J Mol Model (2008) 14:135–148
DOI 10.1007/s00894-007-0257-9

ORIGINAL PAPER

Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment

Alexandre Beautrait · Vincent Leroux · Matthieu Chavent · Léo Ghemtio · Marie-Dominique Devignes · Malika Smaïl-Tabbone · Wensheng Cai · Xuegang Shao · Gilles Moreau · Peter Bladon · Jianhua Yao · Bernard Maigret

Received: 31 August 2007 / Accepted: 15 November 2007 / Published online: 3 January 2008

Abstract Numerous methods are available for use as part of a virtual screening strategy but, as yet, no single method is able to guarantee both a level of confidence comparable to experimental screening and a level of computing efficiency that could drastically cut the costs of early phase drug discovery campaigns. Here, we present VSM-G (virtual screening manager for computational grids), a virtual screening platform that combines several structure-based drug design tools. VSM-G aims to be as user-friendly as possible while retaining enough flexibility to accommodate other in silico techniques as they are developed. In order to illustrate VSM-G concepts, we present a proof-of-concept study of a fast geometrical matching method based on spherical harmonics expansions surfaces. This technique is implemented in VSM-G as the first module of a multiple-step sequence tailored for high-throughput experiments. We show that, using this protocol, notable enrichment of the input molecular database can be achieved against a specific target, here the liver-X nuclear receptor. The benefits, limitations and applicability of the VSM-G approach are discussed. Possible improvements of both the geometrical matching technique and its implementation within VSM-G are suggested.

Keywords Multiple-step virtual screening · VSM-G · Structure-based drug design · Geometrical matching · Spherical harmonics surfaces · SHEF · GOLD · Molecular database enrichment

A. Beautrait · V. Leroux · M. Chavent · L. Ghemtio ·
M.-D. Devignes · M. Smaïl-Tabbone · B. Maigret (✉)
Nancy Université, LORIA, Groupe ORPAILLEUR,
Campus scientifique, BP 239,
54506 Vandœuvre-lès-Nancy Cedex, France
e-mail: bernard.maigret@loria.fr

W. Cai · X. Shao
Department of Chemistry, Nankai University,
Tianjin 300071, People's Republic of China

G. Moreau
330 Avenue Jean Jaurès,
94220 Charanton, France

P. Bladon
Interprobe Chemical Services,
Gallowhill House, Larch Avenue, Lenzie Kirkintilloch,
Glasgow G66 4HX Scotland, UK

J. Yao
Laboratory of Computer Chemistry and Chemoinformatics,
Shanghai Institute of Organic Chemistry,
354 Fenglin Road,
Shanghai 200032, People's Republic of China

Introduction

The search for new drugs is time-consuming and expensive [1], thus any method that speeds up this process is beneficial. Recently, the use of virtual screening (VS) techniques [2] in many drug development strategies has attracted much interest [3]. VS has two obvious advantages: (1) the speed with which a large library of compounds can be screened, and (2) the small initial capital investment compared to the cost of an in vitro high-throughput screening (HTS) program. The first aim of HTS and VS is to reduce a molecular database to a few hit compounds for a protein target. VS, combined or not with HTS, is considered to have been successful when it leads to confirmed hits at lower cost than with HTS alone. Research in this area is particularly active and several success stories have been reported [4–7]. Thus it is now widely accepted that VS calculations can complement HTS experiments [8, 9].

VS methods can have two distinct purposes, with one being the exclusion of a large number of compounds with little or no activity, leading to a limited set of molecules that are more probable hits [10]. Such a method is referred to as a "filter". In the literature, database filtering against a given target is often referred to as "enrichment" [11, 12]. A second purpose is to identify, by ranking input compounds, a small number of candidates likely to be potent. In all VS filters there is a trade-off between speed and accuracy. Filters are optimised for speed; the fastest filters can handle up to a few million molecules, but are notoriously imprecise in reducing this number to less than a thousand while retaining all potential hits. More costly techniques, which can be used in lead optimization strategies, can tackle this problem [13, 14] but not with several million molecules as input and sensible computation times [15]. Therefore, VS protocols are often based on a single or a few fast filters, and are used prior to experimental screening. However, in the latter case, VS usage is limited to that of a pre-filter for HTS, reducing the number of compounds to be tested experimentally, and hence the cost of experiments, by at least one order of magnitude [16, 17].

We have devised a platform for virtual screening called VSM-G (virtual screening manager for computational grids). Our objective with VSM-G is to provide a user-friendly tool that will provide scientists with a large range of in silico strategies for finding hits. Two kinds of approaches can be employed here: ligand-based and structure-based [18, 19]. At present, VSM-G uses structure-based methods to rank input compounds according to their affinity for a target. Thus it can prioritise them for experimental testing. Ligand-based modules, such as substructure searches, can be involved as pre-processing steps to screen molecular databases and reduce the number of compounds requiring subsequent consideration. This initial operation can precede the central element of the platform, the screening funnel, a multi-step structure-based filtering process that hierarchically combines several docking methods.

After describing the VSM-G platform, we will present a proof-of-concept study in the filtering/enrichment context using the liver-X-receptor β (LXRβ) as a target for a screening calculation against a diverse ligand database. The VSM-G screening funnel, consisting of a fast geometrical matching filter preceding flexible docking, was used. This approach is compared to using flexible docking alone for VS. The benefits and limitations of geometrical matching as part of the screening funnel approach, in terms of computing efficiency, applicability and relevance, are discussed.

Overview of the VSM-G platform

Aims and scope of VSM-G

The first step of the pre-clinical drug discovery process can be simplified as a work of exploration at the intersection of distinct spaces [20]. The first of these is the proteome, whose exploration in the drug design context involves its restriction to the sub-space of proteins whose interactions could be therapeutically significant as novel targets—the target space. The second space starts from the even larger ensemble of synthesizable small chemical structures. The exploration here involves sorting out molecules with no, or unwanted, biological effects, restraining the chemical space [21] to the so-called "drug space" [22]. Eventually, merging the target space with the drug space leads to a third ensemble of receptor–ligand associations that have to be explored successfully in order to solve the drug discovery problem. Even if the ensembles of targets and candidate molecules have been previously reduced efficiently to avoid a combinatorial explosion, this is still a long and arduous process.

VSM-G rationalises these searches by focusing on the exploitation and management of current knowledge of the

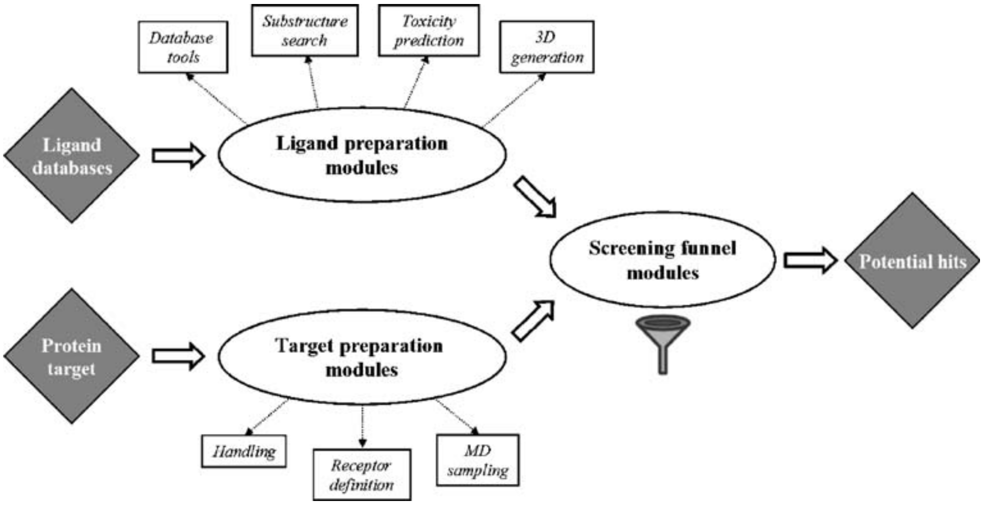


Fig. 1 Basic VSM-G (virtual screening manager for computational grids) workflow for hit discovery

proteome-to-target and chemical-to-drug steps. It also relies on a specific protocol exploiting structure-based VS methods regarding the final ligand-to-hit process. The VSM-G workflow (summarised in Fig. 1) has been designed to match the processes described above. The basic organisation of the platform is therefore divided into three distinct parts: two for the preparation of input data (ligands and protein targets respectively), with the third part being a multi-layer funnel for in silico screening.

