High-throughput and virtual screening: core lead discovery technologies move towards integration

Andrew C. Good, Stanley R. Krystek and Jonathan S. Mason

In addition to high-throughput screening (HTS), the main lead discovery technology employed by most pharmaceutical companies today is virtual screening (VS). Although the two techniques have somewhat different philosophical origins, they contain many synergies that can potentially enhance the lead discovery process. Here, we describe many of the latest developments in VS technology with particular emphasis on their potential impact on HTS in, for example, focussed screening and data mining. In addition, we highlight key issues that need to be addressed before the potential of such efforts can be fully realized.

***Andrew C. Good**
Bristol-Myers Squibb
5 Research Parkway
PO Box 5100, Wallingford
CT 06492, USA
*tel: +1 203 677 6761
fax: +1 203 677 7702
e-mail: andrew.good@bms.com
Stanley R. Krystek and Jonathan S. Mason
Bristol-Myers Squibb
PO Box 4000, Princeton
NJ 08543, USA

▼ The techniques that encompass VS have provided key technologies within the pharmaceutical industry for some time. The primary VS premise is to screen a database of molecules computationally using structural descriptors that relate in some way to potential biological activity. A subset of database molecules found to match these descriptors can then be selected for subsequent biological analysis. Many methodologies are employed, from 2-D substructure- and topological-similarity searches using active ligand structures, to analyses within the 3-D constraints of a target active site. Although VS technology has been extensively reviewed[1–4], relevant illustrations of application and utility are still warranted.

Pharmacophore-constrained searches

In terms of novel lead discovery, pharmacophore searching has perhaps proven the most widely applied VS method. A pharmacophore in this context is generally defined as a critical arrangement of molecular fragments or features creating a necessary, but not sufficient, condition for biological activity[5]. Such functionality is usually atom-based and defined in terms of generic chemical properties (e.g. acid, base, hydrophobe or aromatic), although other properties (e.g. planes, normals and potential target-atom positions) are also used. The method provides an excellent paradigm for discovering novel active chemotypes based on potential ligand binding modes, with hit rates for selected data sets from 1–20% (depending on the quality of the pharmacophore and care taken in compound filtering)[6–9]. Figure 1 illustrates a successful application of this technique.

Pharmacophore application has also been extended to encompass another important VS technique – structure-based VS. It is possible to discover ligands with both diverse chemotypes and binding modes by exploiting structural information taken directly from the target active site. As a result, structure-based VS is potentially the most powerful form of VS (Refs 10–12). There are, however, significant challenges facing structure-based searching tools, because the thousands of putative ligand–receptor orientations that must be tested make for an extremely computationally intensive process. In the past, it was not uncommon to spend one or two weeks searching a database of 100,000 structures (single-ligand conformer only). By combining pharmacophore constraints into the search model[13–15], particularly in conjunction with crucial active-site regions[16] such as a crucial salt bridge interaction, the resulting VS calculation becomes

1359-6446/00/$ – see front matter PII: S1471-1931(00)00015-X

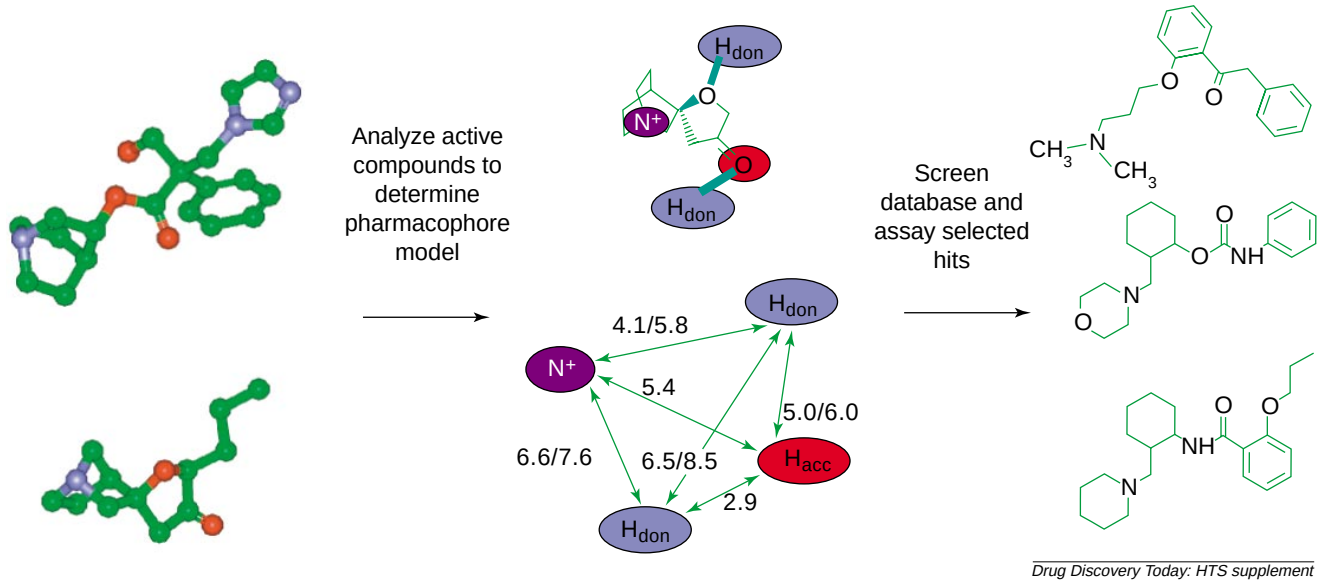

Figure 1. Muscarinic M_3-receptor antagonist virtual screening (VS) study undertaken by Marriott *et al.*[9] The pharmacophore shown in the center of the diagram was derived from known M_3-receptor antagonists shown on the left. Three of the hits found in the resulting VS calculation are shown on the right. H_{don} represents potential active-site donor-atom position, ligand acceptor is designated by H_{acc} and N^+ refers to ligand base. Adapted from Ref. 9.

