Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Physiology - Paris 99 (2006) 232–244
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

Development and virtual screening of target libraries

Didier Rognan *

Bioinformatics of the Drug, CNRS, UMR 7175, 74 route du Rhin, F-67400 Illkirch, France

Abstract

The concomitant development of in silico screening technologies and of three-dimensional information on therapeutically relevant macromolecular targets makes it possible to navigate in the structural proteome and to identify targets fulfilling user-defined queries. This review illustrates some in-house recent advances in the development of target libraries and how they can be browsed to unravel chemogenomic information.

Keywords: Virtual screening; Docking; Chemogenomics

1. Introduction

Virtual screening of compound libraries (Schoichet, 2004) has recently gained considerable importance in early hit finding programs, notably when technological or economic hurdles disfavor experimental screening. Numerous successful applications of either ligand-based (Bajorath, 2002) or structure-based (Kitchen et al., 2004) in silico screening have been reported in the literature. Quite unexpectedly, the inverse paradigm still has not been deeply investigated. Given a set of ligands, is it possible to prioritize their most likely targets for experimental validation? Answering this question first requires the development of a library covering the most reliable target space (Lipinski and Hopkins, 2004). By target library, we mean here a collection of macromolecules for which either the amino acid sequence and/or three-dimensional (3-D) coordinates are available and can be browsed using simple queries. Then, an appropriate screening method has to be set up which is able to select a panel of targets fulfilling requirements imposed by either a ligand structure or a specific fingerprint (Attwood et al., 2003) or an evolutionary trace (Lichtarge et al., 1996). Once a target library has been developed, several applications can be foreseen: (1) simply compare targets and whenever possible relevant ligand binding sites, (2) predict the most likely target(s) of a given ligand, (3) predict a selectivity profile for either a target or a ligand, (4) predict the 'druggability' of a given target from a structural point of view. All these issues require early answers in the evaluation of drug discovery programs. We will try to review each of these applications in the coming sections.

2. Setting up target libraries

When developing a target library, a first compromise between available information (notably at the structural level) and the therapeutical relevance of selected targets has to be made. Many proteins for which fine structural details are known (e.g. toxins, antibodies) are not 'druggable'. Conversely, some important protein families for the pharmaceutical industry (e.g. G-protein-coupled receptors) are poorly understood at the 3-D level. Next, a scope has to be assigned to the library. Which target space has to be covered? Last, which kind of data (amino acid sequences, 3-D atomic coordinates) is browsed for defining a target list?

2.1. sc-PDB: a collection of active sites from the Protein Data Bank

2.1.1. Setting up the database

To establish the proof-of-concept that a protein library might be of screening interest, we have chosen the Protein

* Fax: +33 3 90 24 42 35.
E-mail address: didier.rognan@pharma.u-strasbg.fr

0928-4257/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2005.12.084

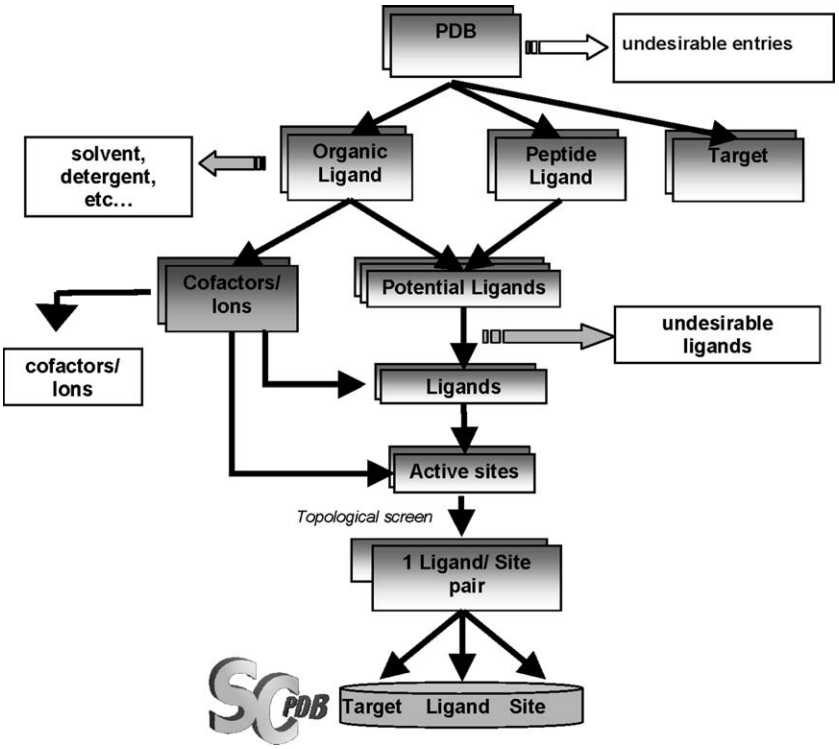

Fig. 1. Flowchart for developing the sc-PDB databank (http://bioinfo-pharma.u-strasbg.fr/scpdb/scpdb_form.html).

