J Mol Model (2008) 14:393–401
DOI 10.1007/s00894-008-0286-z

ORIGINAL PAPER

SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces

Wensheng Cai · Jiawei Xu · Xueguang Shao · Vincent Leroux · Alexandre Beautrait · Bernard Maigret

Received: 19 October 2007 / Accepted: 6 February 2008 / Published online: 11 March 2008

Abstract SHEF (spherical harmonic coefficient filter), a geometrical matching procedure constituting a preliminary step in the virtual high throughput screening of large databases of small drug-like molecules, is demonstrated. This filter uses a description of both the binding site of the target and the ligand surfaces using spherical harmonic polynomial expansions. Using this representation, which is based on limited sets of spherical harmonic coefficients, considerably reduces the complexity of surface complementarity calculation. As a first test, 188 known protein–ligand complexes were used, and the results of docking the abstracted ligands into the bare proteins using SHEF were compared to the original X-ray structures. The ability of SHEF to retrieve known ligands "hidden" in a virtual library of 1,000 randomly selected drug-like compounds is also demonstrated.

Keywords Protein-ligand interactions · Virtual screening · Spherical harmonic expansions · Molecular surfaces · Surface complementarity and similarity · Drug discovery · SHEF

W. Cai (✉) · X. Shao
Department of Chemistry, Nankai University,
Tianjin 300071, People's Republic of China
e-mail: wscai@nankai.edu.cn

J. Xu
Department of Chemistry,
University of Science & Technology of China,
Hefei, Anhui 230026, People's Republic of China

V. Leroux · A. Beautrait · B. Maigret
LORIA, Groupe ORPAILLEUR, Campus scientifique,
Nancy Université,
BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France

Introduction

Recent progress in high-throughput screening (HTS) and combinatorial chemistry has greatly improved the hit-rate and cost-effectiveness of drug discovery campaigns, and has radically changed the chemist's approach to drug design. Virtual high-throughput screening (vHTS) using computers is gaining use in drug discovery as a complementary approach to experimental techniques [1–4].

Associated with vHTS strategies, numerous docking algorithms have been reported in the literature, and their merits have been summarized in several reviews [5–9]. These algorithms use more or less accurate physico-chemical representations of both receptor and ligand structures. These are associated with scoring functions [10] (necessarily approximate) to measure docking efficiency. The docking processes are finally driven by search strategies that, due to the complexity of the problem, are usually not exhaustive. These "classical" docking methods can give good results in the hit discovery context [11], but the time (and cost) of computation is too great to screen millions of compounds.

Preliminary crude but fast filters are thus required in large vHTS campaigns in order to reduce the number of candidate molecules to be passed to more elaborate docking calculations. In this context, several filtering methods, using for example shape [12] or fingerprint [13] signatures, have already been proposed. The main goal of such approaches is to overcome the time bottleneck of accurate docking methods in structure-based drug design strategies [14].

The spherical harmonics shape descriptor was originally proposed and further applied by Ritchie et al. [15–17]. Another recent application of spherical harmonics has been reported by Kahraman et al. [18], who used this shape descriptor to compare the shapes of protein binding pockets and that of their ligands. Here, we describe the SHEF filtering method (spherical harmonic coefficient filter) whose aim is to fulfill the fast docking objective. The core of the SHEF method is the generation of a set of spherical harmonic coefficients that convert 3D surface information into a 1D coefficient vector. A scoring function that uses only these coefficients to compare surfaces is then used. In order to test the effectiveness of the method, the following experiments were carried out: (1) SHEF was applied to a test system consisting of 188 protein–ligand complexes selected from the PDB database [19]; (2) SHEF was tested for its capabilities to retrieve known active ligands hidden in a database of randomly selected compounds; (3) SHEF computational costs were evaluated and compared with those of another vHTS method.

Methods

Representations of molecular surfaces using spherical harmonic expansions

Spherical harmonics (SH) are single-valued, continuous bounded, complex functions of the spherical coordinates (θ, ϕ), which can be considered as "standing waves on a sphere". They are characterized by two "quantum numbers", l and m, which together determine the number and spatial arrangement of nodes in each function. SH functions [20–23] are evaluated using Eq. 1,

$$Y_l^m(\theta,\phi)=\sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m(\cos\theta)e^{im\phi} \quad (1)$$

where l and m are integers (with $-l\leq m\leq l$), and $P_l^m(\cos\theta)$ the associated Legendre functions, which form a complete orthonormal basis set.

For a given protein–ligand complex, the molecular surfaces of a ligand and of the cavity in its binding site region can be modeled by our deflation and inflation techniques [24, 25]. Any single-valued three dimensional surfaces can be approximated by encoding the radial distance of surface points from the origin as a sum of SH functions as follows:

$$r(\theta,\phi)=\sum_{l=0}^{L}\sum_{m=-l}^{l}C_{lm}Y_l^m(\theta,\phi) \quad (2)$$

In this equation, $r(\theta,\phi)$ is the distance function of surface points from the origin inside. C_{lm} is the expansion coefficient of SH arranged by l and m ($0\leq l\leq L$; $-l\leq m\leq l$). L is the order that determines the degree of accuracy of the representation.

