J Mol Model (2010) 16:1223–1230
DOI 10.1007/s00894-009-0633-8

ORIGINAL PAPER

Improving performance of docking-based virtual screening by structural filtration

Fedor N. Novikov · Viktor S. Stroylov · Oleg V. Stroganov · Ghermes G. Chilov

Received: 6 August 2009 / Accepted: 16 November 2009 / Published online: 30 December 2009

Abstract In the current study an innovative method of structural filtration of docked ligand poses is introduced and applied to improve the virtual screening results. The structural filter is defined by a protein-specific set of interactions that are a) structurally conserved in available structures of a particular protein with its bound ligands, and b) that can be viewed as playing the crucial role in protein-ligand binding. The concept was evaluated on a set of 10 diverse proteins, for which the corresponding structural filters were developed and applied to the results of virtual screening obtained with the Lead Finder software. The application of structural filtration resulted in a considerable improvement of the enrichment factor ranging from several folds to hundreds folds depending on the protein target. It appeared that the structural filtration had effectively repaired the deficiencies of the scoring functions that used to overestimate decoy binding, resulting into a considerably lower false positive rate. In addition, the structural filters were also effective in dealing with some deficiencies of the protein structure models that would lead to false negative predictions otherwise. The ability of structural filtration to recover relatively small but specifically bound molecules creates promises for the application of this technology in the fragment-based drug discovery.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-009-0633-8) contains supplementary material, which is available to authorized users.

F. N. Novikov · V. S. Stroylov · O. V. Stroganov · G. G. Chilov
MolTech Ltd,
119992 Moscow Leninskie gory, 1/75A,
Russia

O. V. Stroganov · G. G. Chilov (✉)
N.D. Zelinsky Institute of Organic Chemistry,
119992 Moscow Leninsky pr., 47,
Russia
e-mail: Ghermes@moltech.ru

Keywords Docking · Focused library · Fragment-based · Structural filtration · Virtual screening

Introduction

The evaluation of protein-ligand interactions is a key technology in the drug discovery and virtual ligand screening is one of the practical methods, usually used in conjunction with various experimental techniques. Despite the concerns about the cost and accuracy of the currently available wet-lab techniques, virtual screening is still far from replacing the experimental methods [1]. The fundamental reason for that is the fact that no one can reliably predict the free energy of protein-ligand binding. From the practical point, the pursuit of better accuracy in in silico evaluation of protein-ligand binding has evolved into the emergence of two loosely bound branches—QSAR and molecular modeling, each having its own strengths and weaknesses. The ligand-based QSAR is traditionally viewed as a fast tool to produce chemically meaningful results, however the main disadvantages of this method are the lack of structural interpretation and its high dependence on the training data. The addition of structural information onto the QSAR basis (3D-QSAR) may be a promising direction, however there is no general recipe for doing that. For instance, a recently introduced 3D-QSAR technique performed well in virtual screening on some protein targets, but appeared to be worse than a random ligand selection on other targets [2]. On the contrary, molecular modeling techniques, such as docking, claim to reproduce the structure of a protein-ligand complex and its energy from the first principles. However, in the real world molecular modeling faces the theoretical chemistry problems of energy representation and the computational problems of

global optimization, so the output of the modeling techniques, in particular docking, fails to satisfy the researchers [3]. Moreover, one of the recent studies revealed that a simple substructure search was generally more successful in virtual screening compared to docking (given enough ligands were included in the training set) [4]. Interestingly, the results of substructure search and docking could hardly be coupled to yield a better technique, outlining the fundamental difficulties in linking the ligand-based and structure-based approaches [4].