a series of active-site-constrained pharmacophore searches. Using such an approach dramatically decreases the number of non-productive (low chemical complementarity) ligand–receptor orientations, producing increases in search speed (up to two orders of magnitude) and better hit rates[4] (A.C. Good, pers. commun.). New variants of such technology continue to be developed at a rapid rate, from novel search methods[17] to alternative scoring-function application[18]. By combining such technology with other new developments, for example new scoring functions and docking methods[19,20], the full potential of structure-based VS can be unleashed.

New descriptors for lead discovery

The development of VS technology continues in answer to the questions raised by new technologies adopted within the pharmaceutical industry. In general, the techniques described here require either a postulated crucial pharmacophore for binding or a target active site. Often this amount of structural information is not available however, with only a single lead (e.g. competitor ligand or peptide substrate) accessible. With technologies such as genomics providing an ever-increasing queue of potential targets with limited biological information, this scenario is likely to become an increasingly common problem. Furthermore, the data-set sizes that must be analysed for HTS and combinatorial chemistry libraries, are often beyond the more popular 3-D VS search techniques that require extensive molecular superpositions[17,21]. As a result, alternative technologies have been developed that employ topological and/or pharmacophore information to create molecular descriptors in the form of binary fingerprints[22–26] and histograms[27]. Figure 2 summarizes how several of these descriptors are derived. These techniques permit rapid screening of large data sets (100,000 compounds can be screened in minutes) and have been shown to cluster and extract active compounds from databases using the fingerprint of a single lead[23–27].

Searching versus the RGD motif of fibrinogen

Pickett *et al.*[23] employed a 100,000 compound database, for which four-point-pharmacophore fingerprints had been calculated to study their utility. The fingerprint of the RGD motif that fibrinogen uses to bind to its receptor[28] was then applied to screen the database, which was seeded with 12 fibrinogen-receptor antagonists covering a variety of structural classes[29] (Fig. 3). The RGD motif provides an interesting test case, because it is known to be in a disordered loop region of the fibronectin Type-III domain [as identified by solution nuclear magnetic resonance (NMR) and crystal structures[30]], suggesting a degree of flexibility. It would thus be difficult to identify a reasonable single pharmacophore from knowledge of the crystal structure alone. An analysis of the resulting hit list found all the actives were in the top 3% of the data set of 100,000 compounds, illustrating the ability of pharmacophore descriptors to extract novel active chemotypes using the minimum of structural information. This highlights a useful feature of the conformational-ensemble descriptor, because when the conformer fingerprints of a particular molecule are combined, the resultant descriptor provides an