Therefore, the C_{lm} set of coefficients, considered as "surface descriptors", can completely define and represent the 3D surface shape, as approximated by SH expansions. It is possible to attain any degree of accuracy by adjusting the expansion order of the coefficients. Thus, any 3D surface shape can be converted into a 1D vector and, consequently, the comparison of different 3D molecular surfaces can be achieved by matching their corresponding 1D coefficient vectors.

Surface comparison using the expansion coefficients

Representation of the molecular surfaces of a target binding site and ligand by their expansion coefficients allows a shape comparison between the two surfaces to be achieved. For this purpose, considering the surface of the target as rigid and fixed, the coefficients of the ligand molecule are rotated in order to obtain the minimal root-mean-square distance (RMSD) of these coefficients and those of the target. The rotation matrix used for this purpose has been described by Ritchie and Kemp [15, 16].

The difference, D, of coefficients [15] is applied to measure the shape similarity in this study. If vectors $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{B}'$ are the SH surface representations of the target receptor active site, the ligand and the rotated ligand, respectively (those vectors having $(L+1)^2$ SH coefficients), and $\mathbf{A}_1$, $\mathbf{B}_1$ and $\mathbf{B}'_1$ the centroid vectors restricted to $l=1$ (thus possessing three coefficients, representing the surfaces' average orientation in the Cartesian space), then:

$$D(\mathbf{A},\mathbf{B},\mathbf{B}')=\sqrt{f(\mathbf{A},\mathbf{B})-2g(\mathbf{A},\mathbf{B},\mathbf{B}')} \quad (3)$$

with:

$$f(\mathbf{A},\mathbf{B})=\|\mathbf{A}\|^2+\|\mathbf{B}\|^2+\tfrac{1}{3}\|\mathbf{A}_1-\mathbf{B}_1\|^2$$
$$g(\mathbf{A},\mathbf{B},\mathbf{B}')=\mathbf{A}\bullet\mathbf{B}'+\tfrac{1}{3}(\mathbf{A}_1-\mathbf{B}_1)\bullet\left(\mathbf{A}_1-\mathbf{B}'_1\right)$$

A global optimization of three Euler angles to rotate the coefficients of the ligand is needed. Obviously, a systematic search through a finer angular grid with, for instance, an interval of 1°, is time-consuming, whereas a search through only a coarse angular grid probably misses some important regions. In order to quickly find those potential regions and further converge optimization to the corresponding local minima, a three-step optimization strategy was designed to minimize D (hence maximizing g) for one screening process. In the first step, a grid exploration is performed, using the Euler angles to rotate the coefficients of the molecular surface by regular increments (30° was used here). In the second step, for each of the best 10 orientations found previously, its 27 neighbors (each of the three Euler angles kept or varied by ± 10°) are also calculated, and the new best 10 solutions are selected from the total of 270 orientations. In the last step, the local minimizer L-BFGS [26] is applied to optimize these 10 orientations. From this procedure, an optimal set of Euler angles giving the best similarity score to the pair surfaces can finally be obtained. The final coefficient string related to the molecule surface obtained this way can be used to evaluate its similarity to the target surface.

Construction of the filter

In this method, for a given target binding site surface, the flexibility of the ligand partners can be modeled by considering different conformers as separate docking candidates. The first step of our procedure is therefore to generate a set of low energy conformers for each ligand and the calculation of the associated SH surfaces. The resulting coefficients constitute our ligand-coefficients database. An analogous target-coefficients database can also be generated; this can encompass several active site conformations obtained either from diverse experimental structures or by conformational sampling methods (molecular dynamics or Monte Carlo simulations).

Note that the databases (both ligands and proteins) are reusable for further applications; the calculation of SH coefficients has to be done only once. Moreover, new molecules and receptors can easily be added. The size of such databases is reasonable compared to storing molecular structures as atomic coordinates and bond information records. More interestingly, provided the same expansion order is used for all the conformers in the database, the size of each record is constant. This allows implementation of efficient schemes for storing and accessing data.

Matching the surface of each candidate conformer to a given target is realized by minimizing the difference function between two coefficient sets as stated in Eq. 3. Another scoring function is then used to evaluate the optimized pose. If $\mathbf{A}_{\bar{0}}$ and $\mathbf{B}'_{\bar{0}}$ are the $\mathbf{A}$ and $\mathbf{B}'$ vectors minus the first coefficient (for which $l=m=0$, representing the average radius of the SH expansion volume), then:

$$Score(\mathbf{A},\mathbf{B}') = \frac{\|\mathbf{A}-\mathbf{B}'\|}{\|\mathbf{B}'\|} + w\left(1-\cos\left(\mathbf{A}_{\bar{0}},\mathbf{B}'_{\bar{0}}\right)\right) \tag{4}$$

with:

$$w = \frac{\max(\|\mathbf{A}\|,\|\mathbf{B}'\|)}{\min(\|\mathbf{A}\|,\|\mathbf{B}'\|)} \qquad \cos\left(\mathbf{A}_{\bar{0}},\mathbf{B}'_{\bar{0}}\right) = \frac{\mathbf{A}_{\bar{0}}\bullet\mathbf{B}'_{\bar{0}}}{\|\mathbf{A}_{\bar{0}}\|\left\|\mathbf{B}'_{\bar{0}}\right\|}$$

Similar to D in Eq. 3, the first term in Eq. 4 is also used to evaluate the difference of coefficients. It should be noted that, as the radial coefficient ($l=m=0$) is usually much larger than the others, the first term in Eq. 4 is most sensitive to the size matching between the two surfaces described by the $\mathbf{A}$ and $\mathbf{B}'$ vectors. In the case of only docking a rigid molecule to a protein binding cavity, D is enough to distinguish different docking poses. However, when considering the flexibility of the molecule or screening a group of molecules, different conformers or molecules are needed to be compared after docking. The case where the size is good but the shape is bad may occur, which may be assigned a very low value evaluated by the first term. In order to select the candidate with the best match in size and shape, the second term is added in Eq. 4, which mainly delineates shape similarity.