However, the search for novel ideas improving the accuracy of docking-based virtual screening methods continues. For example, the application of a pharmacophore filter to the list of docked and rank-ordered ligands, in which the ligand was moved to the bottom of the list if its pose (produced by docking) did not meet the set of pre-defined criteria, improved the results [5]. Given the deficiencies of the current scoring functions, one may think of such kind of post-docking filtration as a useful and physically meaningful way to remedy scoring errors that worsen performance of the docking-based screening. A reasonably designed structural criterion may filter out irrelevant ligands that receive a high score due to the deficiencies of the scoring function. The main idea behind the structural filtration suggested in this work is the following: if a set of specific interactions (*e.g.*, H-bonds) is observed for all known complexes of a particular protein with its bound ligands, then we anticipate the same interactions must be present in a ligand recovered in virtual screening; if not, the ligand is discarded. This stated principle of the structural filtration differs from that of the work [5], in which the occurrence of particular types of ligand atoms in particular positions of the protein binding site, rather than the protein-ligand interactions, was monitored. Our approach also differs from the structural interaction fingerprint (SIFt) [6] and interaction fingerprint (IFP) [7] techniques. Both SIFt and IFP generate a bit string coding for all (typically several tenths of) protein-ligand interactions, and ligand filtration is performed by calculating Tanimoto distance (or other measure of string similarity) between two strings. We argue that counting for the small set of specific (rather than all possible) interactions gives a more realistic and clear way of performing structural filtration. It is worth mentioning that an approach very similar to structural filtration has been recently implemented in FlexX-Pharm docking with pharmacophore constraints [8]. However, despite the fact that the structural filters and pharmacophore constraints are introduced in a very similar way, FlexX-Pharm uses them immediately in the docking process to favor ligand poses that satisfy such constraints. In the current formulation we see structural filtration as a post-docking tool that aids the user in selecting ligands that satisfy some (knowledge-based, flexible) criteria, but does not bias the docking capability itself. In addition, the structural filtration can be combined with any type of docking software.

The current state of the theory does not allow us to provide an automatic tool to choose these crucial interactions for structural filtration, however some knowledge-based approaches exist for this purpose [9, 10]. In this work we demonstrate that by applying clear intuitive criteria the choice of crucial interactions for structural filtration can be made. The current formulation of structural filtration has been evaluated on the set of 10 diverse proteins: adrenaline receptor beta 2 (ADRB2), dihydrofolate reductase (DHFR), dihydroorotate dehydrogenase (DHODH), 3-hydroxymethylglutaryl-CoA reductase (HMG-CoAR), leukotrien A4 hydrolase (LTA4H), metabotropic glutamate receptor 3 (mGluR3), orotidine-monophosphate decarboxylase (OMPDC) from *E.coli*, peroxisome proliferator-activated receptor gamma (PPAR-g), phospholipase A2 (PLA2) and phosphotyrosine phosphatase 1B (PTP1B). The impact of the structural filtration on virtual screening performance is benchmarked using our Lead Finder software, which has recently demonstrated high accuracy in ligand docking and binding energy calculations [11].

Experimental section

Preparation of protein structure models

The full-atom protein models were prepared from the corresponding raw PDB structures (Table 1) by adding hydrogen atoms and assigning ionization states of the amino acids using the Model Builder module of the Lead Finder software package. The coordinates of heavy protein atoms were left unchanged. Each protein structure model was validated by docking a set of its ligands (available from the PDB) and comparing the docked and crystallographic ligand positions. More details on the model preparation and validation can be found in the Supporting Information.

Selection of active ligands

For each protein, the sets of active ligands (experimentally validated inhibitors, agonists or antagonists) that were used in the current benchmarking study, were extracted from either PDB [12] or KiBank [13]. The resulting ligands are provided in the Supporting Information.

Virtual screening

Virtual screening calculations were performed with the Lead Finder software v. 1.1.10 under its default configuration parameters. A reference ligand for mapping the ligand