Data Bank (PDB) (Berman et al., 2000) as it is the major 3-D protein database for which experimentally determined protein coordinates are available. Several protein–ligand databases derived from the PDB have been recently described (Golovin et al., 2005; Hendlich et al., 2003; Kitajima et al., 2002; Kramer et al., 1999; Nissink et al., 2002; Roche et al., 2001). Relibase (Hendlich et al., 2003) easily allows retrieving protein–ligand complexes from a user-defined query focusing on specific molecular interactions. MSDsite (Golovin et al., 2005) is a database search and retrieval system for listing PDB entries fulfilling user-defined queries based on ligand information. The LPDB (Roche et al., 2001) stores 195 high-resolution protein–ligand complexes and related physicochemical descriptors as well as binding constants. Its main purpose, as well as related protein–ligand datasets (Kramer et al., 1999; Nissink et al., 2002) is to provide reliable 3-D information for calibrating docking algorithms and scoring functions. The ProLINT database (Kitajima et al., 2002) contains about 20,000 interaction data for two protein families (kinases, proteases) with attached information about the ligand, the protein, experimental binding constants and published literature. It has been used to derive structure–activity relationships and predict binding constants. LigBase (Stuart et al., 2002) is a database of ligand binding sites aligned with related protein structures and sequences containing 50,000 binding sites for heterogeneous ligands (ions, solvent, co-factors, inhibitors, etc.).

However, none of the above-mentioned databases are directly usable to generate a collection of 'druggable' protein active sites customized to accommodate small molecular-weight 'drug-like' ligands. Generally, no differences between solvent, detergent, co-factors and ligands (in the pharmaceutical sense) are made in the above-mentioned databases. To fill this gap, we recently developed a relational database (sc-PDB) (Bairoch et al., 2005) specifically customized for screening purposes (Fig. 1).

Starting from 27,000 PDB entries, a series of hierarchical filters has been applied to constitute the database as following:

- removal of undesirable entries: low resolution (>2.5 Å) X-ray structures, NMR structures;
- on-the fly detection of the molecule to which each referenced PDB atom belongs to (target, organic ligand, peptide ligand, co-factor, ion, solvent, detergent) thanks to knowledge-based rules and preexisting lists of 'HET' codes defined in the PDBsum database (Laskowski et al., 2005);
- removal of undesirable small molecular-weight ligands (solvent, detergents, ions and co-factors exhibiting atom types not recognized by classical docking algorithms);
- definition of putative ligands (organic or peptidic molecules, co-factor if present alone);
- definition of the binding site for each ligand (collection of amino acids for which any heavy atom is closer than 6.5 Å from any ligand atom);
- prioritization of a single ligand/active site for each PDB entry by calculating the buried surface area of the ligand and of the site, and selecting the ligand/site pair for which the percentage of burial is the highest;

- storage, for each selected PDB entry, of 3-D atomic coordinates in readable PDB format (target, active site) and SD/MOL2 formats (ligand, co-factors, ions).

2.1.2. Annotating the database

The current version of the sc-PDB database contains 5947 ligand-binding sites for 2626 small molecules; In total, the database refers to 5947 PDB entries. We assigned a unique UniProt (Bairoch et al., 2005) accession number to each protein, thereby identifying 1628 different proteins in the database. Additional information collected from both UniProt and PDB databanks was collected to obtain the source organism and the biological function of each protein. A functional classification of the database entries is shown in Fig. 2. Entries were separated into two superfamilies, namely enzymatic and non-enzymatic proteins. Out of the 5947 different entries of the database, ca. 85% are enzymes with a well-referenced EC (Enzyme Commission) number (Bairoch, 2000). The distribution of enzyme families displayed in Fig. 2 reveals that the most populated family is that of hydrolases (35% of the enzymes). This is correlated to the high number of proteases in the sc-PDB database. Fig. 2B gives an overview of the redundancy of current database entries. In most cases, less than 10 copies of an active site corresponding to a given protein are available in the database. The uneven protein entries distribution, which reflects the intrinsic PDB redundancy, is of great interest for application like virtual screening. Indeed, conformational differences between several copies of an active site reflect the local protein flexibility.