The *Score* value is the criterion used for screening using SHEF. For each ligand molecule, the conformer providing the lowest score is retained, and the relative effectivness of the ligands are compared using these values in our vHTS procedure. The whole virtual screening process is shown in Fig. 1.

Docking protocol

A test set of 188 protein–ligand complexes has been chosen from the PDB database on the basis of their diversity and non-redundancy. For each complex, the ligand was detached from the protein active site and redocked using SHEF. The goal here is to compare the poses obtained after the SHEF coefficient optimization procedure and those from the X-ray structures of the corresponding complexes.

The calculation time and precision obviously depend on the value of the SH expansion order L. In order to measure this behavior, different values of L were used for the docking calculations. Since, for surface matching, large L (>10) is not necessary [15], values from 3 to 10 were tested.

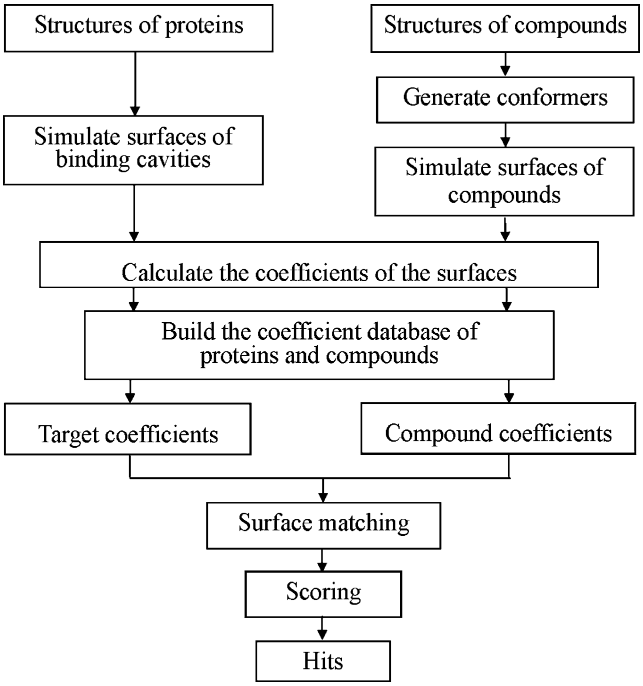
Structures of proteins
Structures of compounds
Generate conformers
Simulate surfaces of binding cavities
Simulate surfaces of compounds
Calculate the coefficients of the surfaces
Build the coefficient database of proteins and compounds
Target coefficients
Compound coefficients
Surface matching
Scoring
Hits

Fig. 1 The flowchart to build a SHEF (spherical harmonic coefficient filter) filter in virtual high-throughput screening (vHTS)

Input data for evaluating filtering efficiency in the virtual screening context

Another important test is to check how good SHEF is as a screening filter. The significance and efficiency of a filter depend on how effectively it can sort out suitable compounds from the input database. An efficient filter is therefore one that can reduce the database to a manageable size for subsequent, more precise, experimental measurements and/or structure-based drug discovery techniques. Methods used for this purpose have to be able to select the most probable inhibitors from randomly chosen drug-like molecules.

A random database with 1,000 compounds, randomly selected from 10,000 drug-like compounds in the NCI 3D database [27, 28], was constructed. The average number of atoms per compound is about 32, and the average molecular weight is 237.6. The conformational flexibility of molecules was considered by storing multiple conformers for each molecule. The corresponding structures were first generated using OMEGA [29], giving an average of 34 conformers per molecule. Next the SH expansion coefficients of each conformer were calculated (L=5, giving a vector of 36 coefficients) and stored in the ligand-coefficients database. These data were used as the decoys in the ligands database.

Metrics for measuring filtering performance

Given a test database composed of n structures, divided into active a molecules (with known activity for the reference target) and decoy d random molecules (with presumably no affinity for the target), screening also divides n into two groups: those predicted to be active (h hits) and those that are filtered out (f). Using a virtual screening program such as SHEF to rank molecules using a scoring function for evaluating affinity, the h value is a parameter set by the user. Screening performance is related to the number of retrieved actives, h_a, and inversely related to the number of false positives, f_+, and false negatives, f_-. These definitions are summarized in Fig. 2a.