Table 1 Structural filters used in virtual screening

Protein	PDB structures[a]	Specific interactions with protein[b]		Structural filter[c]
ADRB2	**2rh1**, 3d4s	I[d]	N312:ND2, N312:OD1 D113:OD1, D113:OD2	I
DHFR	1boz, 1dhf, 1dlr, **1dls**, 1hfp, 1kms, 1kmv, 1mvs, 1pd8, 1s3u, 1s3w, 2dhf, 1ohj, 1boz	I	E30:OE1, E30:OE2 V115:O I7:O	I
DHODH	1d3g, 1d3h, 2b0m, 2bxv, 2fpt, 2fpv, 2fpy, **2fqi**	I	H56:ND1 (d) Q47:OD1 R136:NE	I
HMG-CoAR	3bgl, 2r4f, 2q6c, 2q6b, 2q1l, **1hwl**, 1hwk, 1hwj, 1hwi, 1hw9, 1hw8	I II	S684:OG (d) K282:NZ K637:NZ D635:OD2 R535:NH2	(I AND II) OR (I AND III) OR (II AND III)
		III	R590:NH2 E559:OD2 K691:NZ N755:ND	
LTA4H	1gw6, **2vj8**, 3cho, 3chp, 3chq, 3chr, 3chs	I II	Zn^{2+} OR Zn^{2+} -H_2O[e] E271:OE1 E318:OE2 Q136:OE1	I AND (II OR III)
		III	K565:NZ R563:NE G268:N G269:O Y267:OH (d)	
mGluR3	2e4u, 2e4v, **2e4w**, 2e4x, 2e4y	I II	R68:NH2 K389:NZ S151:OG (d), S151:NT174:N	(I AND II) OR (I AND III) OR (II AND III)
		III	D310:OD1 D194:OD1 A172:O T174:OG1 (a)	
OMPDC	**1eix**, 1jjk	I	R192:NH2 R222:NH1, R222:N Q201:NE2	(I AND II) OR (I AND III) OR (II AND III)
		II	D22:OD2 D76:OD2 D71:OD2	
		III	D71:OD2 T131:N T131:OG1 (a) N201:NE2 K73:NZ	
PPAR-g	**1fm9**, 1i7i, 1k74, 1knu, 1nyx, 1wm0, 1zeo, 2ath, 2f4b, 2g0g, 2g0h, 2gtk, 2hwq, 2hwr	I II	Y473:OH (d) H449:NE2 S289:OG (d) H323:NE2	(I AND II) OR (I AND III) OR II
		III	Y327:OH (d)	
PLA2	1ayp, 1db4, 1db5, 1dcy, 1j1a, **1kqu**, 1kvo, 1poe	I II	H47:ND1 (d) G29N H47:ND1 (a) G29N	1) I AND III AND IV 2) II AND III AND IV
		III	Ca^{2+} OR Ca^{2+}-H_2O[e]	
		IV	G31:N	
PTP1B	1aax, 1bzc, 1bzj, 1c83, **1c84**, 1c85, 1c87, 1c88, 1ecv, 1g7f, 1g7g, 1gfy, 1jf7, 1kak, 1kav	I II	R221:NE, R221:N F182:N G220:N I219:N D179:OD2 K118:NZ Y44:OH (d)	I AND (II OR III)
		III	S215:OG (d) S215:OG (d) A217:N	

[a] structures considered for structural filters development. Structures, from which a model for virtual screening was developed, are typed in bold

[b] protein atoms participating in specific interactions with ligands. When protonation state of amino acid (like histidine) may vary, or protein group may both donate and accept H-bonds (like hydroxyl group) clear indication of a H-bond donor (d) or acceptor (a) is provided. Residue numbering corresponds t the PDB files used as a source for model development

[c] structural filter is denoted as a logical expression, in which distinct groups of specific interactions are combined as either mandatory or obligatory

[d] distinct (spatially separated) groups of specific interactions with protein are enumerated

[e] ligand either coordinates metal ion, or does not replace water molecule form its coordination sphere

binding site was taken from the corresponding PDB structure (Table 1) in each case. The size of the grid box for ligand docking was set to span 6 Å (the default value) in each direction from the reference ligand. The VS-score produced by Lead Finder was used to rank-order ligands in virtual screening (not to be confused with the dG-score that estimates the free energy of protein-ligand binding). Only the top-ranked poses were used for all analyses.

A set of 300,000 ligands comprising the STK library of Vitas-M Laboratory [14] was used as a decoy ligand set to benchmark the performance of Lead Finder in virtual screening. For the quantitative characterization of virtual screening efficiency, the enrichment factor (EF) was used. For a certain fraction of the screened library EF equals the number of recovered active ligands divided by the number of active ligands which could be found in the mentioned fraction of the library by chance. Thus EF can be represented by a continuous curve or by a discrete set of values calculated for certain fractions of recovered active ligands (for example, commonly used indicators EF20, EF40, and EF70 denote the enrichment factors at 20%, 40%, and 70% of recovered active ligands correspondingly).

Structural filtration of docked ligand poses

For each top-scoring ligand pose obtained in virtual screening the formation of a particular set of specific protein-ligand interactions (H-bonds or coordination to metal ion) was monitored. For each protein the set of prerequisite specific interactions (Table 1) was assigned by visual inspection of available PDB structures of corresponding protein-ligand complexes. The structural filtration was done automatically using the structure_filter module of Lead Finder. This module takes a protein structure file and a list of protein residues for which it checks the formation of hydrogen bonds with potential inhibitors from an sdf output file with docked ligand poses produced by Lead Finder. The formation of a hydrogen bond was deemed to occur when the distance between the hydrogen bond donor and acceptor was found to be less than 3.5 Å and the angle was greater than 150°. The coordination with metal ion (in case of LTA4H or PLA2) was detected when the distance between the ligand donor atom and metal was less than 2.5 Å. Finally, the list of ligands ranked by their VS-score was reorganized by moving the ligands that did not satisfy the structural criteria to the end of the list.