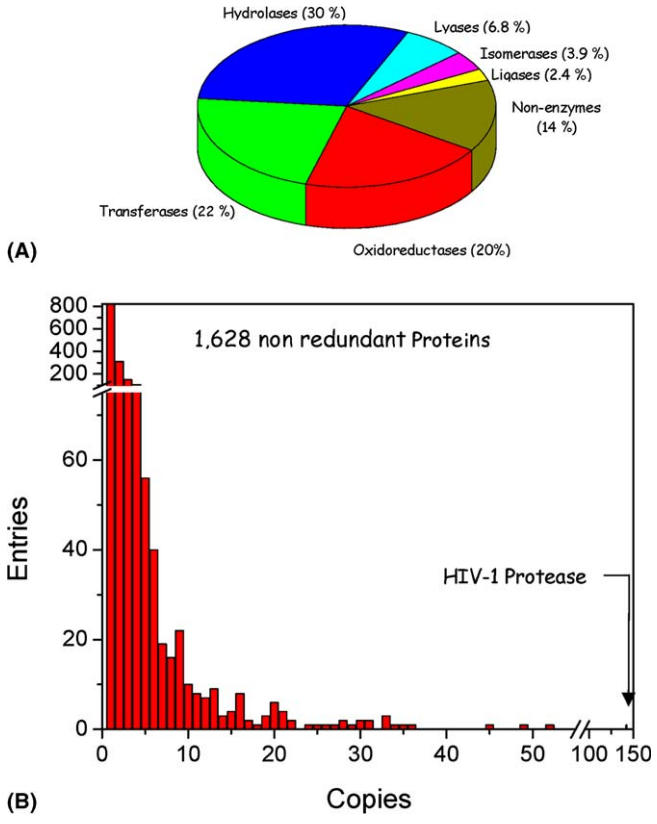

Fig. 2. sc-PDB content (release 3, March 2005): (A) distribution of enzymes and non-enzymes; (B) observed redundancy.

2.2. hGPCRDb: a collection of human non-olfactory GPCRs

2.2.1. Setting up the database

G Protein-Coupled Receptors (GPCRs) constitute a superfamily of membrane receptors of outmost importance in pharmaceutical research (Schwalbe and Wess, 2002). Hence, GPCRs are the macromolecular targets of ca. 30% of marketed drugs (Wise et al., 2004). The first draft of the human genome suggests that over 800 genes encode for a GPCR (Venter et al., 2004) out of which only a few (ca. 30) are currently addressed by marketed drugs. If one excludes the family of sensory receptors, about 400 GPCRs are potentially 'druggable' with ca. 120 proteins being still considered as orphan targets (Wise et al., 2004). Traditionally, the first stage in the design of GPCR ligands has focused on the potency of the ligands for the selected receptor target. Selectivity towards the host receptor is usually considered once potency has already been reached. It would however be highly desirable to consider selectivity as soon as possible in the design process. Ideally, one would like to consider the GPCR universe for designing a ligand with the desired selectivity profile. As addressing this issue by high-throughput screening is currently impossible, 'in silico' screening could provide a reasonable start. Indeed the recently described 2.8 Å-resolution X-ray structure of bovine rhodopsin (Palczewski et al., 2000) provides a possible 3-D template for modeling other GPCRs. Recent reports unambiguously demonstrated that rhodopsin-based GPCR homology models are accurate enough to propose reliable 3-D models of receptors very different from bovine rhodopsin (Petrel et al., 2003; Malherbe et al., 2003) and to identify new ligands by structure-based virtual screening (Becker et al., 2004; Evers and Klabunde, 2005; Evers and Klebe, 2004a; Varady et al., 2003). Of course, using classical homology modeling to establish a 3-D target library including ca. 400 reliable 3-D models is not possible. We therefore designed a chemoinformatic tool (GPCRMod) specifically dedicated to high-throughput GPCR modeling (Bissantz et al., 2004). From the very beginning, several considerations were taken in the design of the code: (i) the target library should cover all human non-olfactory GPCRs, (ii) a reliable multiple alignment of all investigated GPCRs should be obtained at the seven-transmembrane (7-TM) domain only, acknowledging that high-throughput modeling of intra- and extra-cellular loops is not feasible, (iii) the 7-TM binding cavity of every 3-D model should not be biased by the X-ray structure of bovine rhodopsin.