From this, a number of metrics for evaluating virtual screening performance (bound from 0 to 1, and that can be expressed as percentages) can be formulated. The filtering amount F (taken as the screening parameter), the coverage C, the yield of actives Y, the efficiency E and the Güner-Henry score GH [30] are defined as:

$$\begin{aligned}
&F = \left(1 - \tfrac{h}{n}\right) \times 100\% \\
&C(F) = \tfrac{h_a}{a} \times 100\% \\
&Y(F) = \tfrac{h_a}{h} \times 100\% \\
&E(F) = \tfrac{Y(F)}{Y(0)} = \tfrac{nh_a}{ah} \times 100\% \\
&GH(F) = wY(F) + (1 - w)C(F)
\end{aligned} \tag{5}$$

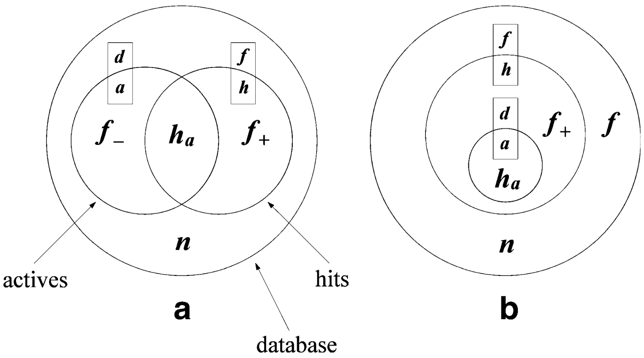

Fig. 2a,**b** Definition of molecular database sub-groups. The main circle represents the whole database (containing n structures), while the two inner circles represent the actives molecules (containing a molecules) and the hits as defined by the screening program (containing h molecules). The number of random molecules d is equal to n−a. The number of the filtered-out molecules f is equal to n−h. The variants h_a, f_+ , and f_- denote the number of retrieved actives, the number of false positives and false negatives, respectively. Hence we have f_+= h−h_a and f_-=a−h_a. **a** General case. **b** Full coverage of the actives by the screening, where f_-=0

In order to have a single value for a given method, only the filtering amount F^*, the maximum value giving full coverage (h_a=a; f_-=0) ,was computed in our tests:

$$F^{*} = \max\left(F/C(F) = 1\right) \tag{6}$$

This particular case is represented in Fig. 2b. In order to better express the screening accuracy, w is set to 0.75, and the GH-score is weighted using the ratio of false positives

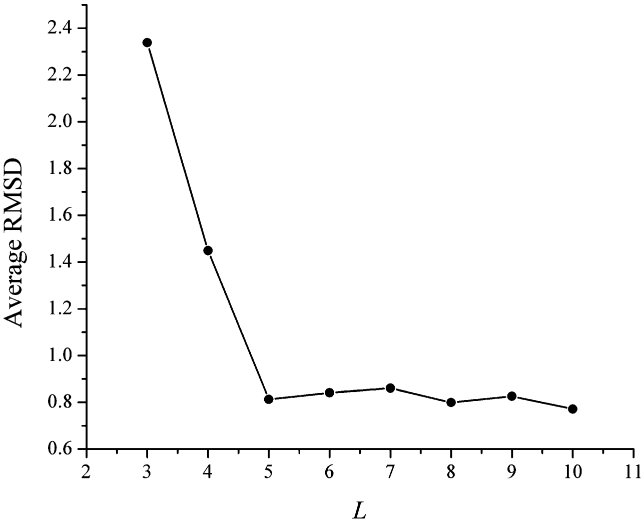

Fig. 3 The average root-mean-square distance (RMSD) between the original crystal structures and the docking results over 188 complexes with different values of the spherical harmonic expansion order L

Table 1 Docking results of 188 complexes: root-mean-square distance (RMSD) values between the ligands from experimental structures and SHEF (spherical harmonic coefficient filter) docking predictions