Results and discussion

Let us consider the structural filter design principles and their application on a few particular examples. Consider an ADRB2 receptor, which structure has been resolved recently [15] and immediately spurred a wealth of modeling studies. The ligand binding site of ADRB2 represents a well-shaped cavity that can easily accommodate big ligands, both flexible and rigid, therefore making the discrimination of true binders from decoy ligands a challenging task. However, an inspection of the two available PDB structures (2rh1 and 3d4s) of ADRB2 with its ligands (antagonists) suggests that hydrogen-bonding to the receptor's residues N312 and D113 may be a structurally conserved pattern of protein-ligand recognition. Upon binding to ADRB2, the ligand forms four H-bonds with residues D113 and N312 and replaces water molecules (bound in this overall hydrophobic sub-site) into bulk solution. Obviously, such binding mode of a ligand is energetically favorable since it would not form as many H-bonds in water solution (only at the cost of entropic losses of the solvent pre-organization). Alternatively, if a ligand does not form the said four H-bonds its binding is hardly energetically favorable. So, on the basis of experimental evidence (crystallographic structures) and the qualitative explanations, we could suggest that formation of four H-bonds (two–with D133 and two–with N312) constitutes a prerequisite structural criterion for ligand binding to ADRB2 (Fig. 1a).

However, specific protein-ligand interactions intended to filter out non-binders are not restricted to H-bonds. The coordination to protein-bound metal ions may be viewed as another structural filter. For example, consider LTA4H enzyme that contains Zn^{2+} ion, crucial for ligand binding and catalysis, in its active site. An analysis of the ten available PDB structures of liganded LTA4H revealed three major groups of specific interactions: coordination to Zn^{2+} ion (structures 1h19, 1gw6, 1hs6, 1sqm, 2vj8, 3chp, 3chq, 3chs); H-bonding of ligand's $-NH_3^+$ group with residues E271, E318 and Q136 (structures 1gw6, 1hs6, 3chr, 2vj8, 3cho, 3chq, 3chr, 3chs); H-bonding of ligand's $-COO^-$ group with G269, G268, K565, R563, and Y267 (structures 1gw6, 1hs6, 2vj8). Due to the spatial separation of the mentioned clusters of interactions, ligands of LTA4H are free to choose between some of them, unlike in the ADRB2 case, where crucial protein residues comprised an inseparable network of interactions. If we exclude structures 1h19, 1sqm, and 1hs6 containing acetate as a ligand (due to its very small size), we mention that LTA4H binders realize the following binding patterns. First, the ligand either coordinates Zn^{2+}, or does not contact it at all, leaving the water molecule in the ion's coordination sphere (like in structures 3cho and 3chr). Second, the ligand may form contacts with the second group of residues or with the third one, or with both of them, but at least one group of contacts must be saturated (Fig. 1b). Of course, taking into account the approximate nature of the structural filter definition, one may choose more or less stringent criteria to regulate the number of suitable compounds. For example, one might want to select only those LTA4H ligands that satisfied all structural criteria simultaneously.

To address an additional issue of structural filtration, consider the PLA2 case. Ligands of PLA2 use to coordinate Ca^{2+} ion, present in the enzyme active site, and accept H-bonds from the main chain of G29 and G31, and the side chain of K62. Additionally, PLA2 ligands, which contain an amide group, donate a hydrogen bond to H47, as in the case with PDB structures 1ayp, 1db4, 1db5, 1j1a, 1kqu, 1kvo. Meanwhile, some of the ligands contain a carboxyl or phosphonic group instead of the amide (as in the case with PDB structures 1dcy and 1poe respectively), and cannot donate a hydrogen bond to H47. However, the distance between the ligand's oxygen atom and the ND1 atom of H47 (2.66 Å in 1dcy and 3.25 Å in 1poe) suggested the existence of a H-bond that could take place only if the ND1 atom was protonated and thus H47 was in the charged form. So, proper modeling of ligand binding to PLA2 must account for two protein states: one with neutral H47 that accepts a H-bond from the ligand, another—with charged H47 that donates a H-bond to the ligand. To foresee these