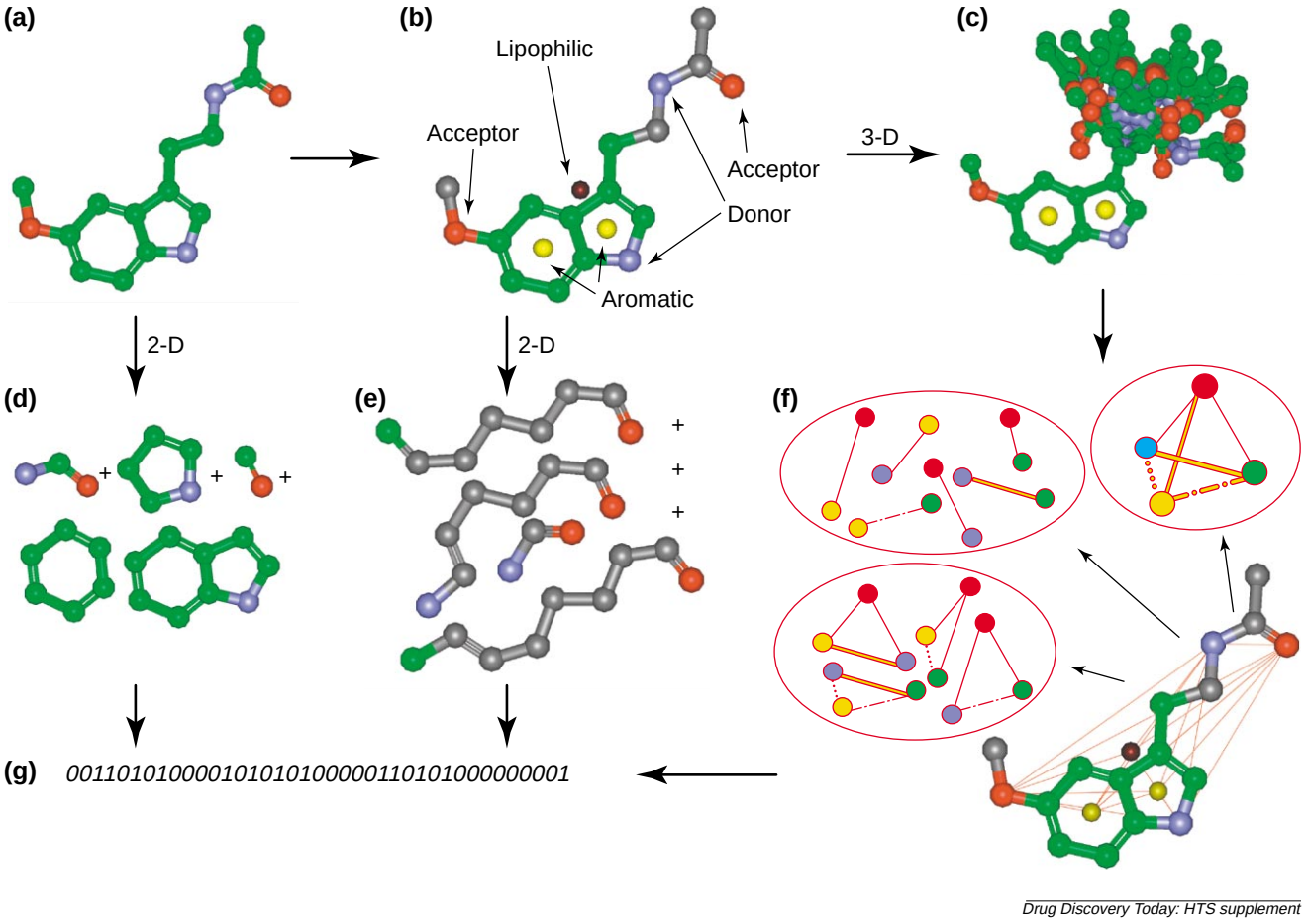

Figure 2. Schematic illustration of primary methods used in molecular fingerprint creation. **(a)** Create 2-D and 3-D model of molecule; **(b)** deconstruct the molecule into pharmacophoric elements; **(c)** generate conformational models; **(d)** deconstruct the molecule into topological/substructural elements; **(e)** determine distance between pharmacophoric groups using bond counts; **(f)** determine 2-, 3- or 4-center distance combinations of pharmacophoric groups for each conformer; and **(g)** determine the presence or absence of each descriptor element and combine to create a binary fingerprint.

overall measure of the pharmacophores available to that structure. Such a property is of particular value when the bioactive conformation is not known (as in this example), as no pharmacophore model is required to undertake the search.

ADMET studies

More recently, the idea of extending VS technology to the problem of adsorption, distribution, metabolism, excretion and toxicity (ADMET) has begun to be studied. The molecular properties of logP and molecular weight have been used for some time as descriptors for filtering compound selections[31]. Recently, this methodology was extended and placed into the context of the known drug universe by Lipinski and co-workers[32], who highlighted the shift away from 'drug-like' compounds with the increased use of combinatorial chemistry and HTS. Additional tools continue to be developed – a good example being the application of polar surface area calculations for the prediction of blood–brain-barrier penetration[33]. Such tools lend themselves to rapid calculation and thus provide a convenient screening methodology for compound filtering. More sophisticated techniques are also being advanced, for example through extensive comparison of large drug-like [World Drug Index (WDI; developed and distributed by Derwent Publications, London, UK)] and non-drug-like [Available Chemicals Database (ACD; developed and distributed by MDL Information Systems, San Leandro, CA, USA)] databases with neural networks and multiple structural descriptors[34]. The ultimate aim of such studies is to improve our understanding of the number and importance of molecular descriptors that determine the drug-like nature of a molecule. For an up-to-date review of this area of VS, see the work of Blake[35].

Synergies between VS and HTS technology

There are many areas in the HTS process where suitable application of VS technology (and vice versa) can impact favorably

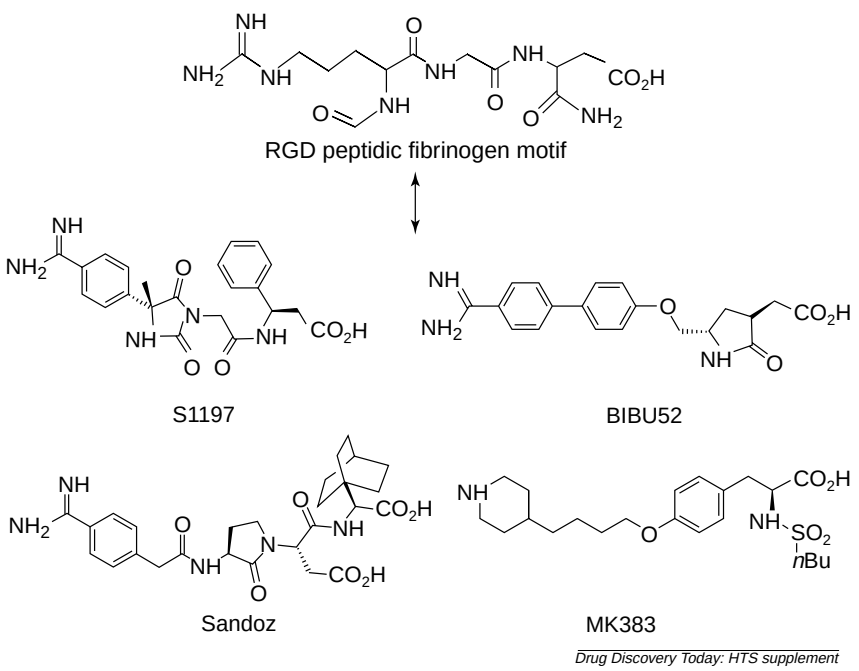

Figure 3. RGD peptidic fibrinogen motif and associated small-molecule inhibitors illustrating the major structural differences between the peptide-binding motif and corresponding small-molecule inhibitors. Adapted from Ref. 23.

on lead discovery. In the following section we summarize some of these synergies.