No	PDB	RMSD	No	PDB	RMSD	No	PDB	RMSD	No	PDB	RMSD
1	1A8G	0.510	48	1DRF	0.140	95	1LNA	0.469	142	1TNH	1.308
2	1A9 M	0.285	49	1DWB	2.034	96	1LST	0.355	143	1TNI	0.338
3	1ABE	1.907	50	1DWD	0.611	97	1LYB	0.885	144	1TNJ	0.754
4	1ABF	0.671	51	1ELA	0.705	98	1MBI	0.320	145	1TNK	1.957
5	1ACJ	0.677	52	1EPO	0.652	99	1MCF	0.688	146	1TNL	1.213
6	1ACM	1.921	53	1EPP	0.807	100	1MCH	0.521	147	1TPH	0.463
7	1ADD	0.333	54	1ETR	0.779	101	1MFC	1.405	148	1TPP	1.469
8	1AHA	0.276	55	1ETS	0.842	102	1MMB	0.411	149	1UKZ	1.041
9	1AJV	0.412	56	1ETT	0.487	103	1MRK	0.912	150	1ULB	0.354
10	1AJX	0.153	57	1FKG	0.161	104	1MUP	0.413	151	1WAP	0.267
11	1APT	0.704	58	1FLR	0.835	105	1NCO	2.072	152	2ACK	0.400
12	1APU	0.788	59	1GHB	0.624	106	1NRR	0.394	153	2ADA	0.236
13	1APV	0.493	60	1GPY	2.725	107	1ODW	0.450	154	2AK3	0.689
14	1APW	0.559	61	1H8D	0.810	108	1ODX	0.539	155	2CGR	1.246
15	1ATL	0.543	62	1HBV	0.330	109	1OKM	0.604	156	2CHT	2.629
16	1B5G	0.506	63	1HDC	0.646	110	1OS0	0.383	157	2CPP	0.344
17	1BAP	1.913	64	1HDT	1.052	111	1PE8	0.194	158	2CTC	0.410
18	1BBP	0.381	65	1HEW	0.312	112	1PHF	0.198	159	2DBE[a]	8.768
19	1BNM	1.061	66	1HIH	0.105	113	1PHG	0.404	160	2GBP	0.462
20	1BNN	0.658	67	1HIV	0.293	114	1POC	0.452	161	2IFB	0.991
21	1BNQ	1.228	68	1HOS	0.199	115	1PPB	0.316	162	2R04	0.417
22	1BNT	0.745	69	1HPS	0.443	116	1PPC	0.543	163	2R07	0.527
23	1BNU	0.350	70	1HPV	0.145	117	1PPK	0.562	164	2TMN	0.797
24	1BNV	0.569	71	1HRI	0.519	118	1PPL	0.920	165	2TSC	0.727
25	1BRA	2.387	72	1HSL	2.369	119	1QBR	0.143	166	3ER3	1.341
26	1BYB	0.799	73	1HTF	0.907	120	1QBT	0.144	167	3HVT	0.445
27	1BYG	1.602	74	1HTG	0.548	121	1QBU	0.198	168	3PTB	0.595
28	1C2T	0.370	75	1HVI	0.212	122	1QF0	0.252	169	3TMN	0.480
29	1C83	0.281	76	1HVJ	0.134	123	1QF1	0.056	170	3TPI	0.467
30	1CBS	0.832	77	1HVK	0.138	124	1QF2	0.585	171	4AAH	0.508
31	1CBX	0.568	78	1HVL	0.074	125	1RGK	0.629	172	4DFR	0.771
32	1CIM	1.330	79	1HVR[a]	10.031	126	1RGL	0.709	173	4HVP	0.275
33	1COM	2.875	80	1HXB	0.412	127	1RJK	0.705	174	4PHV	0.443
34	1COY	0.664	81	1HYT	0.891	128	1RK3	0.532	175	4TMN	0.269
35	1CPS	0.918	82	1ICN	0.559	129	1RKG	0.237	176	5ER2	1.657
36	1CTT	1.728	83	1IDA	0.115	130	1RKH	0.632	177	5HVP	0.128
37	1D3H	0.255	84	1IE8	0.588	131	1RNE	0.652	178	5P21	0.602
38	1D4P	0.915	85	1IE9	0.500	132	1ROB	0.381	179	5TLN	1.030
39	1DB1	0.514	86	1IGJ	1.046	133	1S0Z	0.743	180	6ABP	0.836
40	1DBJ	0.611	87	1INC	0.748	134	1S19	0.579	181	7CPA	0.224
41	1DBK	0.572	88	1JAP	0.875	135	1SNC	0.462	182	7HVP	0.414
42	1DBM	0.61	89	1KEL	0.695	136	1STP	2.210	183	7LPR	1.251
43	1DD7	0.354	90	1KR6	0.420	137	1THL	0.245	184	7TIM	0.527
44	1DID	0.584	91	1KS7	0.860	138	1TKA	1.095	185	8ATC	2.502
45	1DIE	2.699	92	1LAH	0.308	139	1TLP	0.507	186	8CPA	0.492
46	1DIF	0.130	93	1LDM	1.646	140	1TMN	0.564	187	8GCH	0.976
47	1DMP	0.077	94	1LIC	1.242	141	1TNG	0.995	188	9HVP	0.159

The average RMSD over the 188 complexes is 0.813 Å

[a] The 1HVR and 2DBE ligands have symmetrical structures and were" flipped" upon docking, hence the large RMSD values, which thus do not correspond to a poor prediction

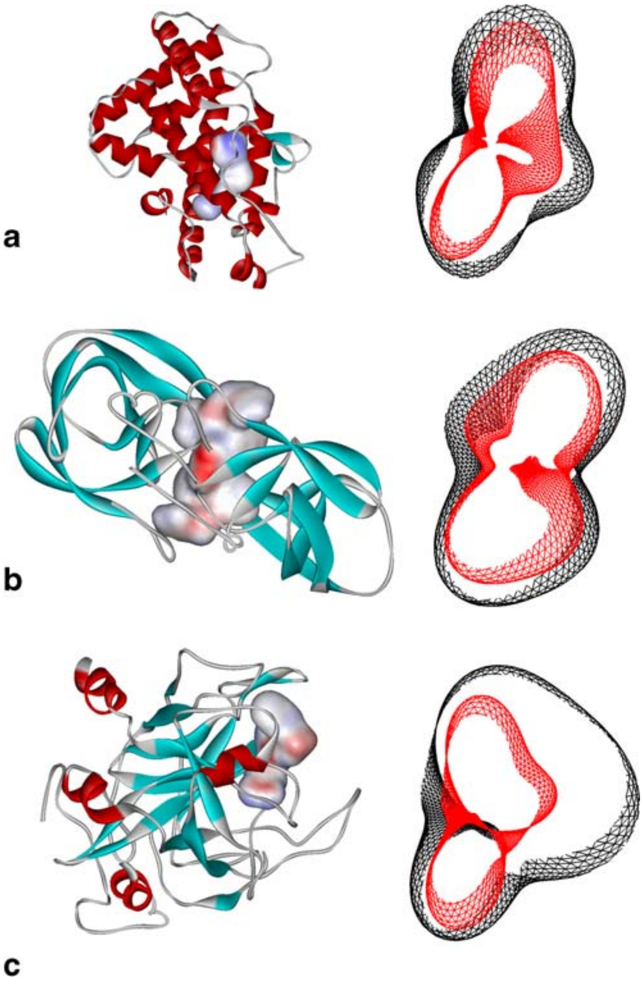

Fig. 4 **a**–**c** The crystal structures of the three complexes (*left*), and their interface section figures generated by their optimized coefficients with L=5 (*right*: *red* ligand, *black* cavity). **a** 1IE9, **b** 1HVK, **c** 1ETS