Current status

The key features of VSM-G are as follows:

1. Wide coverage of the VS process, from ligand and target preparation to the screening setup, monitoring of calculation processes, and final analysis of the results.
2. Unified and user-friendly graphical interface (see Fig. 2). Seamless integration of the modules, e.g. intercommunication procedures, such as file format conversions, are automatic and transparent to the user.
3. Easy code maintenance, with modular design and choice of widely used programming languages (Java, C, C++ and Fortran).
4. Access to grid technology to take advantage of distributed computing involving computer- and cluster-grids.
5. VSM-G relies on third-party software for performing specific tasks, or in order to provide several choices of techniques for a given purpose. Due to its modular design, VSM-G is readily useable even if those external programs are not installed on the host computer. One of the main development goals of VSM-G is to provide at least one free, open-source solution for each task, which is not currently the case (e.g. at the moment GOLD is the only choice for performing flexible docking).

Charts 1 and 2 list the VSM-G features regarding the ligand database preparation and its target-related capabilities, respectively. Current development is concentrated mostly on the screening funnel.

Chart 1

Current VSM-G features: ligand database preparation.

Database creation and handling

- generation of virtual combinatorial libraries from chemical scaffolds and fragments
- merging of molecular files, with detection of duplicate structures
- support for different file formats, the most popular SDF [23] and MOL2 [24] as output
- conversion between formats using in-house code or OpenBabel [25]
- implementation of the MarvinBeans library [26] and VIDA [27] for database browsing (if available)

Substructure search

- flexible criteria through combinations of simple operators (and, or, not, have, at least, at most…)
- support for SMILES [28], SDF and RDF [23] as input
- internal use of a canonical topology coding that greatly reduces the complexity of the requests
- quickly searches through millions of compounds on desktop computers once the coding is performed

Toxicity prediction

- implementation of PCT [29], a carcinogenicity prediction program based on SAR
- exclusion of presumably toxic compounds
- possible enrichment of the database of substructures associated with poor chemical stability or toxicity

3D structure generation

- fragment-based 3D structure generation program
- the fragment database (> 10,000 structures) can be enhanced / extended by the user
- CORINA [30], which shares the same concept, can be used alternatively (if available)
- post-processing options: protonation (at pH = 7); conformational sampling using OMEGA [27] (if available)

Chart 2

Current VSM-G features: target preparation.

Handling of protein structures

- automatic checking and cleaning of input PDB files with respect to PDB standards [31]
- protein structures can be checked using the MOLPROBITY server [32]
- correction of protonation states: link to the H++ web server [33]
- relaxation of the hydrogen positions upon energy minimization
- link to the STING [34] web-based suite of programs for data mining

Receptor definition

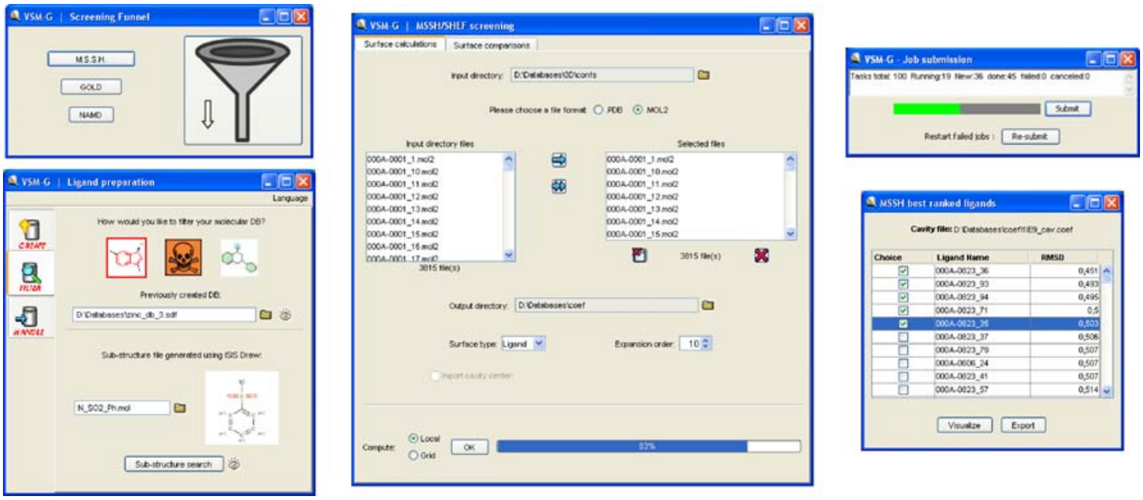
- holoproteins: receptor assumed to be located at the center of mass of the ligand
- apoproteins: generation of an interactive protein 2D map with MSSH [35,36] and VMD [37] for picking up surface receptors
- manual definition of receptors can be imported from VMD selections, and exported to funnel modules
- handling of resident water molecules, potentially useful with some docking programs [38]

Multiple target conformations management

- handling of multiple X-ray structures
- enrichment through MD sampling [15,39], using VMD and NAMD [40]
- clustering, averaging and minimization of conformations from NMR data or MD sampling

The screening funnel: a multiple-step strategy

A wide variety of virtual screening programs are currently available, and it is generally assumed that a well-chosen combination of methods will give better results than any single method alone. The interest in such multiple-step VS protocols, often as a combination of a single structure-based docking calculation with ligand-based approaches as pre-filters, has been stressed in several papers [5, 41]. Post-processing refinements starting from docking results have also been reviewed [15, 42]. Alternatively, several methods can be employed at different stages within a given docking program [43]. The use of several docking programs in the same protocol [44] is less frequent. Moreover, most



How would you like to filter your molecular DB?

Previously created DB: D:\Databases\zinc_db_3.sdf

Sub-structure file generated using ISIS Draw: N_SO2_Ph.mol

Sub-structure search

Input directory: D:\Databases\3D\confs

Please choose a file format ○ PDB ● MOL2

Input directory files:
000A-0001_1.mol2
000A-0001_10.mol2
000A-0001_11.mol2
000A-0001_12.mol2
000A-0001_13.mol2
000A-0001_14.mol2
000A-0001_15.mol2
000A-0001_16.mol2
000A-0001_17.mol2
3815 file(s)

Selected files:
000A-0001_1.mol2
000A-0001_10.mol2
000A-0001_11.mol2
000A-0001_12.mol2
000A-0001_13.mol2
000A-0001_14.mol2
000A-0001_15.mol2
3815 file(s)

Output directory: D:\Databases\coef

Surface type: Ligand

Expansion order: 10

□ Import cavity center:

Compute: ● Local ○ Grid OK 83%

Tasks total: 100 Running:19 New:36 done:45 failed:0 canceled:0

Restart failed jobs :

Cavity file: D:\Databases\coef\1E9_cav.coef

Choice	Ligand Name	RMSD
☑	000A-0823_36	0,451
☑	000A-0823_93	0,493
☑	000A-0823_94	0,495
☑	000A-0823_71	0,5
☑	000A-0823_35	0,503
□	000A-0823_37	0,506
□	000A-0823_79	0,507
□	000A-0606_24	0,507
□	000A-0823_41	0,507
□	000A-0823_57	0,514

Fig. 2 Screenshots of the VSM-G graphical interface

programs require significant expertise in setting up and analysing the results. More generally, each technique features a specific balance between the speed of calculations and the reliability of results [45]. Open software tools overcoming such limitations are lacking. The virtual screening implementation in the VSM-G platform is constituted by a series of different structure-based methods, organised sequentially in a funnel strategy. The techniques range from simple methods to more sophisticated ones, profiting from the speed of the former and the accuracy of the latter. At each step of the process, the filter discards inappropriate compounds. The simplest and quickest filters are used at an early stage in the filtering process, allowing the more time consuming processes to be used at later stages. The multiple-step screening funnel strategy is shown in Fig. 3.