Augmenting compound collections

One of the key requirements for HTS is a large and diverse source of compounds. Smaller pharmaceutical companies tend to have limited compound inventories, and although many of the major pharmaceutical firms have access to larger molecular collections, these often contain inherent historical bias with many compounds of venerable age and questionable purity. Furthermore, there is no obvious route for the re-supply of many such historical molecules. As a consequence, augmentation of in-house compound databases continues to be a high priority throughout the industry. The two main methods for accomplishing this are in-house combinatorial chemistry and outsourced compound acquisition. In both cases VS techniques are widely applied to aid in compound selection. VS descriptors are often used to determine the structural diversity of molecules purchased both within the compound selection, and with respect to the current HTS inventory[36,37]. In addition, these descriptors have been combined with those used in ADMET screening to ensure that selections are not only diverse but also drug-like[36,38]. It is also possible to extend the same technology to the analysis of external supplier databases simply by replacing the compound libraries with supplier lists[37]. Such techniques have been reviewed elsewhere[39].

Compound mixing and pooling

In some HTS departments there is still a demand for mixing compounds into pools for faster screening. A side-effect of this is that many compounds have a tendency to cross-react when mixed together. Hann *et al.*[40] used a number of VS filters and reactivity classifications to select molecules with a low occurrence of highly reactive groups, and pooled them to prevent mixing of molecules prone to cross-reactivity (e.g. nucleophile and electrophile). LC–MS analysis was undertaken on pools derived from mixtures of 160 discrete in-house compounds at time zero, and one and two months after mixing. Pools mixed after sorting by reactive class were found to contain significantly less unexplained new peaks (9% after three months), compared with pools combined using other less 'rational' methods (25–78% after three months).

Similar techniques have also been designed to minimize the occurrence of mass duplicates, therefore improving the ability of mass spectroscopy to identify the active component in a mixture[41].

Focussed, iterative and HTS VS

As we have already alluded to, compound re-supply is a major HTS issue. However, this problem can be mitigated through more efficient use of the existing inventory of a company. One method by which this can be accomplished is through the creation of focussed screening sets. Such sets are generally directed towards target classes for which structure–activity relationships (SARs) exist across the different lead chemotypes or active sites (e.g. kinases, ion channels). Using VS techniques to analyse, for example, known crystal structures and lead inhibitors, allows the creation of compound subsets biased towards the target class SAR. These subsets can then be used as an initial screen in the hope of generating sufficient hits without being forced to screen the whole molecular inventory. Alternatively, the SAR generated from the initial screen can be used to make further compound selection, thus forming an iterative screening strategy.

In a similar vein, HTS robotic-compound-selection technology has the potential to increase the number of hits found in a given VS calculation. Referring back to the RGD VS calculation example of Pickett *et al.*[23], all of the active compounds were found in the top 3% of the VS hit list. For an inventory of 500,000 compounds, this still translates into 15,000 molecules.

Traditionally, one limit of VS in many institutions has been that its application has been restricted to a low-throughput mode (100–1000 compounds). However, HTS robotics, when correctly configured, provides the technology necessary for routine 'cherry picking' of larger screening sets. As a consequence, it is technically possible to select thousands of compounds routinely from an inventory. This allows the creation of real-time focussed screening sets, providing another strategy for iterative screening while simultaneously increasing the potential impact of VS calculations.

HTS data mining

An area of research currently generating a lot of interest is the field of data mining. HTS is capable of generating a huge amount of SAR data. The ability to create SAR models from this data has the potential to provide initial direction in lead optimization efforts, while forming an additional iterative screening protocol. VS descriptors used in combination with existing computer-aided-drug-design (CADD) technology offer useful tools to address the data-mining challenge. In addition to existing CADD methodology, new techniques are being developed to cope with the large amount of data generated and analysed during such calculations[25,42,43] (Calvet, A. and Cowan, G. Validating new techniques for HTS data analysis. Data storage, visualization and mining solutions, *IBC Screentech 2000*, Monterey, CA, USA, 28 February–1 March 2000).

Exploitation of existing CADD technology

A good example of CADD technology derives from the work of McGregor and Muskal[24], who developed a strategy based on three-center-pharmacophore fingerprints for their studies. The six core pharmacophoric types (donor, acceptor, acid, base, aromatic and hydrophobe) were applied in the fingerprint, together with an additional definition of 'other' for all remaining unassigned atoms. The distances between centers in each pharmacophore were divided into six ranges, leaving 10,549 theoretically accessible pharmacophores after removal of those failing the triangle rule (the length of one side cannot exceed the combined length of the other two) or those made redundant by symmetry.

The resultant descriptors were analysed in conjunction with three estrogen-receptor (ER) data sets previously analysed with other quantitative-SAR (QSAR) descriptors[44]. One of the studies mimicked a data-mining analysis in which a partial least squares (PLS) QSAR model was derived from 15 actives (activity set to 1.0), plus 750 'inactives' taken from the non-ER-active structures in the MDL Drug Data Report (MDDR, developed and distributed by MDL Information Systems) (activity set to 0.0). This test was designed to simulate the kind of 'noisy' data seen in HTS screens. Model validation was undertaken by scoring 250 MDDR-ER active molecules, 86 ER actives from a combinatorial library and 8290 inactives from the remainder of the MDDR excluded from model training. Using a cut-off value of 0.2 (determined by initial training set molecule–activity distributions within the model) to define the boundary between active and inactive, >87% of each test set was assigned correctly in all three cases. Figure 4 highlights the highest positively weighted pharmacophore for the QSAR model in the context of two diverse ER-active chemotypes.