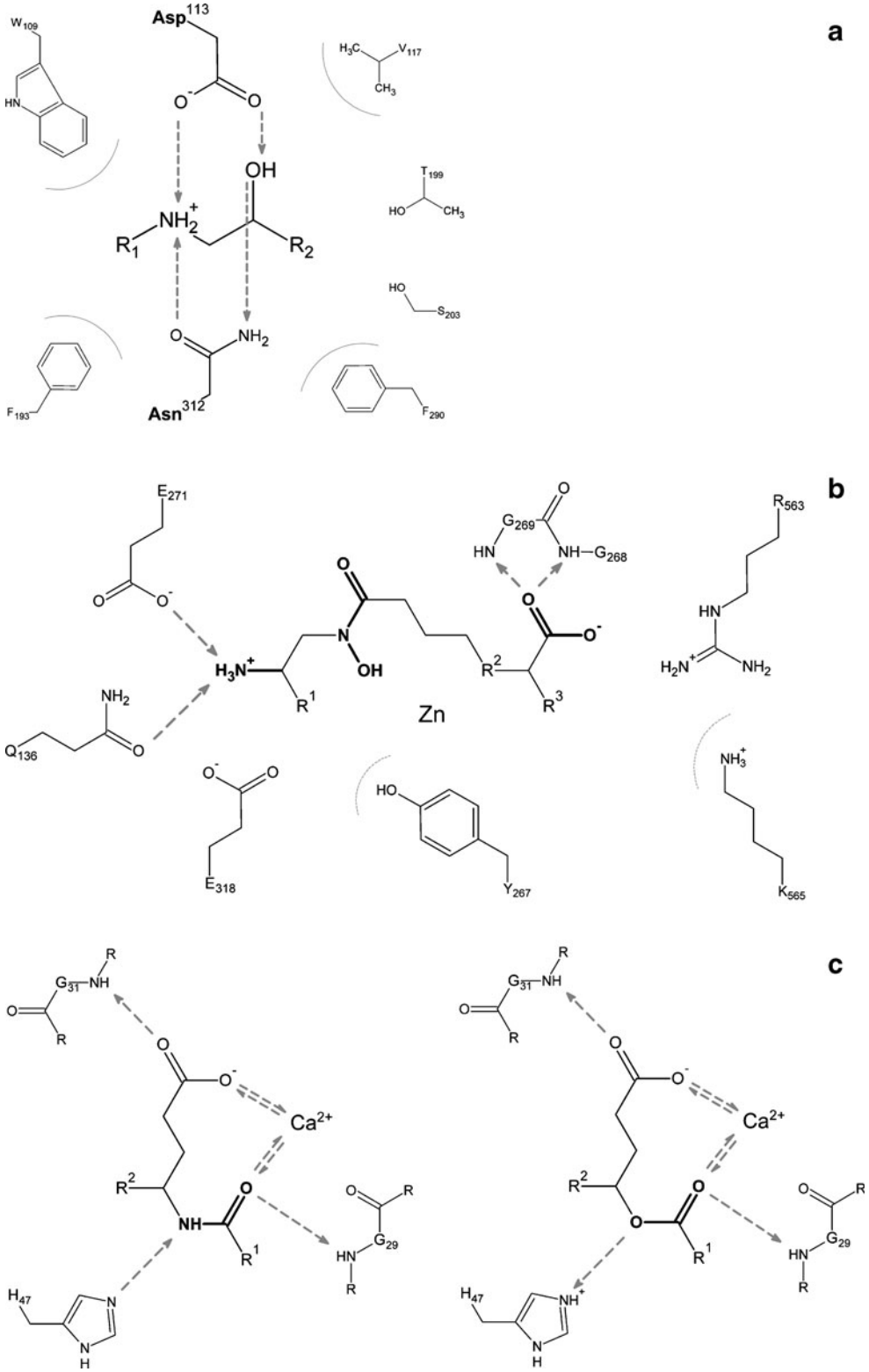

Fig. 1 Illustration of the structural filters definition for (**a**) ADRB2, (**b**) LTA4H, (**c**) PLA2

two possibilities we could construct two models of PLA2 structure (with either neutral or charged H47), dock ligands independently to these models and apply distinct structural filters to the resulting ligand poses. However, we observed that ligands from 1dcy and 1poe were docked correctly to the PLA2 model prepared from the structure 1kqu, in which H47 was in the neutral form (probably, coordination to Ca^{2+} and other interactions resulted in the correct ligand placement). Thus, only one protein model was used by us to dock ligands, while two distinct structural filters—to capture

Table 2 The accuracy of docking native ligands to their targets and the corresponding calculated binding energies