Methodology

Outline of the proof-of-concept study

Most docking methods are not efficient enough for use in high-throughput VS (i.e. the time required to process >10^6 molecules is out of reach with modern hardware). Fast filtering prior to docking might be a workaround. Ligand-based methods can also prove useful here, but unless large training sets are available for the target, they are of limited value. Geometrical matching procedures, which are orders of magnitude faster than common docking methods, can be employed in this particular context [46], and can lead to discovery of hits [47], but few studies estimating their impact in a general VS experiment exist.

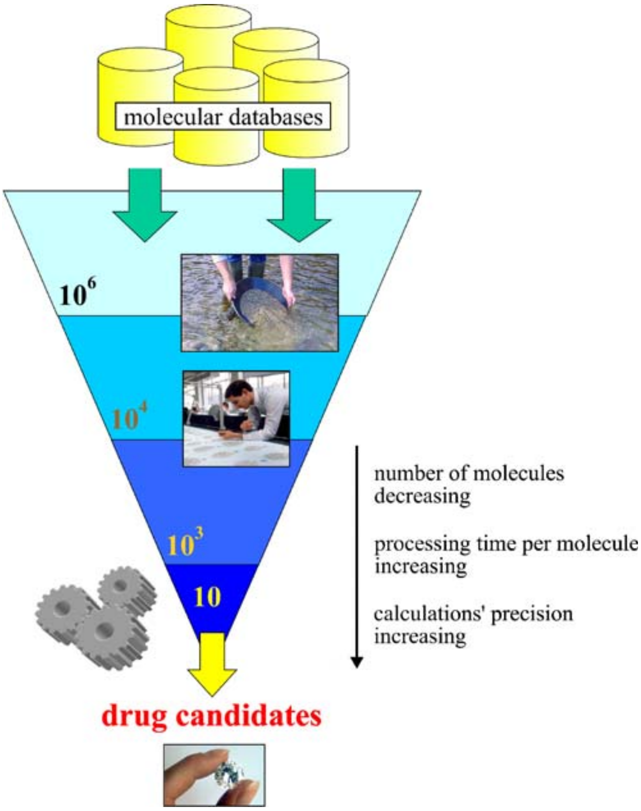

molecular databases
10^6
10^4
10^3
10
number of molecules decreasing
processing time per molecule increasing
calculations' precision increasing
drug candidates

Fig. 3 Basic principle of the virtual screening funnel process

The geometrical matching procedure evaluated here is a two-step process. First, the MSSH program [35, 36] approximates the geometry of molecular structures using a series of spherical harmonics functions. This representation is very compact as all information is contained in the expansion coefficients, while the corresponding surfaces still provide a good level of detail. Additionally, this process can be performed once and for all for each protein and ligand conformer. Afterwards, evaluating the surface complementarity between a target active site and a ligand is performed through simple and efficient operations [48] specific to spherical harmonics algebra. This very fast procedure is performed with the SHEF program [W. Cai et al. manuscript in preparation], which identifies and scores the geometrically optimal orientation of each ligand conformation for the target. These techniques are described in depth by Cai et al. [35, 36, W. Cai et al. manuscript in preparation]

In this paper, we study a VSM-G-operated screening funnel using MSSH/SHEF followed by flexible docking using GOLD [49, 50]. Such an approach involves using SHEF results to filter out part of the input ligand database before proceeding to the second funnel step relying on GOLD. In this proof-of-concept study we did no such filtering; all molecules of the test set are evaluated with both techniques in order to simulate the screening funnel for all levels of filtering between the two steps.

Target preparation

The liver X receptors (LXRs) [51] represent attractive targets for the development of new therapeutic agents for treating multiple (especially cardiovascular) diseases [52]. Several structures of the ligand binding domain of LXR, co-crystallised with various ligands, have been determined by X-ray crystallography. Reports on structural analysis reveal great plasticity of the ligand binding pocket, which is able to accommodate ligands with noticeably different shapes and sizes [53]. In this work, we study more specifically the LXRβ isoform, for which we took as a starting point different X-ray structures available from the Protein DataBank (PDB) [54]: 1P8D [55], 1PQ6 [53] and 1PQ9 [53]. For each of these structures the most complete chain was retained: chain A for 1P8D and chain B for 1PQ6 and 1PQ9. In all cases the binding area was complete and the Cα trace superimposed well, allowing missing fragments to be added using homology modeling. Protonation

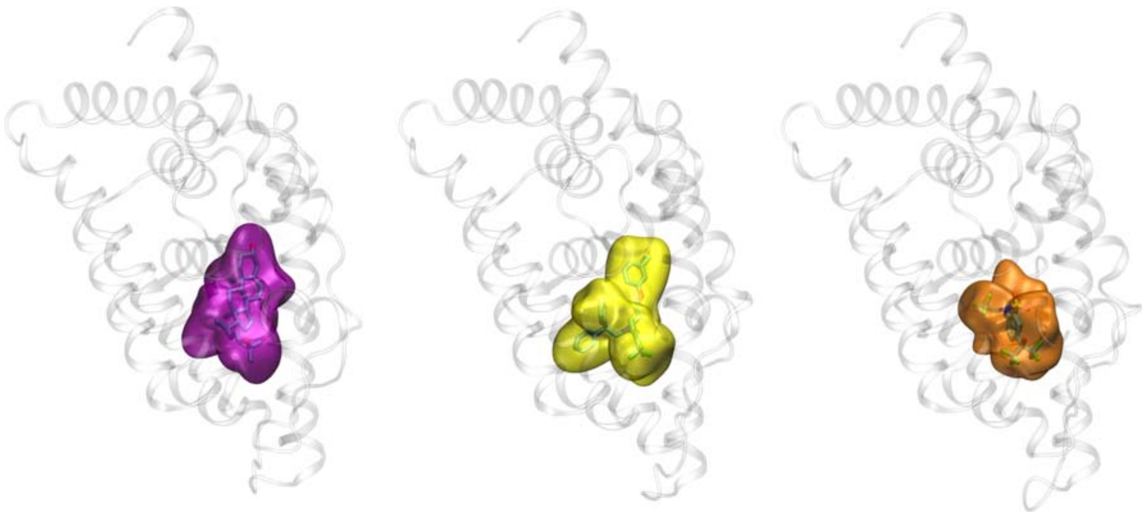


Fig. 4 Shapes of the 1P8D, 1PQ6 and 1PQ9 active sites (from *left* to *right*) as approximated by spherical harmonics expansion surfaces using MSSH. The X-ray ligands filling the active sites are shown

was performed at pH 7 with VSM-G. The imidazole tautomer of the active site histidine residue is the one at $N^{\delta 1}$-H [56].

Figure 4 shows that the three binding site conformations, represented by their MSSH-generated surfaces imported into VMD [37], are clearly distinct geometrically. The 1PQ9 cavity is significantly smaller (810 Å^3) than those of 1PQ6 (996 Å^3) and 1P8D (1,014 Å^3). 1PQ6 has a less-spherical, more specific shape. Therefore, it could be expected that (1) 1P8D is the least selective upon ligand binding, (2) 1PQ6 shape specificity could be overcome by ligand flexibility, and (3) the 1PQ9 conformation should filter out more structures based on their size.