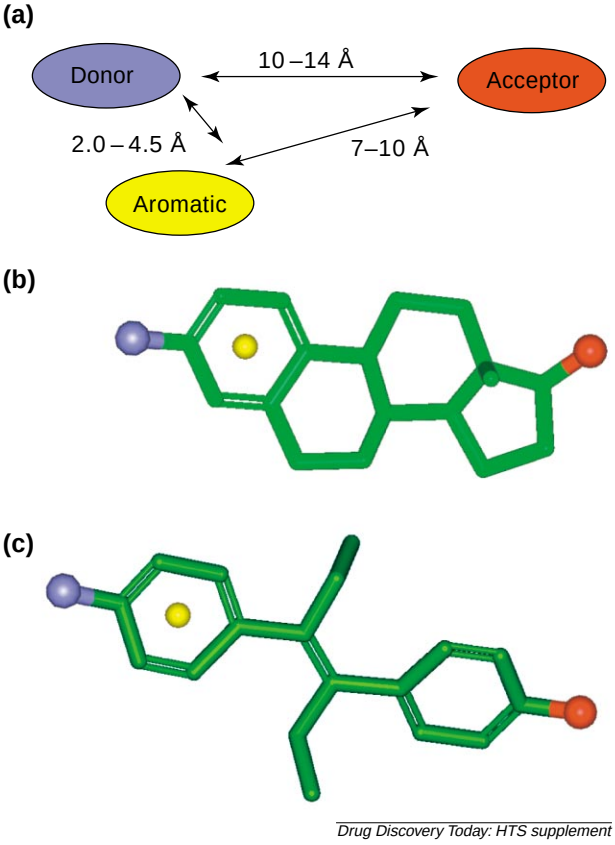

Figure 4. Pharmacophore **(a)** with the greatest positive weighting from estrogen-receptor studies by McGregor and Muskal[24], mapped to the natural ligand estradiol **(b)** and the most potent data set compound diethylstilbestrol **(c)**. Adapted from Ref. 24.

Development of new methodology

Rusinko *et al.*[25] extended the statistical-classification technique of recursive partitioning[45] (RP) to analyse large biological data sets. This method has the advantage that it scales linearly with the number of descriptors analysed. As a consequence, the technology is extremely fast; it is possible to analyse a data set of 300,000 compounds associated with 2 million descriptors in ~1 h of CPU time on a typical workstation.

For this implementation, the recursive partitioning technique SCAM was constructed to analyse a variety of binary fingerprint descriptors. The technique works by building a classification