Target	Ligands		dGcalc, kcal/mol[c]		
	total[a]	PDB[b]	min	max	average
ADRB2	48 (41)	0 (0)	−13.1 (−13.1)	−5.2 (−5.2)	−10.0 (−9.8)
DHFR	25 (21)	16 (13)	−13.6 (−13.6)	−5.7 (−5.8)	−10.1 (−10.7)
DHODH	16 (15)	11 (10)	−13.9 (−13.9)	−5.4 (−5.4)	−10.7 (−10.7)
HMG-CoAR	16 (14)	9 (8)	−12.7 (−12.7)	−8.7 (−9.3)	−10.5 (−10.8)
LTA4H	31 (29)	6 (5)	−15.3 (−15.3)	−6.7 (−9.7)	−12.5 (−12.9)
mGluR3	14 (12)	5 (5)	−10.9 (−10.9)	>0 (−7.0)	−3.3 (−9.3)
OMPDC	12 (12)	8 (8)	−13.0 (−13.0)	−10.1 (−10.1)	−11.3 (−11.3)
PPAR-g	18 (18)	14 (14)	−13.9 (−13.9)	−9.7 (−9.7)	−11.6 (−11.6)
PLA2	15 (15)	13 (13)	−12.7 (−12.7)	−5.1 (−5.1)	−9.6 (−9.6)
PTP1B	27 (27)	24 (24)	−11.6 (−11.6)	−8.0 (−8.0)	−9.6 (−9.6)

[a] Total number of native ligands used in the virtual screening experiment for the corresponding target. The number of native ligands, which were correctly docked under default settings of Lead Finder is provided in the parenthesis (ligands, which lack PDB structure, were considered correctly docked if their calculated pose satisfied the structural filter criteria for the corresponding target)

[b] The number of native ligands for which the PDB structure was available. The number of ligands which were correctly docked under default settings of Lead Finder is provided in the parenthesis

[c] Minimal (min), maximum (max) and average binding energies calculated with Lead Finder for all native ligands for the corresponding target, and for ligands, which were correctly docked (data in the parenthesis)

ligands accepting a H-bond from H47 or donating it—were applied afterward (Fig. 1c). In our opinion, the example with PLA2 demonstrates flexibility of structural filters with respect to the protein (protonation) state. Probably, multiple filters can compensate, at least to some extent, the limitations of docking caused by the rigid protein approximation.

Similar considerations, both structure- and knowledge-inspired, led us to the construction of structural filters for the currently studied proteins (Table 1).

A quick look at Table 1 suggests that there are proteins (like ADRB2, DHFR, and DHODH), in which the structurally conserved specific interactions are compactly placed in the binding site, and saturation of all these interactions is clearly anticipated for protein-ligand binding. However, in most cases clusters of interactions are distributed over the binding site, therefore a simultaneous saturation of all interactions is no longer mandatory. At the same time, the currently available structural data do not allow exact identification of combina-

Table 3 Performance of docking-based virtual screening (I), and docking-based screening followed by structural filtration (II). Indicators calculated over correctly docked native ligands only are provided in parenthesis

Target Name	First native ligand[a]		Last native ligand[b]		EF40		Selected ligands[c]
	I	II	I	II	I	II	
ADRB2	117	4 (4)	109475	177 (177)	25	2356 (2403)	1476
DHFR	5	1 (1)	20018	341 (291)	564	10922 (10922)	3907
DHODH	263	17 (17)	28513	3422 (3422)	139	3215 (3274)	3622
HMG-CoAR	93	17 (17)	19921	1372 (1200)	194	1876 (1876)	3520
LTA4H	4	1 (1)	29984	2341 (1594)	191	1783 (1783)	2574
mGluR3	93	33 (33)	2420	1488 (389)	524	1396 (1396)	1488
OMPDC	1	1 (1)	55	33 (33)	24009	24009 (24009)	5664
PPAR-g	171	9 (9)	9304	6718 (6718)	39	259 (259)	13572
PLA2	34	6 (6)	21013	7415 (7415)	15	350 (350)	10415
PTP1B	16	3 (3)	1895	294 (294)	3001	9235 (9235)	2043

[a] Position of the first native ligand in the rank-ordered library after virtual screening

[b] Position of the last native ligand in the rank-ordered library after virtual screening

[c] The number of ligands in the screened library, which satisfied the structural filter criteria for the corresponding target