The protein–ligand binding modes depicted in the three experimental structures have also been analysed. The shared characteristics are dominated by hydrophobic interactions with F_{271}, F_{329} and F_{340}. 1PQ6 allows for a possible specific charge–charge interaction with R_{319}. R_{319} already makes an internal interaction with E_{281} in the 1P8D conformation, dampening the strength of possible ligand interaction. In the case of 1PQ9, neither of those residues is accessible as the pocket size is restricted by a particular F_{329} orientation.

Ligand database preparation

The starting database is composed of compounds commercially available from three suppliers, ChemDiv [57], Enamine [58] and Comgenex [59], in March 2006. Filtering using Lipinski's rule-of-five [60] was performed, allowing a single violation for each structure, giving a total of 598,375 unique molecules. In order to reduce the database size while retaining as much chemical diversity as possible, we used ScreeningAssistant software [61]. This tool characterises each molecule of the database using SSKey-3D 54-bit fingerprints [62], allowing for similarity estimation between pairs by computing Tanimoto coefficients [63]. Database clustering can then govern the generation of diversity-maximised subsets. In our case, we targeted a 10,000 molecule subset and obtained a database of 8,383 compounds.

A reference diverse database was defined by merging the initial database of 598,375 molecules with the Chimiothèque Nationale (CN) database [64, 65]. Diversity of each of the three subsets (the 598,375 database, the 8,383 diversity set and the 31,220 CN) was expressed as a fraction of the total diversity [61] (Table 1). It appears that the 8,383 subset and

Table 1 Diversity analysis of the reference database used in this paper, here referred to as the diversity subset of 8,383 compounds. In the table, 100% diversity is that of the union of the large-scale and the CN (Chimiothèque Nationale) databases. All values are computed by the ScreeningAssistant software. Please refer to Monge et al. [61] for details of how drug-like and lead-like compounds are defined, and how molecular database diversity is measured

Database	Number of compounds	Drug-like compounds	Lead-like compounds	Drug-like diversity	Lead-like diversity	Global diversity
Large-scale	598,327	563,777 (94.2%)	195,332 (32.6%)	84.3%	82.3%	81.8%
Diversity subset	8,383	7,875 (93.9%)	3,178 (37.9%)	50.0%	43.5%	48.3%
CN	31,220	27,403 (87.8%)	20,295 (65%)	41.4%	44.8%	43.7%

the larger CN database are of comparable diversity. The former is therefore suitable as input data for a VS validation experiment. Interestingly, from the large scale database to the diversity subset, we traded only ~40% of the diversity for a 98.6% size reduction.

The 8,383-compound database was pre-processed into VSM-G ligand preparation modules, which made it suitable for subsequent docking programs. Molecules were first converted into 3D, and their protonation state was then set arbitrarily at pH=7. As MSSH/SHEF is a rigid shape-matching procedure, a conformational search was performed (retaining at most 400 conformers per compound), giving 1,102,299 conformers.

Parameterisation of the virtual screening programs

A total of 1,102,299 conformers were docked using SHEF in the three target conformations, giving 3,306,897 rigid docking calculations. Using GOLD, 8,383 molecules were docked, giving 25,149 flexible docking calculations. The program parameters used, which favoured reliability over speed, are listed in Chart 3.

Chart 3

Parameters for MSSH, SHEF and GOLD used for the validation study simulating the use of MSSH/SHEF for filtering prior to GOLD calculations.

MSSH [35,36] **/ SHEF** [48, W. Cai et al. manuscript in preparation]

- spherical harmonics expansion of order 10
- cavity coordinates defined using the ligand center of mass

GOLD [49,50]

- default genetic algorithm parameters
- 50 dockings / molecule
- early termination option: docking stopped if the top 5 conformations fall within 1.5 Å RMSD range
- cavity definition: flood fill (works well when the receptor is not open and extended)
- same cavity coordinates as with MSSH/SHEF
- scoring function: GoldScore

Definition and relevance of reference data

The reference data for evaluating SHEF performance is constituted by GOLD results and not by experimental data. Like all docking programs, GOLD does not provide 100% success in reproducing conformations and binding free energies of protein–ligand complexes [66]. Hence the reference set is approximate and cannot be used to measure SHEF performance precisely. However, our aim here is simply to demonstrate SHEF usefulness as part of the VSM-G screening funnel, in a large-scale VS context. Consequently, a chemically diverse reference set that is large enough statistically seems appropriate despite GOLD-related limitations.

In order to evaluate filtering, the reference molecular database has to be divided in two subsets, the first corresponding to the (presumably) most potent molecules (referred to as the hit compounds subset) that will be conserved upon filtering, and the second being considered as inactive structures for the target. GOLD score values are used to rank ligands against the three target conformations, and the top 10% best-ranked ligands are selected from each of the three sets. This cutoff value is set arbitrarily. Ranks are used to select ligands instead of score values because molecular dynamics simulations performed in our laboratory on LXRβ indicate that important induced fit effects [67] could occur upon ligand binding. This suggests that the GOLD scoring function, which does not account for receptor internal energy, may correlate only with the global free energy of binding across a single receptor conformer [1].

As shown in Fig. 5, the three ensembles of 838 selected structures overlap, giving a classification of hits into different families regarding their selectivity for the three target conformations. Out of a total of 1,414 molecules, 670 (47%) bind specifically to one of the three conformations, 356 (25%) bind to all three conformations, with the remainder binding two out of three. The amount of selective molecules for each conformation is 20%, 24%, and 36% for 1P8D, 1PQ6 and 1PQ9, respectively, which is in agreement with the structural specificities highlighted previously.

Analysis of results

An in-house program was created for representing relationships between the screening results of two different techniques for the same set of input data. Figure 6 explains the principles of the generated graphical representation. Both rank ranges are divided in 20 5% blocks—a sensible trade-off between graphical clarity and the amount of information represented. Three particular cases are provided as examples. Figure 6a depicts random selection, while

[1] Redocking experiments of LXRβ reference ligands present in the X-ray structures back up this hypothesis. Using GOLD, the 1PQ6 ligand redocked in the 1PQ6 binding pocket conformation yields a significantly higher score than the 1PQ9 ligand redocked in the 1PQ9 conformation. However, according to experimental data, the 1PQ9 ligand is indeed clearly more potent on LXRβ than the 1PQ6 ligand, further indicating that the protein–ligand interaction could not be the dominant term in the free energy of binding

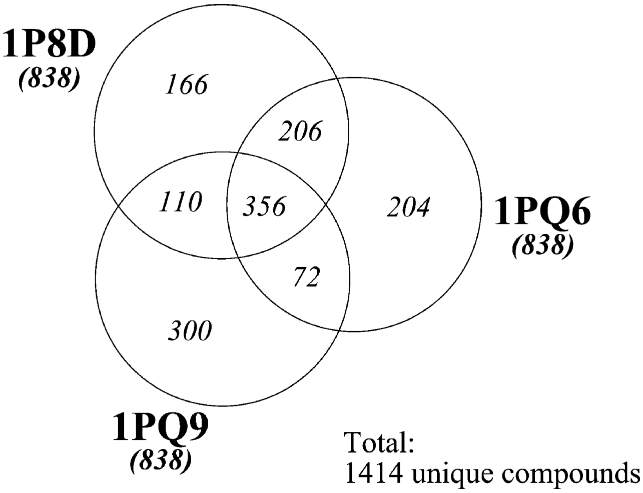


Fig. 5 Populations of hits defined from GOLD results of the 8,383-compound diverse database. For each target conformation (1P8D, 1PQ6 and 1PQ9), the top-scoring 10% structures are defined as hits. The overlap of these three sets is represented. A total of 1,414 hit compounds were defined as the target subset that has to be conserved through the filtering process

Fig. 6c corresponds to a perfect correlation. The results of any given filtering process will obviously lie between these two. Figure 6b illustrates another ideal case for filtering, but only for a precise filtering amount (which may or may not be satisfactory).