on decoys. Finally GH^*, a value derived from the GH-score expressing the screening efficiency is obtained:

$$GH^* = \frac{3Y\left(F^*\right)+1}{4}\left(1-\frac{f_+\left(F^*\right)}{d}\right) \tag{7}$$

Results and discussion

Rigid docking test of 188 complexes

After performing SHEF for the 188 ligands and their corresponding binding sites, the atomic coordinates corresponding to the obtained minimal D value for each complex (see Eq. 3) were compared to the original X-ray structure. Figure 3 shows the relationship between the docking results and the order L. It can be seen that the average RMSD between the experimental ligand-bound conformation and the docking results for 188 complexes decreases rapidly until L is equal to 5, and then changes slightly when L is between 5 and 10. Consequently, a value of L=5 is recommended and was used in the docking tests presented here.

Docking results for L=5 are given in Table 1. The RMSDs of all but 2 of the 188 entries are smaller than 3.0 Å, giving an average RMSD of 0.813 Å. Because of a symmetry problem, two complexes, namely 1HVR and 2DBE, have much larger RMSDs (10.031 Å and 8.768 Å for 1HVR and 2DBE, respectively), although the two SHEF poses in fact fit quite well with the X-ray data. The flipped orientations for the ligands in 1HVR and 2DBE give the best results when using the scoring function in Eq. 3. Both unflipped orientations of the two ligands are also found in the optimization results as the second-best solutions, giving scores very close to the best scores. The corresponding RMSD values for 1HVR and 2DBE are calculated to be 0.015 Å and 0.210 Å.

It should be noted that the optimized value of D for each complex reflects the degree of solvent exposure of the corresponding binding cavities. Indeed, small D means that the ligand is entirely embedded in a relatively closed binding cavity, whereas larger values indicate that the protein holds a more open binding cavity, so that a limited match exists between the ligand and the binding site. Most of the complexes in our test set have relatively closed or partly opened cavities and therefore exhibit good complementarities. As an example, the crystal structures of three complexes, namely 1IE9, 1HVK and 1ETS, and their

Table 2 PDB codes of the structures comprising the two protein families used in the filtering test

Protein family	Known ligands/active compounds	PDB ID of complexes in the family
I: Vitamin D receptor complexes	9	1DB1 1IE8 1IE9 1RJK 1RK3 1RKG 1RKH 1S0Z 1S19
II: HIV-1 protease complexes	30	1A8G 1A9M 1AJV 1AJX 1DIF 1DMP 1HBV 1HIH 1HIV 1HOS 1HPS 1HPV 1HTF 1HTG 1HVI 1HVJ 1HVK 1HVL 1HVR 1HXB 1ODW 1ODX 1QBR 1QBT 1QBU 4HVP 4PHV 5HVP 7HVP 9HVP

In families I and II, there are 9 and 30 complexes, respectively. Each complex contains one known ligand or active compound