In a first step, 372 human GPCR amino acid sequences were aligned at the 7-TM by browsing the target sequence for family-specific fingerprints and motifs (Bissantz et al., 2004) (Fig. 3). Then, alignments were converted into 3-D

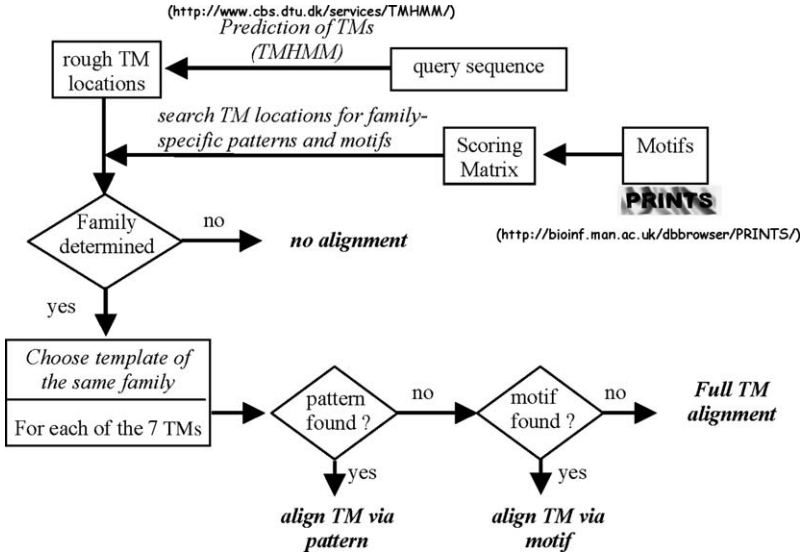

Fig. 3. Multiple alignment flowchart in GPCRMod.

model using a comparative modeling tool that uses a set of ligand-biased GPCR models as main chain templates, and two rotamer libraries for side chain positioning (Fig. 4). A key point of the modeling procedure is that 7-TM cavities are modeled starting from templates which prove useful to discriminate known ligands from decoys. Resulting 3-D models are qualitatively quite similar to those obtained by ligand-assisted comparative modeling (Evers and Klebe, 2004a,b; Evers and Klabunde, 2005) but obtained at throughput allowing the fast generation of hundreds of targets.

2.2.2. Annotation of the hGPCRdb

Assuming that similar targets recognize similar ligands, an accurate annotation of all entries should consider similarities/differences at their binding cavity. As most small molecular-weight ligands probably bind to the 7-TM core, all GPCR entries have been annotated using a chemogenomic procedure considering a fingerprint characterizing their 7-TM binding cavity. Thirty positions lining the retinal binding site in bovine rhodopsin, were extracted from all entries and concatenated into ungapped sequences out of which a phylogenetic tree could be derived using the standard UPGMA clustering method (Surgand et al., 2006) (Fig. 5).

Twenty two clusters could be unambiguously detected from the present analysis of 30 amino acid positions (Fig. 5). These clusters were defined in order to encompass the maximum number of related entries within a branch characterized by the highest possible statistical bootstrap value. Thirty four out of 372 entries could not be assigned to one of the existing 22 clusters and are defined as singletons. The herein presented tree is very similar to the most complete phylogenetic tree (GRAFS classification) known to date (Fredriksson et al., 2003) although the latter has been obtained from full TM sequences. In both classifications, GPCRs of the Frizzled, Glutamate, Secretin and Adhesion families cluster in well-separated groups whereas the large Rhodopsin family can be classified into 18

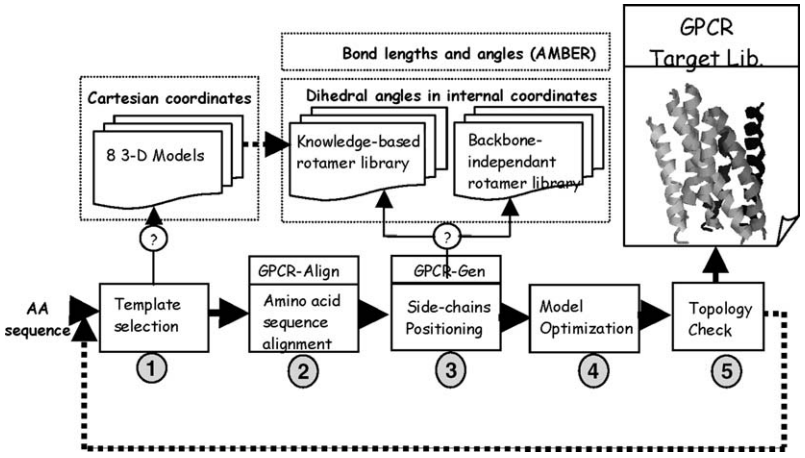

Fig. 4. 3-D model generation flowchart in GPCRMod.