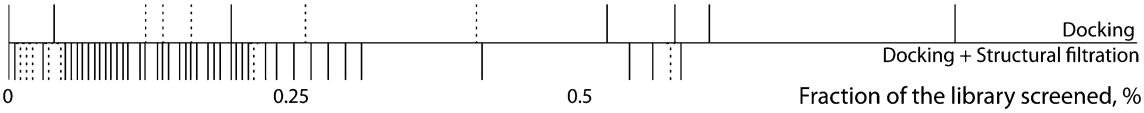

Fig. 2 Illustration of the improvement of virtual screening performance by structural filtration for ADRB2 as a target. Horizontal axis corresponds to the fraction of the library screened; positions of the native ligands obtained during virtual screening are depicted by vertical bars. Upper part of the plot corresponds to the performance of docking-based screening; lower part corresponds to the performance of docking-based screening followed by structural filtration. Ligands that were docked suggestively correctly (satisfied structural filter criteria) are depicted by continuous bars, the remaining (suggestively incorrectly docked) ligands are depicted by dotted bars. Lower part of the plot contains all (48) native ligands of ADRB2, which fall into the top 0.6% of the screened library

tions of structural criteria, which and only which are necessary for ligand binding. So, there is some space for experimenting with particular definitions and physical rationalization of structural filters.

The results of the application of structural filters to the docked ligand poses are summarized in Tables 2 and 3, while standard enrichment curves for all 10 targets are provided in the Supporting Information.

It can be seen that the structural filtration improves enrichment from several to several hundred folds of magnitude depending on the target. One of the most dramatic improvements is observed for the ADRB2 receptor, where accounting for the specific H-bonds is crucial for discriminating true binders from big and flexible decoys capable of filling the protein's binding cavity. The same factor reduced the false positive rate in cases of HMG-CoAR, LTA4H and PPAR-g, which were also characterized by a spacious active site. Pictorial illustration of the virtual screening improvement achieved by structural filtration in case of ADRB2 is presented in Fig. 2.

Another aspect of the positive impact of structural filtration is illustrated by such proteins as DHODH and mGluR3. Most of the native ligands of these proteins have medium binding potency and are quite small in size, so by default it is more difficult to recover such ligands in virtual screening. In those cases, the structural filtration reduces the false negative rate by moving correctly docked ligands with moderate score to the top of the list.

The structural filtration adds apparently little for OMPDC and PTP1B compared to the background enrichment. Obviously, this is due to the fact that the currently used native ligands for these proteins are quite potent binders, already receiving a high enough score (given they were correctly docked). However, the benefit becomes clear when potent binders receive a low score for some reason. For example, some of the native ligands of DHFR were scored relatively low probably due to a steric overlap with the protein (Table 2). However, due to the softness of the Lennard-Jones potential implemented in Lead Finder, these ligands could be docked correctly and hence were successfully recovered by the structural filtration. Thus, the deficiencies of the protein structure model or, more generally, the limitations of the rigid-protein approximation could be overcome (to some extent, of course) by applying structural filtration. An additional illustration of this hypothesis comes with PLA2, for which we were able to account for two alternative protonation states of H47 by using a single structure model of the protein in combination with two distinct structural filters.

Finally we have to admit that structural filtration works properly only for correctly docked ligands, that is for (active) ligands which docked pose conforms the designed structural filter. When the (active) ligand is misdocked, it does not form the obligatory interactions anymore, so the structural filtration will be helpless to recover this ligand. Thus the accuracy of docking imposes certain restrictions on the quality of virtual screening, which exploits structural filtration. Our current assessment of the influence of docking errors on the virtual screening quality can be traced from Tables 2 and 3, where the integral benchmarks and the benchmarks obtained using only correctly docked active ligands are provided. One can see that both estimations are quite similar, probably due to high docking success rate achieved by Lead Finder.

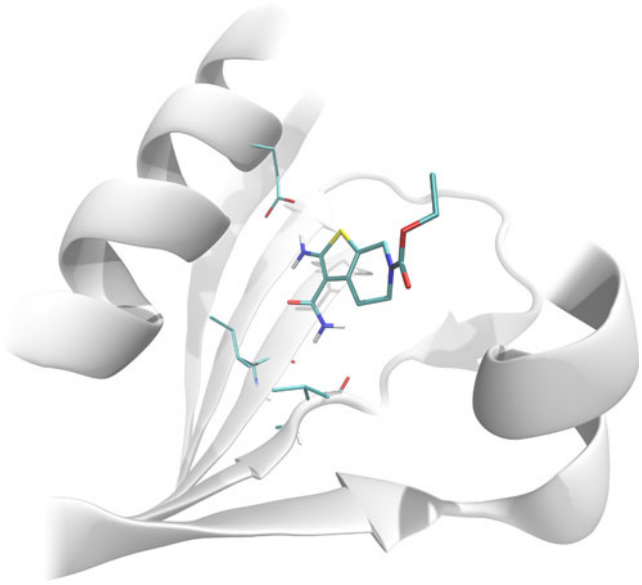

Fig. 3 An example of novel potential DHFR inhibitor recovered by virtual screening with subsequent structural filtration. The structure is disclosed with permission from the Vitas-M Laboratory