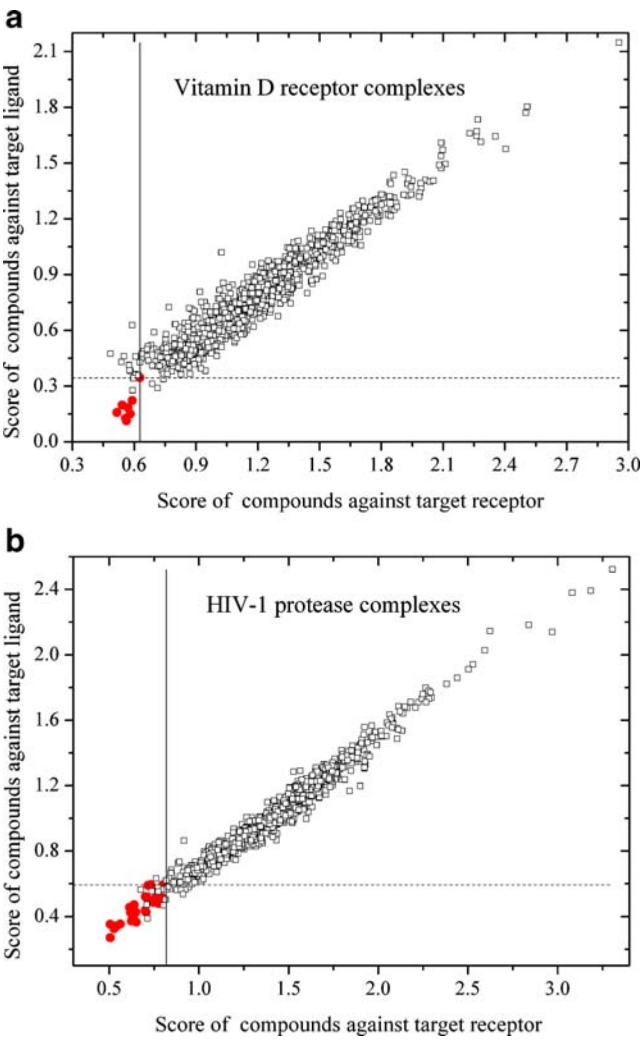

Fig. 5a,**b** SHEF filtering results against target receptor and target ligand, respectively. In each filtering test, all active compounds must be retrieved, corresponding to the case in Fig. 2b. Known ligands (actives) are represented as *solid circles*, while *open squares* denote the decoys. *Solid* and *dashed lines*: largest (worst) score amongst the actives against the reference binding site and its ligand, respectively. SHEF hits are left to the solid line. **a** Vitamin D receptor complexes (target and reference ligand from PDB structure 1IE9). **b** HIV-1 protease complexes (target and reference ligand from PDB structure 1HVK)

interface sections generated by their optimized coefficients are shown in Fig. 4.

Filtering performance in virtual screening

To perform the screening test, two representative groups of protein–ligand complexes were selected from Table 1 according to their binding cavity characteristics. They possess closed and half-closed cavities and are classified into two distinct protein families (Table 2): vitamin D (9 complexes) and HIV-1 protease receptors (30 complexes). The corresponding ligands are considered as actives in the filtering process, and are merged into two composite databases based on the 1,000 random drug-like decoys. The vitamin D and HIV-1 databases have 1,009 and 1,030 compounds, with 33,765 and 38,015 conformers, respectively. In order to assess the robustness of the method, the X-ray conformations of the known ligands were removed from the database in our test experiments, leaving only OMEGA-generated conformers.

The filtering results are shown in Fig. 5. The X-axis denotes the optimized *Score* (Eq. 4) of the compounds in the composite database against the target cavity, and the solid line indicates the corresponding cutoff score (the largest score among all active compounds) used to filter the docking poses in order to retrieve all active compounds (C=1; F=F^*). The Y-axis denotes the corresponding *Score* of the compounds against the target ligand, and the dashed cutoff line is used to recover all known ligands. The lower left rectangle formed by these two cutoff lines and two axes recovers all active compounds when screening against both target receptor and target ligand.

It can be seen from Fig. 5 that the distribution of most of the points in each figure is mostly linear. It means that the more complementary a candidate molecule is to the receptor target, the more similar this molecule is also to the reference ligand. It can also be seen that the distribution of the points in Fig. 5b is better than that in Fig. 5a. This is due to the

Table 3 Effectiveness of SHEF and FRED measured after maximum filtering upon total coverage (all actives recovered), Metrics are calculated by Eqs. 5, 6, 7 and expressed as percentages

Vitamin-D receptor complexes					
a=9; d=1000	F^* (%)	h	Y (%)	E (%)	GH^* (%)
SHEF	97.9	21	42.9	48.0	56.5
FRED	92.0	81	11.1	12.5	30.9
HIV-1 protease complexes					
a=30; d=1000	F^* (%)	h	Y (%)	E (%)	GH^* (%)
SHEF	96.0	41	73.2	25.1	79.0
FRED	95.2	49	61.2	21.0	69.6

a Number of known ligands, d number of random molecules in the database, h number of hits, F^* maximum filtering amount, Y yield of actives, E enrichment efficiency. The important measurement to indicate the effectiveness of the filtering is the Güner-Henry score (GH^*), which suggests that the performance of SHEF is superior to that of FRED higher sensitivity of the SH procedure to complicated shapes presenting several clear lobes and holes in the ligands and the receptor binding site, which are clearer in HIV-1 protease complexes than in the Vitamin-D complexes.

The effectiveness of the filtering was measured using the GH^* value (see Eq. 7); results are shown in Table 3. The corresponding a, d, F^*, $h(F^*)$, $Y(F^*)$ and $E(F^*)$ values (Eqs. 5, 6) are also shown. In order to compare SHEF results with those of a classical rigid docking method recognized for performing exhaustive and fast calculations, virtual screening on the reference dataset was also done using FRED [29]. FRED, a well-known rigid docking algorithm based on shape and chemistry, is considered an effective and very fast filtering method; therefore, it is an appropriate choice of comparison program for screening effectiveness in this study. The results clearly indicate that SHEF is superior to FRED regarding filtering performance.

CPU time used for computing the coefficients and screening the coefficient database

In SHEF, the total computational time comprised two components: the CPU time required to calculate the coefficients to build the ligand- and target's pocket-coefficient databases, and the CPU time required for the screening itself. The average CPU time on a computer composed of a AMD MP2200 + processor with 1 Gb memory (with a computing speed comparable to CPUs currently at the lower-performance end of PC desktops) required to calculate the coefficients with $L=5$ for one ligand conformer (with 32 atoms) is about 1 s. For one protein cavity (with 350 wall atoms) about 20 s are required. Both these calculations need to be done only once.

The filter will then work from the pre-constructed coefficient databases without any atomic coordinate information. Using SHEF, the average screening time for one conformer is 0.046 s, which is about 2.4 times faster than FRED on the same computers. The average time to screen a compound (i.e., 34 conformers) is 1.564 s, which is much faster than the new technique proposed by Putta et al. [31].