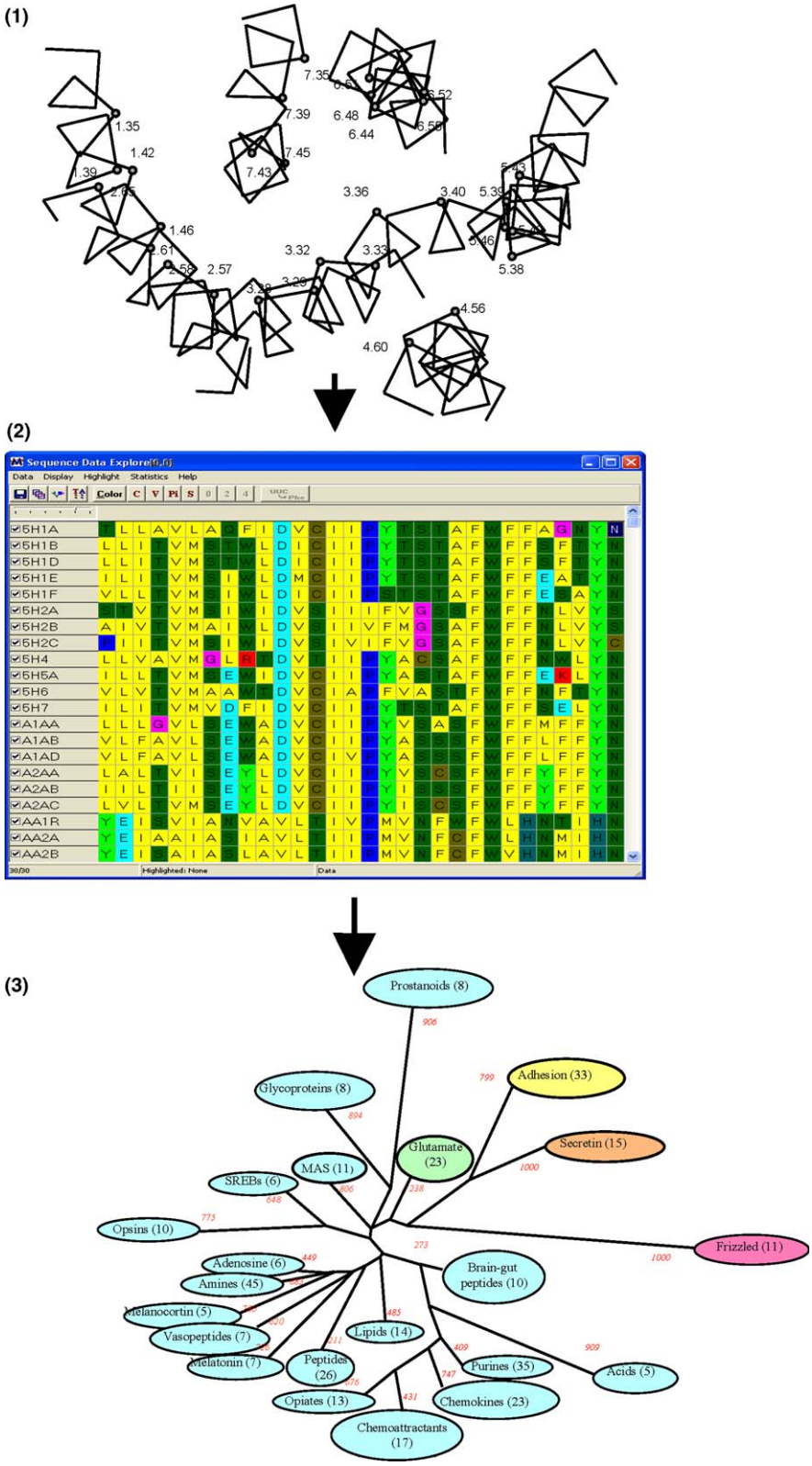

Fig. 5. Two-step protocol to generate a TM cavity-driven phylogenetic tree: (1) selection of 30 critical positions, (2) definition of ungapped sequences describing the 7-TM cavity, (3) TM cavity-derived phylogenetic tree for 372 human GPCRs. The consensus tree was derived from 1000 replicas using amino acid identity within a set of 30 discontinuous positions to measure protein distances. Numbers in commas indicate the number of entries in each cluster. Numbers in italic represent bootstrap values to assess the statistical significance of the tree. Receptors classified as singletons (see text) are not displayed here for sake of clarity. Glutamate, Rhodospin, Adhesion, Frizzled and Secretin subfamilies are colored in green, cyan, yellow, pink and orange, respectively.