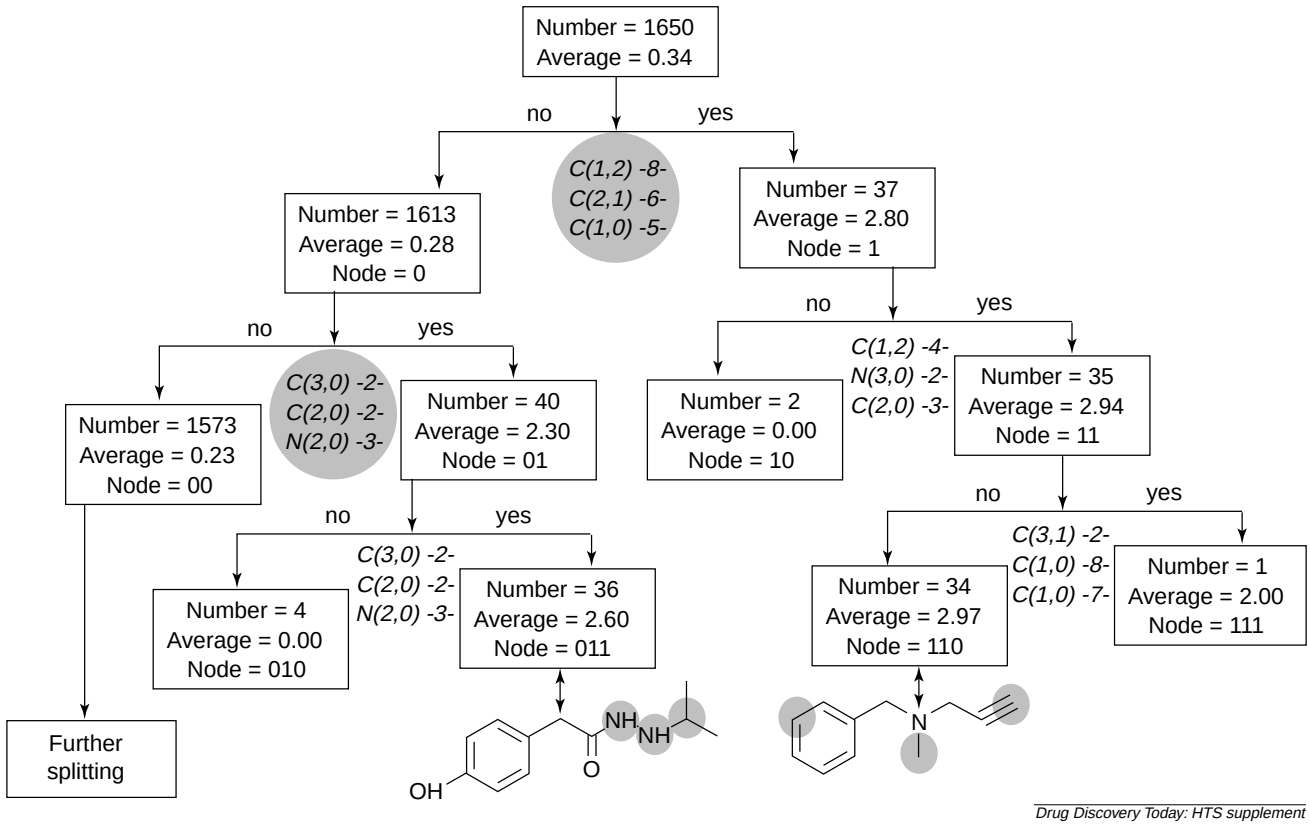

Figure 5. Results of SCAM run on MAO inhibitors using atom triple descriptors. Node split atom triple information and selected associated statistics are included. Two active molecules associated with specific nodes are included in the display, and selected atom triples associated with these nodes are highlighted. Definitions: number, number of compounds in split; average, average activity; and node, node number. Adapted from Ref. 25.

tree. At each branch of the tree the program selects descriptors whose presence or absence permits differentiation between inactive and active molecules. These descriptors and the resultant splitting of actives and inactives, are described at each tree node, permitting easier results analysis.

An example of this technique in action using atom triples (similar to the three-center pharmacophores described above, but with topological bond count rather than Cartesian space used to measure the distance between centers) is shown in Fig. 5. For this study, 1650 monoamine oxidase (MAO) inhibitors were studied[46]. Various molecular descriptors were analysed and VS calculations run on the resulting classification tree to select further compounds for screening. In one such calculation, the WDI (35,631 structures) was searched for MAO inhibitors [72 known to exist in the database (0.2% of total)]. Using the RP tree classification rules derived from the training set, 227 structures were identified as having potential MAO activity. Of these, seven were found to be classified as such (a 3% 'hit' rate), representing a 15-fold enrichment over random selection.

Target prioritization

All the methods described so far refer to VS approaches for lead discovery. Recent bioinformatics developments in the discovery and validation of new protein targets also has the potential to aid HTS. Although not VS in the traditional sense, they are nonetheless 'computational screening' from the target perspective, and are thus worthy of mention. The distribution of current drug-target biochemical classes[47] (Box 1) illustrates that current therapies are focussed within certain groups of proteins. Furthermore, it is clear that specific target classes [e.g. enzymes, G-protein-coupled receptors and nuclear hormone receptors (NHRs)] lend themselves more easily to drug discovery than others (e.g. protein–protein interactions). One of the major genome-project hurdles for pharmaceutical companies is to determine which genes will provide the best drug targets. This is of crucial importance, because it is not possible for HTS to screen all proteins deemed potentially important. Rather, companies must carefully select the targets most likely to yield a novel and useful therapy. This is where bioinformatics techniques step in, using a variety of protein-sequence, -motif or -fold

recognition comparison technologies. The goal of such methods is to mine genomic database sequence information for structural homologs, thus allowing target class assignment, and hence target prioritization.

Many techniques have been developed for this problem, including multiple sequence alignments[48,49], motif searching to identify functionally relevant regions[50] and profile-based methods such as psi-BLAST (Ref. 51). In addition, when similarity searching or motif analyses fail to assign function, recent advances in 'protein threading' (fold recognition) techniques can nevertheless make it possible to predict 3-D structures from novel protein sequences[52].

Antimicrobial targets provide a good example of technique application. There are currently 30 complete bacterial and one complete fungal genome published[53]. In addition, >100 bacterial, >10 fungal, and eight protozoal genomes are currently being sequenced. Genomic selection is being used on the microbial genomes as a strategy for the identification of potential antibacterial targets. As well as being essential to cell survival, antimicrobial targets are generally proteins found in multiple pathogenic genomes and not in eukaryotic organisms. Bioinformatics tools are being applied in the search for targets with these selectivity profiles using techniques such as concordance analysis[54].

Conclusions: remaining challenges to integration

Clearly, bioinformatics target-validation tools already have the ability to impact new project prioritization and hence help direct the resource application of HTS departments. Furthermore, the examples already detailed clearly demonstrate the potential complementary nature of HTS and traditional VS techniques. This potential was recognized early on in the history of HTS development[55], and efforts to create and deploy relevant enabling technologies are ongoing[56]. In addition, the emerging importance of techniques such as data mining are such that alliances between pharmaceutical and software companies continue to be spawned (Box 2). For such synergies to be applied to full effect however, a number of issues must be addressed. For example, both data-mining methods described in this article attempt to simulate HTS data using molecules whose biological activities have been more clearly defined (for example from the MDDR or confirmed IC_{50} data). Although some noise is introduced through the creation of binary activity profiles (0 inactive and 1 active), the overall accuracy of the data, and hence the SAR pattern inherent within it, is not in question. At the core of these, and all other 'flavors' of data mining, is some form of pattern-recognition technology. Consider this in the context of the two hypothetical assays detailed in Fig. 6. The assay with high reproducibility shows a distinct inhibition pattern across the duplicate runs and should lend itself well to data mining.