Conclusions

An efficient filter SHEF for vHTS using the SH coefficients of molecular surfaces has been presented. Both the rigid docking and filtering performance tests of this method gave satisfactory results. The accuracy of the flexible docking depends on the pre-generated conformers in the database. The aim is to eliminate most of the compounds or conformers that do not fit to the target binding cavity, rather than to identify the best binders. More accurate docking calculations based on binding energy estimation should be applied to the selected ligands. SHEF is therefore a method that can be used as a potential fast and efficient filter prior to more efficient techniques in the vHTS context. As such, it confirms that techniques using purely geometrical representations of the active site and the candidate ligands can provide positive results [32].

In this paper, basic test experiments have been performed. In fact, we have implemented SHEF into an integrated package for vHTS, the VSM-G platform. The combined use of SHEF with a classical docking program using this software was validated as a relevant enrichment technique in large-scale virtual screening experiments [33]. Additionally, although SHEF focuses on geometrical complementarity, it could be extended to include chemical features so as to provide a more extensive measure of protein–ligand binding. Such an extension of SH molecular surfaces has already yielded good results for similarity-based ligand-based drug design approaches [17]. Work to expand SHEF in a similar fashion is in progress.

Acknowledgments This study is supported by National Natural Scientific Foundation of China (Nos. 20325517 and 20573102), and the Franco-Chinese PRA (Projet de Recherche Avancée) project (B02–04). 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 postdoc fellowship of INCa (Institut National du Cancer). We thank Openeye for providing free access to OMEGA, FRED and VIDA software according to an academic license

References

1. Walters WP, Stahl MT, Murcko MA (1998) Drug Discov Today 3 (4):160–178
2. Cavasotto CN, Orry AJ (2007) Curr Top Med Chem 7(10):1006–1014
3. Reddy AS, Pati SP, Kumar PP, Pradheep HN, Sastry GN (2007) Curr Protein Pept Sci 8(4):329–351
4. Seifert MHJ, Kraus J, Kramer B (2007) Curr Opin Drug Discov Dev 10:298–307
5. Brooijmans N, Kuntz ID (2003) Annu Rev Biophys Biomol Struct 32:335–373
6. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Proteins: Struct Funct Bioinf 57:225–242
7. Leach AR, Shoichet BK, Peishoff CE (2006) J Med Chem 49 (20):5851–5855
8. Sousa SF, Fernandes PA, Ramos MJ (2006) Proteins: Struct Funct Genet 65(1):15–26
9. Joseph-McCarthy D, Baber JC, Feyfant E, Thompson DC, Humblet C (2007) Curr Opin Drug Discov Dev 10:264–274
10. Jain AN (2006) Curr Protein Pept Sci 7(5):407–420
11. Abagyan R, Totrov M (2001) Curr Opin Chem Biol 5:375–382
12. Proschak E, Rupp M, Derksen S, Schneider G (2008) J Comput Chem 29(1):108–114
13. Deng Z, Chuaqui C, Singh J (2004) J Med Chem 47(2):337–344
14. Bleicher KH, Böhm H-J, Müller K, Alanine AI (2003) Nat Rev Drug Discov 2(5):369–378
15. Ritchie DW, Kemp GJL (1999) J Comput Chem 20(4):383–395

16. Ritchie DW, Kemp GJL (2000) Proteins: Struct Funct Genet 39 (2):178–194
17. Mavridis L, Hudson BD, Ritchie DW (2007) J Chem Inf Model 47(5):1787–1796
18. Kahraman A, Morris RJ, Laskowski RA, Thornton JM (2007) J Mol Biol 368(1):283–301
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28(1): 235–242
20. Leicester SE, Finney JL, Bywater RP (1998) J Mol Graph 6 (2):104–108
21. Max NL, Getzoff ED (1988) IEEE Comput Graph Appl 8(4):42–50
22. Duncan BS, Olson AJ (1993) Biopolymers 33(2):219–229
23. Barnett MP (2003) J Chem Inf Comput Sci 43(4):1158–1165
24. Cai W, Zhang M, Maigret B (1998) J Comput Chem 19(16): 1805–1815
25. Cai W, Shao X, Maigret B (2002) J Mol Graph Model 20(4):313–318
26. Liu DC, Nocedal J (1989) Math Program 45:503–528
27. Milne GW, Nicklaus MC, Driscoll JS, Wang S, Zaharevitz D (1994) J Chem Inf Comput Sci 34(5):1219–1224
28. Voigt JH, Bienfait B, Wang S, Nicklaus MC (2001) J Chem Inf Comput Sci 41(3):702–712
29. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Biopolymers 68(1):76–90, OpenEye Science Software: Santa Fe, NM http://www.eyesopen.com/docs/pdf/fred.PDF
30. Güner OF, Henry DR (2000) Pharmacophore perception, development, and use in drug design. IUL Biotechnology Series, La Jolla, pp 195–211
31. Putta S, Lemmen C, Beroza P, Greene J (2002) J Chem Inf Comput Sci 42(5):1230–1240
32. Jiang F, Kim S (1991) J Mol Biol 219(1):79–102
33. Beautrait A, Leroux V, Chavent M, Ghemtio L, Devignes M-D, Smaïl-Tabbone M, Cai W, Shao X, Moreau G, Bladon P, Yao J, Maigret B (2008) J Mol Model 14(2):135–148