The Spearman [68] ρ and Kendall [69] τ coefficients are employed as measures of correlation:

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} \Delta r_i^2$$

$$\tau = -1 + \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta\left(r_j > r_i\right)$$

where r_i is the SHEF ranking of the i–ranked GOLD structure; δr_i is the difference between these two ranks ($\Delta r_i = r_i - i$). δ is the boolean function: δ (true)=1 while δ (false)=0. The rankings, in both cases, are in ascending order from the best predicted binding molecule to the worst. We also have $0 \le \rho \le 1$ and $-1 \le \tau \le 1$, with 0 indicating an absence of correlation (random selection) and 1 a perfect correlation (same rankings).

Other metrics are used in order to evaluate filtering performance. Given a definition of what represents a hit structure and what does not for a specific target, we can describe the quality, q, of a molecular database of *n* structures as the ratio between the number of hit compounds and the total number of structures:

$$q = \frac{n_{hits}}{n}$$

The enrichment, e, of a database by a filtering process and for a given filtering ratio f ($0 \le f \le 1$; f being the number of filtered out candidates) can be defined as the ratio between the quality of the reduced database and the quality of the initial database:

$$e(f) = \frac{q(f)}{q(0)}$$

Enrichment is commonly used to evaluate the efficiency of the molecular database method. By definition, random selection does not affect quality, so its efficiency is 1 for any filtering amount. The maximum enrichment that can be

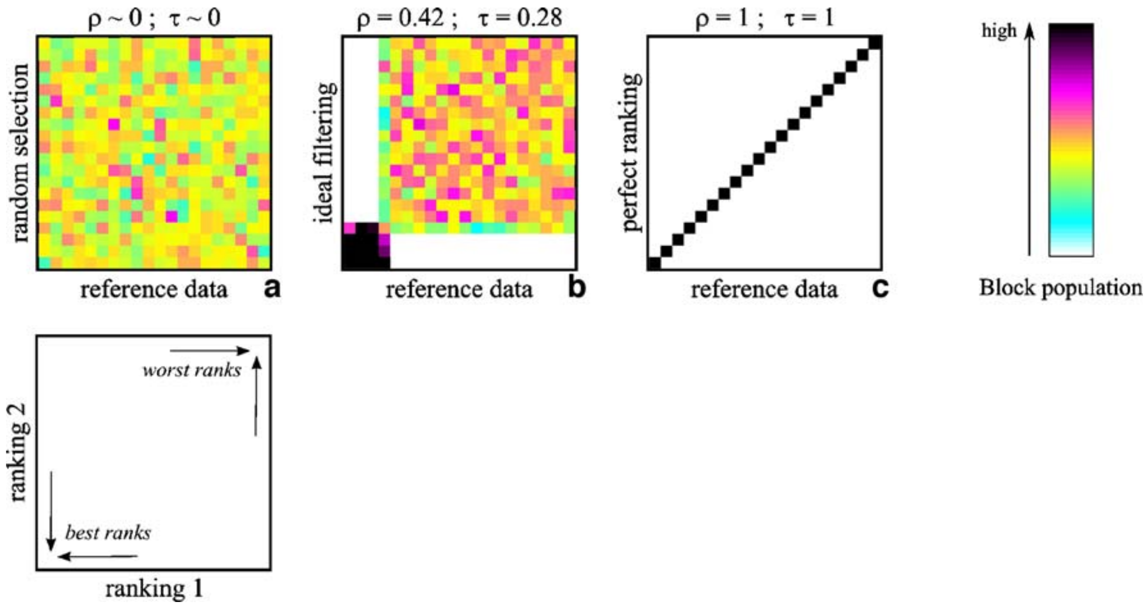


Fig. 6 Explanation of density plot representation of rank correlation, illustrating three particular cases: **a** random selection, **b** ideal filtering, **c** perfect correlation

obtained for a given filtering level is when all hits are retained, which corresponds to:

$$e_{\max}(f) = \frac{1}{1-f}$$

The filtering efficiency, E, is eventually defined as the relative distance of the filtering method from random filtering (E=0) to maximum enrichment (E=1):

$$E(f) = \frac{e(f) - 1}{e_{\max}(f) - 1}$$

Results

Influence of target conformation on GOLD and SHEF results

The density plots shown in Fig. 7 give a picture of how target conformation specificities influence GOLD and SHEF results. The SHEF correlation between 1P8D and 1PQ9 (Fig. 7b) is greater than those between 1PQ6 and both 1P8D (Fig. 7a) and 1PQ9 (Fig. 7c). This is in agreement with the observation that the 1PQ6 shape is the most specific. In the case of GOLD, it first appears that 1P8D and 1PQ6 results are highly correlated (Fig. 7d). The correlations with 1PQ9 (Fig. 7e,f) are lower. A significant number of structures performing well with both 1P8D and 1PQ6 are ranked low with 1PQ9, indicating a group of ligands whose size fits well into the former active site conformations but not into the smaller 1PQ9. Surprisingly, such an expected group does not appear in SHEF results.

The results show that SHEF, which is a surface-based method, appears more sensitive to active site shape specificities than GOLD, which relies on a classical atom coordinate-based representation of molecular structures. However, in contrast to GOLD, SHEF appears unable to assess size constraints correctly. This could be related not to SHEF itself but rather to its current implementation within the VSM-G screening funnel. Indeed, only the best conformer score is retained for ranking each compound; the diversity of geometrically acceptable conformations (referred to as adaptability) is not taken into account. This could lead to SHEF producing false positives with ligands occupying almost all the active site volume. Such ligands might require a minimal adaptability in order to provide a good chance to satisfy chemical constraints upon binding, in addition to geometrical complementarity.

Relationship between SHEF and GOLD classifications

Figure 7g–i depicts the relationships between SHEF and GOLD ranks for 1P8D, 1PQ6 and 1PQ9. Given the fundamental differences between these two programs, it is not surprising to see lower correlation between SHEF and GOLD than between the two different target conformations for either SHEF or GOLD. We are, however, far from the random case depicted in Fig. 6a, thus it is clear that noticeable enrichment using SHEF is already observed at this point.

Although the general profile of the three density plots is similar, they differ regarding the distribution of false positives, i.e. populations located at the bottom right corners, corresponding to molecules whose binding ranks are overestimated by SHEF according to GOLD results. In agreement with previous observations, it appears that SHEF generates most false positives when docking on the 1PQ9 conformation, while correlation between GOLD and SHEF is best in the 1PQ6 case, which presents a more specific shape that should favour SHEF efficiency.

Interestingly, Fig. 7j shows that the correlation between the SHEF and GOLD consensus rankings is higher than the average of the GOLD–SHEF correlation for the three receptor conformations. Additionally, such an approach could be more interesting than the 1PQ6-only filtering of Fig. 7h, which naturally favours ligands more specific to 1PQ6. Even if the corresponding correlation is higher, it is probably more important to favour diversity regarding target conformations if precise information concerning their relative stability is unknown.

SHEF as a first-step enrichment filter in the screening funnel protocol

It should first be noted that the ligands present in the 1P8D, 1PQ6 and 1PQ9 experimental structures, redocked using GOLD, fall within the range of the hits subset as defined above. These reference ligands are also amongst the top 2% structures according to SHEF calculations. Therefore, unless the filtering ratio is set too high, they would be retrieved in a SHEF/GOLD screening funnel experiment.