Box 1. Distribution of drug target biochemical classes for current therapies

G-protein-coupled receptors	45%
Enzymes	28%
Hormones and growth factors	11%
Ion channels	5%
Nuclear receptors	2%
Other	9%

Box 2. URLs of press releases for pharmaceutical–software alliances

http://www.tripos.com/about/press/2000/2000.01.11.html
http://www.tripos.com/about/press/1999/1999.08.26.html

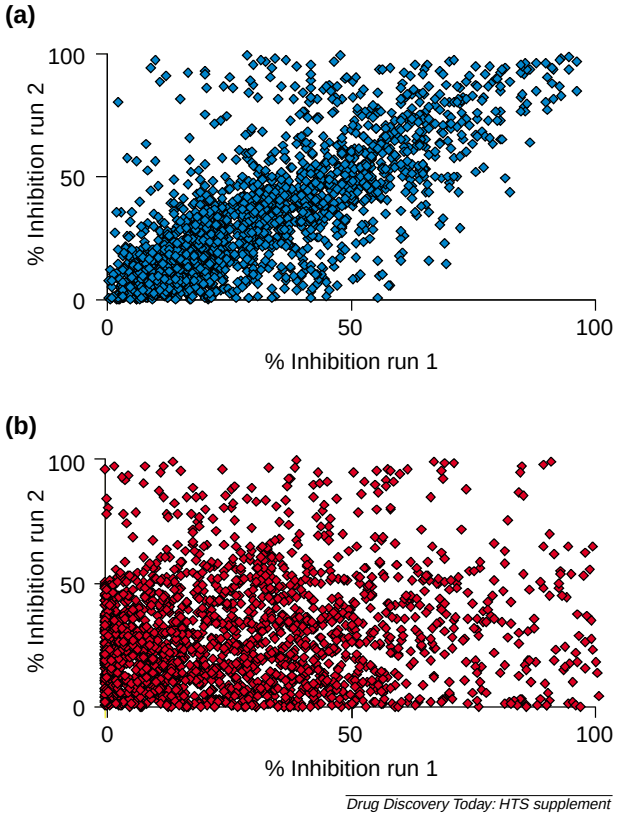

Figure 6. Two hypothetical HTS assays undertaken in duplicate illustrating potential SAR patterns (or the lack of them). **(a)** depicts an assay with high reproducibility and **(b)** depicts an assay with low reproducibility.

By contrast, the assay with low reproducibility shows no such pattern and mining this data would almost certainly prove futile. Clearly, focussed and iterative screening strategies are equally reliant on such SAR patterns for their successful application. The quality and precision of the raw HTS measurements is thus crucial if such patterns are to exist (Calvet, A. and Cowan, G. Validating new techniques for HTS data analysis. Data storage, visualization and mining solutions, *IBC Screentech 2000*, Monterey, CA, USA, 28 February–1 March 2000). In addition, focussed, iterative and HTS VS strategies also require a careful design of compound handling systems to enable routine 'cherry picking' from the inventory. There are many factors that can impact HTS measurement quality (Gunter, B. Quality control of the HTS process. *Innovative techniques for lead discovery and development*, 2000 Charleston Conferences, Isle of Palms, SC, USA, 28 February–1 March 2000), and although these factors are gaining some attention, the primary focus for many engaged in HTS development continues to be on ever-increasing capacity through miniaturization and increased compound inventory[57]. HTS–VS synergies will only be fully exploited when issues of quality and design are given at least equal status to those of quantity.

References

1 Milne, G.W.A. *et al*. (1998) Pharmacophores in drug design and discovery. *SAR QSAR Environ. Res*. 9, 23–38

2 Warr, W.A. and Willett, P. (1998) The principles and practice of three-dimensional database searching. In *Designing Bioactive Molecules* (Martin, Y.C. and Willet, P., eds), pp. 73–95, American Chemical Society

3 Walters, W.P. *et al.* (1998) Virtual screening – an overview. *Drug Discov. Today* 3, 160–178

4 Böhm, H-J. and Schneider, G., eds *Virtual Screening for Bioactive Molecules* Wiley–VCH (in press)

5 Marshall, G.R. (1993) Binding site modeling of unknown receptors. In *3D QSAR in Drug Design* (Kubinyi, H., ed.), pp. 80–116, Escom

6 Wang, S. *et al.* (1996) Discovery of novel, non-peptide HIV-1 protease inhibitors by pharmacophore searching. *J. Med. Chem.* 39, 2047–2054

7 Hong, H. *et al.* (1997) Discovery of HIV-1 integrase inhibitors by pharmacophore searching. *J. Med. Chem.* 40, 930–936

8 Kaminski, J. *et al.* (1997) Identification of novel farnesyl protein transferase inhibitors using three-dimensional database searching methods. *J. Med. Chem.* 40, 4103–4112

9 Marriott, D.P. *et al.* (1999) Lead generation using pharmacophore mapping and three-dimensional database searching: application to muscarinic M_3 receptor antagonists. *J. Med. Chem.* 42, 3210–3216

10 Murcko, M.A. *et al.* (1999) Structure-based drug design. *Annu. Rep. Med. Chem.* 34, 297–306

11 Sarmiento, M. *et al.* (2000) Structure-based discovery of small molecule inhibitors targeted to protein tyrosine phosphatase. *J. Med. Chem.* 43, 146–155

12 Burkhard, P. *et al.* (1999) The discovery of steroids and other novel FKBP inhibitors using a molecular docking program. *J. Mol. Biol.* 287, 853–858

13 Miller, M.D. *et al.* (1994) FLOG: A system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Design* 8, 153–174

14 Burkhard, P. *et al.* (1998) An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 Å. X-ray structure of a thrombin-ligand complex. *J. Mol. Biol.* 277, 449–466

15 Ewing, T.J.A. and Kuntz, I.D. (1997) Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* 18, 1175–1189

16 Desjarlais, R.L. and Dixon, J.S. (1994) A shape- and chemistry-based docking method and its use in the design of HIV-1 protease inhibitors. *J. Comput.-Aided Mol. Design* 8, 231–242