different clusters. Remarkably, all known GPCR subfamilies (e.g. receptors for biogenic amines, purines, and chemokines) are reproduced with high bootstrap support. The five main families (Glutamate, Rhodopsin, Adhesion, Frizzled, Secretin) reported in the GRAFS classification are recovered with no overlaps between the corresponding clusters with the single exception of Q9GZN0 (GPR88), a rhodopsin-like GPCR clustered with class III GPCRs. Interestingly, receptors for which the orthosteric binding site is not located in the TM domain (Adhesion, Secretin and Glutamate families) are nevertheless grouped into homogeneous clusters. Relating cluster members to precise molecular features is here greatly facilitated by the analysis of a small subset of amino acids. For each of the 22 clusters, there is often a clear relationship between known ligand chemotypes (e.g. amines, carboxylic acids, phosphates, peptides, eicosanoids, and lipids) and the cognate TM cavities. For example, receptors for bulky ligands (e.g. phospholipids, prostanoids) have a TM cavity significantly larger than that for smaller compounds (e.g. biogenic amines, nucleotides). Receptors for charged ligands (cationic amines, phosphates, mono and di-carboxylic acids) always present among the 30 critical residues one or more conserved amino acid exhibiting the opposite charge (e.g. $Asp^{3.32}$ for biogenic amines; $Asp^{4.60}/Glu^{7.39}$ for chemokines; $Arg^{3.29}/Lys^{6.55}/Arg^{7.35}$ for nucleotides).

Our clustering approach implies two assumptions: (i) the overall fold of the 7-TM domain around the binding cavity has been conserved along evolution; (ii) critical hotspots spread over the 7-TM domain repeatedly account for ligand binding. Although solid biostructural data for the three most important GPCR classes (class I, class II, class III) are missing, numerous experimental do provide evidence for data in favor of strong similarities among many GPCRs: (i) residues known to affect small molecular-weight ligand binding to unrelated GPCRs are mostly spread among the herein selected 30 residues suggesting a common architecture of the TM pocket, (ii) many known ligands are promiscuous for even unrelated GPCRs and are usually anchored through so-called privileged structures to common subpockets of different GPCRs (Bondensgaard et al., 2004). Of course, we are aware that class II and class III GPCRs exhibit an additional orthosteric site located outside the 7-TM bundle. Therefore, conclusions drawn herein only apply to the 7-TM binding site.

3. Screening target libraries

Provided that a target library has been set up, two screening methods are possible (Fig. 6). In a 1-D screening, a query enclosing amino acid sequence information (e.g. fingerprint) is used to parse family-specific alignments in order to retrieve interesting targets. In a 3-D screening, the 3-D structure of either a ligand or a known active site is used to browse 3-D structures or homology models. Both applications will be detailed in the following section.

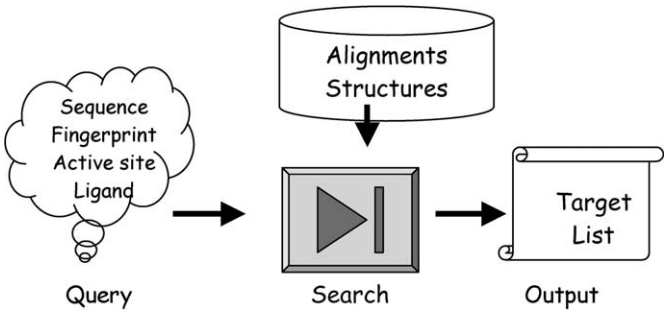

Fig. 6. Target library screening flowchart.

3.1. 1-D screening

Simple 1-D screening is less precise than 3-D screening but also less sensitive to errors. When applied entire target families (e.g. GPCRs, kinases), its accuracy only depends on the quality of the sequence alignment which is generally much higher that of 3-D structural models. Assuming that similar ligands should bind to similar cavities, browsing a database of sequence alignments can easily provide access to reliable information if specific fingerprints are already known. Three possible applications of 1-D screening of a GPCR target library are presented here.

3.1.1. Searching for orthologs/paralogs

The amino acid sequence of GPCRs is extremely variable in length (from 290 to 6300 residues for human GPCRs) notably at extra- and intra-cellular loops. Relying receptor comparisons on full sequence alignment may thus be quite misleading. Comparing the above-defined TM cavity-lining residues is much more appropriate. For any GPCR target of interest, these 30 residues can be identified quite unambiguously at least for rhodopsin-like GPCRs as several class specific TM fingerprints previously identified in this family of receptors can guide the sequence alignment (Bissantz et al., 2004).