Taking the SHEF consensus ranking as a reference, we plotted the variation of the population of GOLD hits as a function of the filtering ratio. The resulting curve is shown in Fig. 8 together with the enrichment curves that would result from random selection and from the ideal case where the 1,414 hits are all ranked before the other 6,969 molecules. A clear enrichment is observed on all ranges of filtering. There is still much room for improvement, but present SHEF performance is interesting considering that SHEF and GOLD are not in the same league in terms of speed and precision. In the virtual screening context, if the number of molecules to screen with GOLD using available computing power is too high, SHEF could provide a rational solution to decreasing the number of candidate molecules without limiting too much the chances of finding novel hit compounds for a given target.

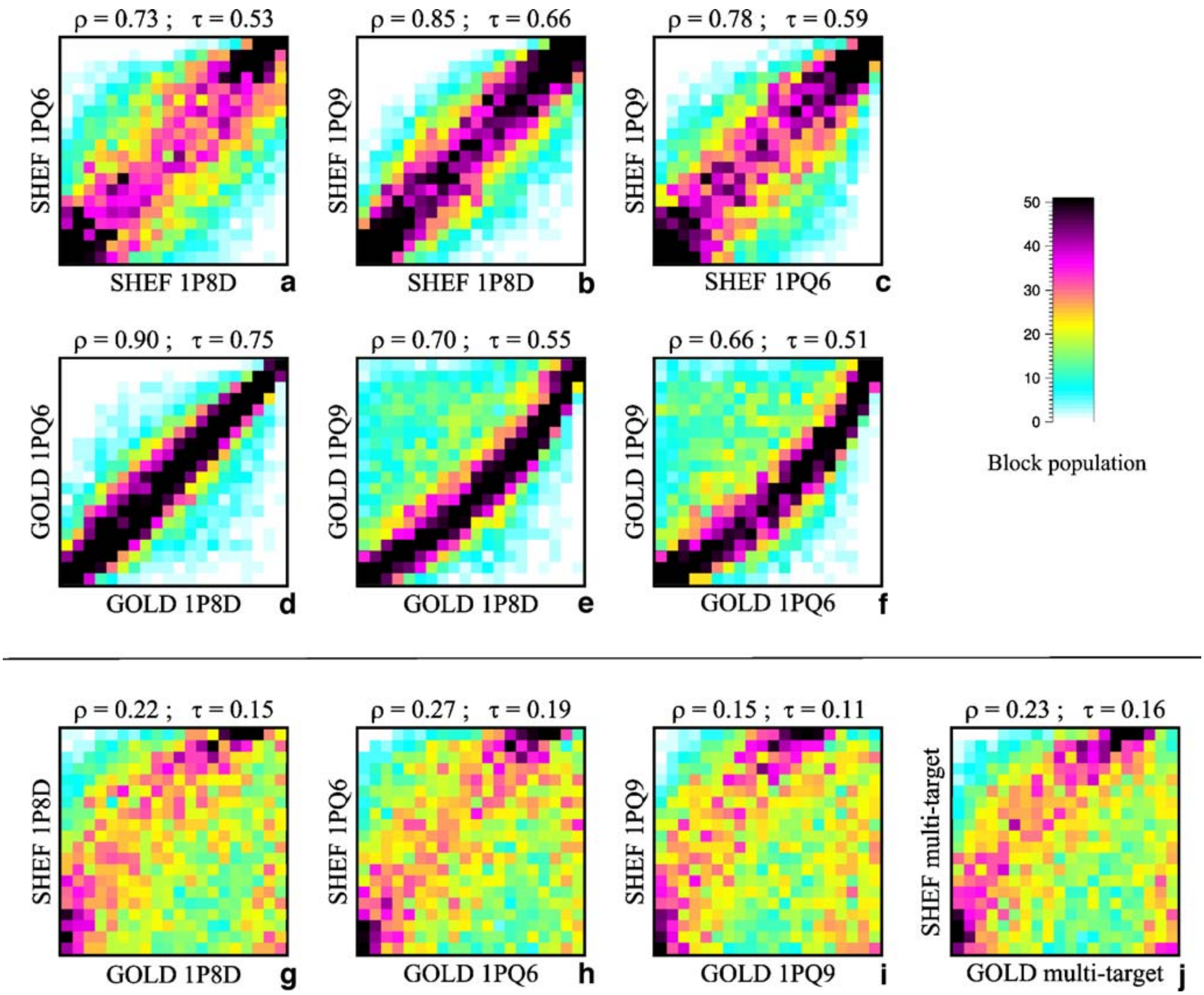


Fig. 7 Density plots between rankings. The first six plots depict the relationships between the different target conformations, for SHEF (**a**, **b**, **c**) and GOLD (**d**, **e**, **f**). Target conformation influence on these two programs can therefore be observed. The last four plots show the relationship between SHEF and GOLD results, for the three target conformations (**g**, **h**, **i**), then using multiple-target rankings (**j**). The scale is set so that the average $(5\%)^2$ block density is 8,383 / 400 ~ 21. Further explanation of these representations can be seen in Fig. 6

Correlation between SHEF efficiency and the nature of the protein–ligand binding mode

We will now focus on results for two specific, arbitrarily chosen, filtering ratios: 0.1 (low filtering, 90% of molecules retained) and 0.5 (50% of molecules filtered out). In order to determine whether particular families of molecules could influence SHEF filtering efficiency, the variation of all hit populations as defined in Fig. 5 for the four possible SHEF rankings (on the 1P8D, 1PQ6, 1PQ9 targets, and multiple-target consensus) was collected. The results are shown in Table 2. This data was translated in terms of filtering efficiency E(f) in Table 3. The main result can be interpreted as follows: if we apply 10% and 50% filtering, respectively, using the multiple-target SHEF filter, amongst all hits we will retain 90.8% and 52.8%, respectively, of what would have been lost using random selection.

The comparison between the four available filters based on SHEF rankings suggests the use of multiple-target consensus ranking as the best choice. This is in agreement with observations made by analysing Fig. 7g–i. More interestingly, analysis of SHEF efficiency of the different hits sub-groups reveals that molecules specific to the 1PQ6 target conformation according to GOLD perform poorly with SHEF (see Table 3, ”1PQ6-specific” line). It has been shown that the specific 1PQ6 shape is taken into account by SHEF, but 1PQ6 also presents a second peculiarity: the accessibility of a charged residue. The corresponding