17 Miller, M.D. *et al.* (1999) SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions. *J. Med. Chem.* 42, 1505–1514

18 Charifson, P.S. *et al.* (1999) Consensus scoring; a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* 42, 5100–5109

19 Muegge, I. *et al.* (1999) Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* 42, 2498–2503

20 Baxter, C.A. *et al.* (2000) New approach to molecular docking and its application to virtual screening of chemical databases. *J. Chem. Inf. Comput. Sci.* 40, 254–262

21 Perry, N.C. and Van Geerestein, V.J. (1993) Database searching on the basis of three-dimensional molecular similarity using the SPERM program. *J. Chem. Inf. Comput. Sci.* 32, 607–616

22 Good, A.C. and Kuntz, I.D. (1995) Investigating the extension of pair-wise distance pharmacophore measures to triplet-based descriptors. *J. Comput.-Aided Mol. Design* 9, 373–379

23 Pickett, S.D. *et al.* (2000) Enhancing the hit-to-lead properties of lead optimization libraries. *J. Chem. Inf. Comput. Sci.* 40, 263–272

24 McGregor, M.J. and Muskal, S.M. (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* 39, 569–574

25 Rusinko, A., III *et al.* (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* 39, 1017–1026

26 Mason, J.S. *et al.* (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* 42, 3251–3264

27 Sheridan, R.P. *et al.* (1996) Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* 36, 128–136

28 Eldred, C.D. and Judkins, B.D. (1999) Fibrinogen receptor antagonists: design and clinical applications. *Prog. Med. Chem.* 36, 29–90

29 Mousa, S.A. and Cheresh, D.A. (1997) Recent advances in cell adhesion molecules and extracellular matrix proteins: potential clinical implications. *Drug Discov.Today* 2, 187–199

30 Leahy, D.J. *et al.* (1996) 2.0 Å crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell* 84, 155–164

31 Livingstone, D.J. (2000) The characterization of chemical structures using molecular properties. A Survey. *J. Chem. Inf. Comput. Sci.* 40, 195–209

32 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

33 Clark, D.E. (1999) Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration. *J. Pharm. Sci.* 88, 815–821

34 Wagener, M. and Geerestein, V.J. (2000) Potential drugs and non-drugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* 40, 280–292

35 Blake, J.F. (2000) Chemoinformatics – predicting the physicochemical properties of 'drug-like' molecules. *Curr. Opin. Biotechnol.* 1, 104–107

36 Good, A.C. and Lewis, R.A. (1997) New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPick. *J. Med. Chem.* 40, 3926–3936

37 Rhodes, N. and Willett, P. (2000) Bit-string methods for selective compound acquisition. *J. Chem. Inf. Comput. Sci.* 40, 210–214

38 Martin, E.J. and Critchlow, R.E. (1999) Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.* 1, 32–45

39 Willett, P. (2000) Chemoinformatics – similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* 11, 85–88

40 Hann, M. *et al.* (1999) Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* 39, 897–902

41 Martin, E. and Wong, A. (2000) Sensitivity analysis and other improvements to tailored combinatorial library design. *J. Chem. Inf. Comput. Sci.* 40, 215–220

42 Ahlberg, C. (1999) Visual exploration of HTS databases. *Drug Discov.Today* 4, 370–376

43 Gund, P. and Sigal, N.H. (1999) Applying informatics systems to high-throughout screening and analysis. In *Pharmainformatics: A Trends Guide* (Owen, M., ed.), pp. 25–29, Elsevier Science

44 Tong, W. *et al.* (1998) Evaluation of quantitative structure–activity relationship methods for large-scale prediction of chemicals binding to estrogen receptor. *J. Chem. Inf. Comput. Sci.* 38, 669–677

45 Quinlan, J.R. (1992) *C4.5 Programs for Machine Learning*, Morgan Kaufmann

46 Brown, R.D. and Martin, Y.C. (1996) Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 36, 572–584

47 Drews, J. (1999) Drug discovery: a historical perspective. *Science* 287, 1960–1964

48 Fasman, K.H. and Salzberg, S.L. (1998) An introduction to biological sequence analysis. In *Computational Methods in Molecular Biology* (Salzberg, S.L. *et al.*, eds), pp. 29–42, Elsevier Science

49 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

50 Baxevanis, A.D. (1998) Practical aspects of multiple sequence alignment. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (Baxevanis, A.D., ed.), pp. 172–188, John Wiley & Sons

51 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

52 Sippl, M.J. (1993) Boltzmann's principle, knowledge based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput.-Aided Mol. Design* 7, 473–501

53 Kyrpides, N. (1999) Genomes online database (GOLD): a monitor of complete and ongoing genome projects world wide. *Bioinformatics* 15, 773–774

54 Bruccoleri, R.E. *et al.* (1998) Concordance analysis of microbial genomes. *Nucleic Acids Res.* 26, 4482–4486

55 Hyde, R.M. and Hersey, A. (1996) Screening and rational drug design – competitive or complementary? In *Bioactive Compound Design* (Ford, M.G., ed.), pp. 37–45, Bios Scientific

56 Hann, M. and Green, R. (1999) Chemoinformatics – a new name for an old problem? *Curr. Opin. Chem. Biol.* 3, 379–383

57 Fox, S. *et al.* (1999) High throughput screening for drug discovery: continually transitioning into new technology. *J. Biomol. Screening* 4, 183–186