From the practical point, the structural filtration can be viewed as a valuable instrument to increase quality of the focused libraries produced by docking-based virtual screening. For example, in the current study we were able to retrieve from 1476 to 13572 compounds (from totally 300000 ligands screened), depending on the target, which fit the designed structural criteria (Table 2). A nice example of such compounds is depicted in Fig. 3. Interestingly, one of the potential DHFR inhibitors (Fig. 3) is a fragment-like molecule, which could hardly be recovered by traditional docking-based virtual screening. Thus we can suggest that the structural filtration will be especially valuable in the fragment-oriented research, where moderate to poor binding affinities would hamper straightforward ligand scoring.

Conclusions

The current study has introduced the methodology of structural filtration of docked ligand poses, which aimed at improving the virtual screening performance. For a particular protein the structural filter is viewed as a set of specific interactions (H-bonds, coordination to metal ions, *etc*.) with ligands that are observed in the available crystallographic structures and can be viewed as important for the protein-ligand recognition. The construction of such structural filters and their application to the virtual screening results on 10 diverse protein targets revealed significant (from several to several hundred folds of magnitude) improvement in the enrichment. The current results suggest a high added value of structural filtration to the background virtual screening, especially in reducing the false positive and false negative rates, and in leveraging the impact of deficiencies of protein structure models on docking and scoring. One of the promising applications of the structural filtration of ligand poses is expected to be in the fragment-based drug discovery.

Acknowledgments We thank Val Kulkov (BioMolTech Corp) for revising the manuscript. Our special thanks to Prof. Viktor Gergel and Alexander Grishagin, the Center of Supercomputer Technologies of the N.I. Lobacevsky State University of Nizhni Novgorod, for our access to the high-performance computing cluster (Nizhni Novgorod segment of the SKIF-grid program). The work was supported by the Foundation of Assistance to Small Innovative Enterprises (Contract №7168/4935r)

References

1. Richon AB (2008) Current status and future direction of the molecular modeling industry. Drug Discov Today. doi:10.1016/j.drudis.2008.04.008
2. Cheeseright TJ, Mackey MD, Melville JL, Vinter JG (2008) FieldScreen: virtual screening using molecular fields. Application to the DUD data set. J Chem Inf Model. doi:10.1021/ci800110p
3. Warren GL, Andrews CW et al (2006) A critical assessment of docking programs and scoring functions. J Med Chem. doi:10.1021/jm050362n
4. Tan L, Geppert H, Sisay MT, Gütschow M, Bajorath J (2008) Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. ChemMedChem. doi:10.1002/cmdc.200800129
5. Muthasa D, Sabnisa YA, Lundborga M, Karlén A (2008) Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering. J Mol Graph Model. doi:10.1016/j.jmgm.2007.11.005
6. Deng Z, Chuaqui C, Singh J (2004) Structural Interaction Fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. J Med Chem. doi:10.1021/jm030331x
7. Marcou G, Rognan D (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model. doi:10.1021/ci600342e
8. Hindle SA, Rarey M, Buning C, Lengauer T (2002) Flexible docking under pharmacophore type constraints. J Comput-Aided Mol Des. doi:10.1023/A:1016399411208
9. Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS (2007) A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. J Chem Inf Model. doi:10.1021/ci600253e
10. Nandigam RK, Kim S, Singh J, Chuaqui C (2009) Position specific interaction dependent scoring technique for virtual screening based on weighted protein-ligand interaction fingerprint profiles. J Chem Inf Model. doi:10.1021/ci800466n
11. Stroganov OV, Novikov FN, Stroylov VS, Kulkov V, Chilov GG (2008) Lead finder: an approach to improve accuracy of protein–ligand docking, binding energy estimation, and virtual screening. J Chem Inf Model. doi:10.1021/ci800166p
12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
13. Zhanga J, Aizawaa M, Amaria S, Iwasawab Y, Nakanoc T, Nakatac K (2004) Development of KiBank, a database supporting structure-based drug design. Comput Biol Chem. doi:10.1016/j.compbiolchem.2004.09.003
14. STK library (2007) Vitas-M Laboratory, Moscow. http://www.vitasmlab.com/compound-libraries-2.htm. Accessed 15 Jul 2009
15. Cherezov V, Rosenbaum DM, Hanson MA et al (2007) High-resolution crystal structure of an engineered human β2-Adrenergic G protein-coupled receptor. Science. doi:10.1126/science.1150577