As an example, we have been looking for the human ortholog(s) of a gene product from *C. elegans* (Y22D7AR_13) in order to predict the functional role of this presumed GPCR. Blasting its full amino acid sequence against human GPCRs leads to ambiguous conclusions because the level of sequence identity with the closest human GPCRs is low (usually in the 15–30% range) and that several candidates are possible (Table 1). Looking at local sequence identity within a set of 30 TM cavity-lining residues provides an answer that is easier to interpret because the sequence identities with the input query are much higher (above 70% for the first three 5-TH receptors, Table 1). Since 7 out of the top 10 ranked candidates were 5-HT receptors, the *C. elegans* gene product was predicted to be a receptor for serotonin, which was further experimentally validated (Segalat, personal communication). The proposed approach has the merit to be extremely fast (a few ms) but requires the a priori identification of the 7-TMs and a good sequence alignment of the latter domain. Therefore, the presence of TM fingerprints (usually

Table 1
Searching for the 10 closest human orthologs of the *C. elegans* Y22D7AR_13 gene product

Full sequence blast[a]			TM-cavity search[b]		
Rank	Receptor	Sequence identity, %	Rank	Receptor	Sequence identity, %
1	5-HT_{1B}	29.6	1	5-HT_{1A}	72.7
2	5-HT_{1D}	29.0	2	5-HT_7	72.7
3	5-HT_{1A}	26.7	3	5-$H5_A$	72.7
4	D_2	25.4	4	α_{1A}	69.7
5	5-HT_{1E}	24.8	5	5-HT_{1B}	69.7
6	α_{2A}	24.0	6	5-HT_{1D}	69.7
7	α_{2C}	23.8	7	5-HT_4	66.7
8	D3	23.6	8	α_{1B}	63.6
9	α_{2B}	21.1	9	α_{1D}	60.6
10	M_2	16.8	10	5-HT_{1E}	60.6

[a] Sequence comparison achieved using standard settings of the BLASTP program (http://services.bioasp.nl/blast/cgi-bin/blast.cgi).
[b] Sequence comparison achieved using our in-house GPCRfind program (http://bioinfo-pharma.u-strasbg.fr/gpcrdb/jss.html).

present in nearly all entries) (Bissantz et al., 2004; Surgand et al., 2006) in the input query is a prerequisite.

3.1.2. Computer-guided target deorphanization

A TM-cavity biased phylogenetic tree offers the opportunity to navigate in target space without the necessity to rely on questionable 3-D structures. Receptors close in target space can be expected to bind ligands close in chemical space. Known GPCR ligands are thus a good starting point to start deorphanizing receptors predicted to be close enough to liganded receptors (Table 2).

For example, focusing our cavity-based tree on two related orphan receptors (GPR19, GPR83) predicts a significant relationship to three tachykinin receptors (NK1R, NK2R, NK3R; Fig. 7). Likewise, GPR54 is predicted to be close to three galanine receptor subtypes (GALR, GALS, and GALT). Therefore, a rational start to find ligands for these three orphans would be first to test known ligands for neurokinine and galanine receptors, respectively. An experimental validation of this approach has been recently reported by scientists at 7TM-Pharma who identified ligands for the CRTh2 (GPR44) receptor by evaluating angiotensin 2 receptor (AG2R, AG2S) ligands, the corresponding targets being close when considering the 7-TM cavity (Frimurer et al., 2005).

3.1.3. Matching target space with ligand space

GPCR ligands sharing a common privileged structure and exhibiting promiscuous binding to unrelated GPCRs are a current important source for GPCR library design. Assuming that conserved moieties of the ligands are likely to bind to conserved subsites of the targets (Bondensgaard et al., 2004), matching privileged structures with TM hotspots can be achieved very easily without biasing the match by a manual or automated 3-D docking.

As an example, biphenyltetrazoles and biphenylcarboxylic acids (Fig. 8) are known to bind to at least six GPCRs (AG22, AG2R, AG2S, GHSR, L4R1, L4R2) (Smith et al., 1993; Reiter et al., 1998; Ji et al., 1994). Fine details of 3-D recognition of this privileged substructure by GPCR hotspots have been recently proposed by a thorough mutagenesis-guided manual docking of several

Table 2
Possible ligand source for some orphan GPCRs

Orphan receptor(s)[a]	Cluster[b]	Source
Q9GZN0, Q9NFN8	Glutamate	GABA-B allosteric ligands
Q8NHZ9, Q8TDU1	Glutamate	CaSR allosteric ligands
LRG4, LRG5, LRG6	Glycoproteins	LH/FSH nonpeptide ligands
Q8TDV5	Lipids	Cannabinoid receptors ligands
GP19, GP83	Peptides	Tachykinin receptors ligands
Q969F8	Peptides	Galanine receptor ligands
GPRA, PKR1, PKR2	Vasopeptides	Oxytocin/vasopressin receptor ligands
O14804, GP57, GP58	Amines	Biogenic amine receptors ligands
GP39	Brain-gut peptides	Neuromedin U receptors ligands
O75307, RDC1	Chemokines	Chemokine receptor ligands
GPR7, GPR8	Opiates	Somatostatine receptor ligands
GP15, GP25, GP44, GPR1	Chemoattractants	Angiotensin II receptor ligands
EBI2, GP92, P2Y5	Purines	LPC/SPC receptor ligands
G171, GP87	Purines	Purinergic nucleotide receptor ligands
GP17, GP34, FK79, P2YA	Purines	Cysteinyl Leukotriene receptor ligands

[a] Receptors are labeled according to their UniProt (Bairoch et al., 2005) entry name.
[b] For cluster definition, see Surgand et al. (2006).