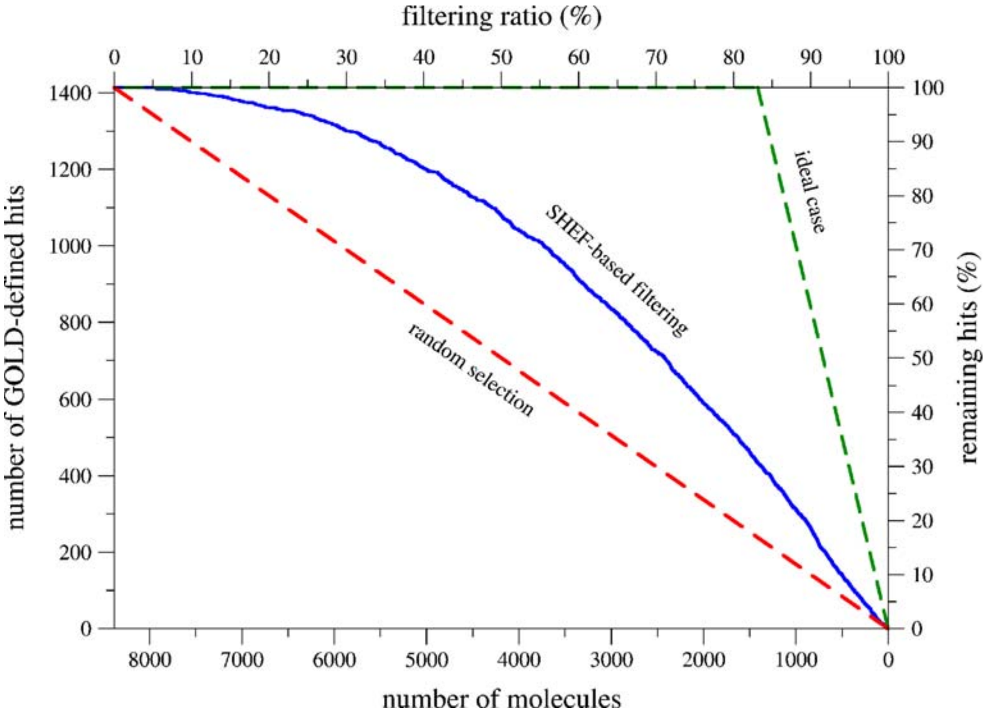


Fig. 8 Enrichment curve of SHEF as measured in the validation experiment, depending on the degree of filtering applied. Reference results are the GOLD rankings, which were used to define a target subset of 1,414 structures (referred to as hits) out of a starting number of 8,383. Multiple-target rankings were used in both cases (*i.e.* the rank of each molecule is the best rank amongst the 1P8D, 1PQ6 and 1PQ9 classifications). The two *dotted curves* represent random selection and perfect correlation (in which SHEF would reproduce GOLD results perfectly); thus the filtering efficiency (E) for a filtering ratio (f) can be measured as the relative position on the y-axis of the SHEF enrichment curve between these two

1PQ6-specific ligands most probably share a binding mode dominated by electrostatic effects that SHEF, which compares only geometries, is unable to assess. In contrast, molecules that are defined as hits for all three of the LXRβ pocket conformations are those for which SHEF filtering is the most efficient for both values of filtering (see Table 3, row “1P8D+1PQ6+1PQ9”). These molecules might have a high degree of adaptability, allowing SHEF to perform well in identifying conformations with the best steric complementarity.

Discussion and concluding remarks

In this study, we wanted to present an overview of VSM-G, and to more precisely evaluate the usefulness of the SHEF geometrical matching procedure as part of the VSM-G multiple-step high-throughput VS procedure. We used score values from the flexible docking program GOLD as reference data, allowing a qualitative assessment of MSSH/SHEF efficiency as a first fast filter for the multiple-step VSM-G procedure. Thus, even considering the limitations of our validation test, the results are clear enough to demonstrate that SHEF, and by extension its association as the first module in the VSM-G screening protocol, can indeed be useful for in silico drug discovery.

This paper has highlighted precisely the conditions required to obtain good performance from MSSH/SHEF. It appears that for flexible receptors prone to induced fit effects upon complexation, a filtering based on a consensus ranking of SHEF results for multiple target conformers should be favoured. More importantly, basic information regarding the types of interactions involved in ligand binding is crucial for deciding if MSSH/SHEF should be used and if so to what extent. Enrichment can be expected only when binding is not largely dominated by chemical interactions such as electrostatic effects or hydrogen bonding. Active sites that are known to favour hydrophobic interactions might be targets of choice for a structure-based drug design strategy involving MSSH/SHEF as part of a multiple-step VS procedure set up using the VSM-G program.

Limitations of the spherical harmonics-based geometrical matching procedure have been pointed out. As with all structure-based in silico techniques, two fundamental aspects govern how protein–ligand binding is modelled. Firstly, the way in which search space is defined, and secondly, how this space is explored. An improvement of SHEF in the former aspect would involve taking into account basic chemical properties to extend the complementarity score that is currently computed. Such an approach has already been tried out in the ligand-based drug design area [70]. Regarding the exploration strategy, in its current implementation in VSM-G, SHEF acts as a rigid docking program that selects only a single conformer from the list of conformers for a given structure; this approach has been shown here to produce significant numbers of false positives in some cases. One alternative could be to use a diverse set of docked conformers for each

Table 2 Evolution of the GOLD hits subsets population when applying 10% and 50% SHEF-based filtering. The groups defined in Fig. 6 are studied separately, while for SHEF filtering, the results for each of the three target conformations are presented as well as those using the multiple-target consensus ranking. Note: the multiple-target / all hits results (bottom right) can be measured directly on the curve in Fig. 8

Population of the different GOLD hit groups after SHEF filtering		Initial population	SHEF-based filters							
			1P8D		1PQ6		1PQ9		Multiple-target	
			10%	50%	10%	50%	10%	50%	10%	50%
GOLD-based *hit* groups	1P8D	838	834	672	835	702	827	652	832	688
	1PQ6	838	828	620	833	664	817	591	826	644
	1PQ9	838	832	632	837	707	835	660	838	687
	1P8D-specific	166	165	130	165	125	164	117	165	130
	1PQ6-specific	204	198	116	201	125	192	96	197	121
	1PQ9-specific	300	295	194	299	222	297	212	300	213
	1P8D+1PQ6	206	203	156	204	151	197	137	201	142
	1P8D+1PQ9	110	110	90	110	97	110	90	110	93
	1PQ6+1PQ9	72	71	52	72	59	72	50	72	58
	1P8D+1PQ6+1PQ9	356	356	296	356	329	356	308	356	323
	all hits	1414	1398	1034	1407	1108	1388	1010	1401	1080

ligand, the selection between them being made by a second module in the screening funnel protocol. Various techniques are being considered in this regard [71–73].

In any case, it is uncertain that improvements of the SHEF algorithm would necessarily be worthwhile. At the present time the main advantage of the MSSH/SHEF approach is its speed. With the safe parameters used in this report, SHEF is typically 2–3 orders of magnitude faster at processing >10^6 conformers than GOLD is for docking the corresponding ~10^4 structures. MSSH is still 1 order of magnitude faster than GOLD, and its calculations can be done once and for all for a given molecular database. Enhancements of the MSSH and SHEF programs should obviously not be made at the cost of the loss of this computing speed advantage, which allows for large-scale structure-based VS to be performed.

In future work, we will focus on selection rather than on filtering capability. This will include a proof-of-concept study of the usefulness of post-docking optimisations and molecular dynamics calculations as funnel modules following geometrical matching and flexible docking. Next, we will illustrate the whole screening funnel strategy through an actual large-scale hit discovery campaign using computer grid architectures. The relevance of using advanced techniques such as target sampling and grid computations in such a context will also be highlighted.

Table 3 Values of the SHEF filtering efficiency E(f) for f = 10% and f = 50%. These values are directly correlated to those of Table 2

SHEF filtering efficiency (%)		SHEF-based filters							
		1P8D		1PQ6		1PQ9		Multiple-target	
		10%	50%	10%	50%	10%	50%	10%	50%
GOLD-based hit groups	1P8D	95.2	60.4	96.4	67.5	86.9	55.6	92.8	64.2
	1PQ6	88.1	48.0	94.0	58.5	74.9	41.1	85.7	53.7
	1PQ9	92.8	50.8	98.8	68.7	96.4	57.5	100	64.0
	1P8D-specific	94.0	56.6	94.0	50.6	88.0	41.0	94.0	56.6
	1PQ6-specific	70.6	13.7	85.3	22.5	41.2	−5.9	65.7	18.6
	1PQ9-specific	83.3	29.3	96.7	48.0	90.0	41.3	100	42.0
	1P8D+1PQ6	85.4	51.5	90.3	46.6	56.3	33.0	75.7	37.9
	1P8D+1PQ9	100	63.6	100	76.4	100	63.4	100	69.1
	1PQ6+1PQ9	86.1	44.4	100	63.9	100	38.9	100	61.1
	1P8D+1PQ6+1PQ9	100	66.3	100	84.8	100	73.0	100	81.5
	all hits	88.7	46.3	95.0	56.7	81.6	42.9	90.8	52.8

Acknowledgements We thank Yesmine Asses, Safia Kellou and Amel Maouche for their feedback. Alexandre Beautrait was supported by grants from INRIA (Institut National de Recherche en Informatique et en Automatique), Région Lorraine, and ARC (Association pour la Recherche sur le Cancer); Vincent Leroux by a post-doctoral fellowship from the INCa (Institut National du Cancer); Matthieu Chavent by a joint fellowship between CNRS (Centre National pour la Recherche Scientifique) and Région Lorraine. We thank Openeye for providing free access to OMEGA and VIDA software according to an academic license, Chemaxon for supplying MarvinBeans Java library, CCDC for the trial version of the GOLD program, and the laboratory of chemoinformatics at the Orléans University for the ScreeningAssistant program.

