Williams, D.H.; Cox, J.P.L.; Doig, A.J.; Gardner, M.; Gerhard, U.; Kaye, P.T.; Lal, A.R.; Nicholls, I.A.; Salter, C.J.; Mitchell, R.C. *J. Am. Chem. Soc.*, **1991**, *113*, 7020.
[85] Williams, D.H.; Bardsley, B. *Perspect. Drug Discov. Design*, **1999**, *17*, 43.
[86] Rose, P.W. *Scoring methods in ligand design. Second UCSF Course in Computer-Aided Molecular Design*, San Francisco, **1997**.
[87] Marrone, T.J.; Luty, B.A.; Rose, P.W. *Perspect. Drug Discov. Design*, **2000**, *20*, 209.
[88] Mackay, J.P.; Gerhard, U.; Beauregard, D.A.; Maplestone, R.A.; Williams, D.H. *J. Am. Chem. Soc.*, **1994**, *116*, 4573.
[89] Böhm, H.J. *J. Comput. Aided Mol. Design*, **1994**, *8*, 243.
[90] Verkhivker, G.; Appelt, K.; Freer, S.T.; Villafranca, J.E. *Protein Eng.*, **1995**, *8*, 677.
[91] Bohm, H.J. *J. Comput.-Aided Mol. Design*, **1998**, *12*, 309.
[92] Head, R.D.; Smythe, M.L.; Oprea, T.I.; Waller, C.L.; Greene, S.M.; Marshall, G.R. *J. Am. Chem. Soc.*, **1996**, *118*, 3959.
[93] Marshall, G.R.; Head, R.D.; Ragno, R. In *Thermodynamics in Biology*; Di Cera, E., Ed.; Oxford University Press: New York, **2000**, pp. 87-111.
[94] Eldridge, M.D.; Murray, C.W.; Auton, T.R.; Paolini, G.V.; Mee, R.P. *J. Comput.-Aided Mol. Design*, **1997**, *11*, 425.
[95] Murray, C.W.; Auton, T.R.; Eldridge, M.D. *J. Comput.-Aided Mol. Design*, **1998**, *12*, 503.
[96] Jain, A. *J. Comput.-Aided Mol. Design*, **1996**, *10*, 427.
[97] Horvath, D. *J. Med. Chem.*, **1997**, *40*, 2412.
[98] Wang, R.; Liu, L.; Lai, L.; Tang, Y. *J. Mol. Model.*, **1998**, *4*, 379.
[99] Wang, R.; Lai, L.; Wang, S. *J. Comput.-Aided Mol. Design*, **2002**, *16*, 11.
[100] Grant, J.A.; Pickup, B.T.; Nicholls, A. *J. Comput. Chem.*, **2001**, *22*, 608.
[101] Wang, R.; Wang, S. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1422.
[102] Wu, J.H.; Batist, G.; Zamir, L.O. *Anti-Cancer Drug Design*, **2001**, *16*, 129.
[103] Available from NCI; http://resresources.nci.nih.gov/database.cfm?id=501.
[104] Rarey, M.; Lemmen, C.; Matter, H. In *Chemoinformatics in Drug Discovery*; Oprea, T.I., Ed.; Wiley-VCH: New York, **2004**, pp. 51-108.
[105] Oprea, T.I.; Matter, H. *Curr. Opin. Chem. Biol.*, **2004**, *8*, 349.
[106] Young, S.S.; Farmen, M.; Rusinko III, A. *Random Versus Rational: Which is Better for General Compound Screening?* Available on-line at http://www.netsci.org/Science/Screening/feature09.html.
[107] Kubinyi, H. *J. Rec. Signal Transd. Res.*, **1999**, *19*, 15.
[108] See the OpenEye Scientific Software, Santa Fe, NM website, http://www.eyesopen.com/.
[109] Weininger, D. *J. Chem. Inf. Comput. Sci.*, **1988**, *28*, 31.
[110] *Daylight Toolkit v4.81*, Daylight Chemical Information Systems, Santa Fe, NM; http://www.daylight.com/.
[111] *OEChem Toolkit v1.3*, Openeye Scientific Software, Santa Fe, NM; http://www.eyesopen.com/.
[112] *Open Babel*; http://openbabel.sourceforge.net/.
[113] *Smi2fp_ascii*, Daylight Chemical Information Systems, Santa Fe, NM; http://www.daylight.com/.
[114] *UNITY® 4.4.2*, Tripos Inc., St. Louis, MO; http://www.tripos.com/.
[115] *MACCSKeys320Generator*, Mesa Analytics and Computing LLC, Santa Fe, NM; http://www.mesaac.com/.
[116] Barnard, J.M.; Downs, G.M. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 141; see also http://www.bci.gb.com/clusteranalysis.html.
[117] MOE: The Molecular Operating Environment from Chemical Computing Group Inc., Montreal, Quebec; http://www.chemcomp.com/.
[118] *CLOGP 4.0*, BioByte Corporation, Claremont, CA; http://www.biobyte.com/.
[119] ACD Labs is available from Advanced Chemistry Development, Inc.; http:// www.acdlabs.com/.
[120] *EPI Suite v3.11*, U.S. Environmental Protection Agency; http://www.epa.gov/.
[121] Tetko, I.V.; Tanchuk, V.Y. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1136; http://146.107.217.178/lab/alogps/index.html.
[122] *Cluster Package*, Daylight Chemical Information Systems, Santa Fe, NM; http://www.daylight.com/.
[123] *Measures*, Mesa Analytics and Computing LLC, Santa Fe, NM; http://www.mesaac.com/.
[124] ChemoMine plc, Cambridge UK; http://www.chemomine.co.uk/.
[125] *MODDE 7*, Umetrics AB, Umeå, Sweden; http://www.umetrics.com/.