References

1. DiMasi JA, Hansen RW, Grabowski HG (2003) J Health Econ 22:151–185
2. Shoichet BK (2004) Nature 432:862–865
3. Stahura FL, Bajorath J (2004) Comb Chem High Throughput Screening 7:259–269
4. Perola E, Xu K, Kollmeyer TM, Kaufmann SH, Prendergast FG, Pang YP (2000) J Med Chem 43:401–408
5. Grüneberg S, Stubbs MT, Klebe G (2002) J Med Chem 45:3588–3602
6. Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P (2003) J Med Chem 46:2656–2662
7. Kraemer O, Hazemann I, Podjarny AD, Klebe G (2004) Proteins: Struct Funct Bioinf 55:814–823
8. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Conolly DT, Shoichet BK (2002) J Med Chem 45:2213–2221
9. Bajorath J (2002) Nat Rev Drug Discov 1:882–894
10. Abagyan R, Totrov M (2001) Curr Opin Chem Biol 5:375–382
11. Xu H, Agrafiotis DK (2002) Curr Top Med Chem 2:1305–1320
12. Krovat EM, Langer T (2004) J Chem Inf Comput Sci 44:1123–1129
13. Huo S, Wang J, Cieplak P, Kollman PA, Kuntz ID (2002) J Med Chem 45:1412–1419
14. Jenwitheesuk E, Samudrala R (2003) BMC Struct Biol 3
15. Alonso H, Bliznyuk AA, Gready JE (2006) Med Res Rev 26:531–568
16. Waszkowycz B, Perkins TDJ, Sykes RA, Li J (2001) IBM Syst J 40:360–376
17. Bleicher KH, Böhm H-J, Müller K, Alanine AI (2003) Nat Rev Drug Discov 2:369–378
18. Veselovsky AV, Ivanov AS (2003) Curr Drug Targets: Infect Disord 3:33–40
19. Jain AN (2004) Curr Opin Drug Discov Dev 7:396–403
20. Ofran Y, Punta M, Schneider R, Rost B (2005) Drug Discov Today 10:1475–1482
21. Dobson CM (2004) Nature 432:824–828
22. Oprea TI, Gottfries J (2001) J Comb Chem 3:157–166
23. MDL, SD file format. http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp
24. Tripos, Mol2 file format. http://www.tripos.com/data/support/mol2.pdf
25. Open Babel project. http://www.openbabel.sourceforge.net
26. ChemAxon Ltd., Budapest, Hungary. http://www.chemaxon.com/products.html
27. OpenEye Science Software: Santa Fe, NM. http://www.eyesopen.com
28. Weininger D (1988) J Chem Inf Comput Sci 28:31–36
29. Liao Q, Yao JH, Li F, Yuan SG, Doucet J-P, Panaye A, Fan BT (2004) SAR QSAR Environ Res 15:217–235
30. Sadowski J (1993) Chem Rev 93:2567–2581
31. PDB file format. http://www.rcsb.org/pdb/static.do?p=file_formats/pdb/index.html
32. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall III WB, Snoeyink J, Richardson JS, Richardson DC (2007) Nucleic Acids Res 35: W375–W383
33. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A (2005) Nucleic Acids Res 33:W368–W371
34. Neshich G, Mancini AL, Yamagishi ME, Kuser PR, Fileto R, Pinto IP, Palandrani JF, Krauchenco JN, Baudet C, Montagner AJ, Higa RH (2005) Nucleic Acids Res 33:D269–D274
35. Cai W, Zhang M, Maigret B (1998) J Comput Chem 19:1805–1815
36. Cai W, Shao X, Maigret B (2002) J Mol Graph Model 20:313–328
37. Humphrey W, Dalke A, Schulten K (1996) J Mol Graph 14:33–38
38. Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) J Med Chem 48:6504–6515
39. Wong CF, Kua J, Zhang Y, Straatsma TP, McCammon JA (2005) Proteins: Struct Funct Bioinf 61:850–858
40. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) J Comput Chem 26:1781–1802
41. So S-S, Karplus M (2001) J Comput Aided Mol Des 15:613–647
42. Lyne PD (2002) Drug Discov Today 7:1047–1055
43. Wang J, Kollman PA, Kuntz ID (1999) Proteins: Struct Funct Genet 36:1–19
44. Miteva MA, Lee WH, Montes MO, Villoutreix BO (2005) J Med Chem 48:6012–6022
45. Leroux V, Maigret B (2007) Comput Appl Chem 24:1–10
46. Yamagishi MEB, Martins NF, Neshich G, Cai W, Shao X, Beautrait A, Maigret B (2006) J Mol Model 12:965–972
47. Singh J, Chuaqui CE, Boriack-Sjodin PA, Lee WC, Pontz T, Corbley MJ, Cheung H-K, Arduini RM, Mead JN, Newman MN, Papadatos JL, Bowes S, Josiah S, Ling LE (2003) Bioorg Med Chem Lett 13:4355–4359
48. Ritchie DW, Kemp GJL (1999) J Comput Chem 20:383–395
49. Jones G, Willett P, Glen RC (1995) J Mol Biol 245:43–43
50. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727–748
51. Lala DS (2005) Curr Opin Investig Drugs 6:934–943
52. Collins JL (2004) Curr Opin Drug Discov Dev 7:692–702
53. Färnegårdh M, Bonn T, Sun S, Ljunggren J, Ahola H, Wilhelmsson A, Gustafsson J-Å, Carlquist M (2003) J Biol Chem 278:38821–38828
54. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
55. Williams S, Bledsoe RK, Collins JL, Boggs S, Lambert MH, Miller AB, Moore J, McKee DD, Moore L, Nichols J, Parks D, Watson M, Wisely B, Willson TM (2003) J Biol Chem 278:27138–27143
56. Steiner T, Koellner G (1997) Chem Commun (Cambridge, UK) 13:1207–1208
57. ChemDiv - The chemistry of cures. http://www.chemdiv.com
58. Enamine - Smart chemistry solutions. http://www.enamine.net
59. Albany Molecular Research - AMRIDirect chemical compound database. http://www.amridirect.com
60. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Delivery Rev 23:3–25
61. Monge A, Arrault A, Marot C, Morin-Allory L (2006) Mol Divers 10:389–403
62. Xue L, Godden J, Bajorath J (1999) J Chem Inf Comput Sci 39:881–886

63. Tanimoto TT (1961) Trans NY Acad Sci 2:576–580
64. Hibert M, Haiech J (2000) M S Méd Sci 16:1332–1339
65. Chimiothèque Nationale. http://chimiotheque-nationale.enscm.fr/
66. GOLD CCDC/Astex validation test set results. http://www.ccdc.cam.ac.uk/products/life_sciences/validate/gold_validation/
67. Koshland D Jr (1994) Angew Chem, Int Ed Engl 33:2375–2378
68. Spearman C (1904) Am J Psychol 15:72–101
69. Kendall M (1938) Biometrika 30:81–89
70. Mavridis L, Hudson BD, Ritchie DW (2007) J Chem Inf Model 47:1787–1796
71. Massova I, Kollman PA (2000) Perspect Drug Discov Des 18:113–135
72. Gilson MK, Zhou H-X (2007) Annu Rev Biophys Biomol Struct 36:21–42
73. Marcou G, Rognan D (2007) J Chem Inf Model 47:195–207