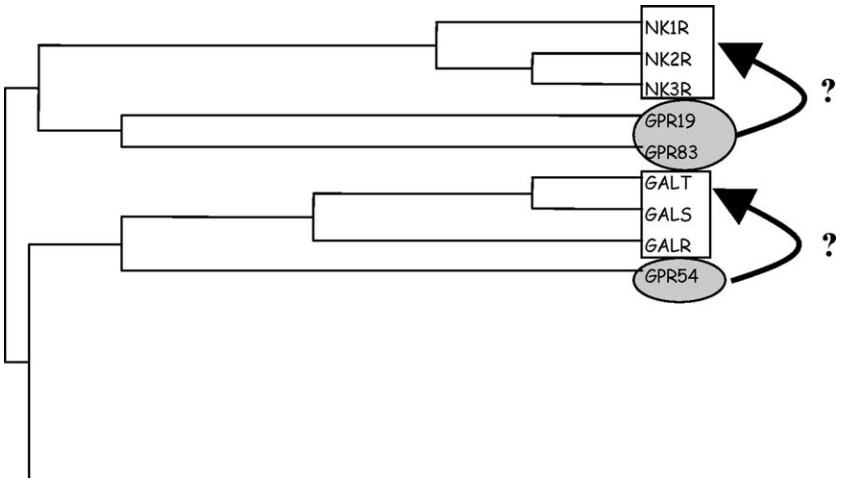

Fig. 7. Close up to the peptide receptors cluster.

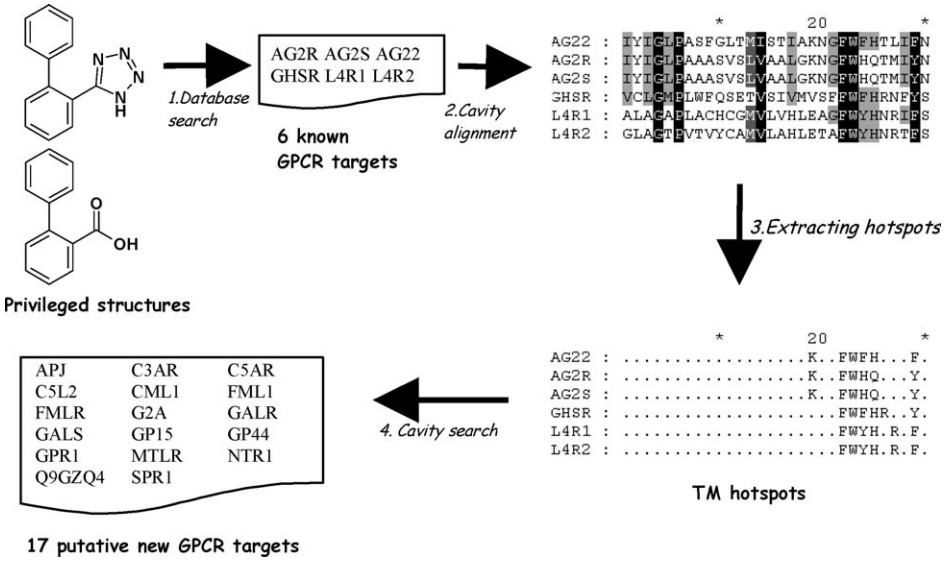

Fig. 8. Matching privileged structures of known GPCR ligands to TM hotspots. An in-house GPCR ligands database is searched to retrieve privileged structures common to multiple GPCRs and to find conserved residues within the 7-TM cavity of selected entries. Browsing the in-house GPCR cavity database (sequence of 30 critical positions lining the 7-TM cavity of 372 human GPCRs) allow to retrieve new GPCR entries satisfying the query and likely to accommodate the privileged structure.

GPCR ligands (Bondensgaard et al., 2004). We propose here a much simpler approach leading to the same outcome; looking at the 30 residues lining the TM cavity of the later six GPCRs allows us to clearly identify putative TM residues able to interact with this substructure (Fig. 8).

Conserved aromatic residues are likely to interact with the biaryl moiety cluster between TMs 6 and 7 (Phe$^{6.44}$, Trp$^{6.48}$, Phe/Tyr/His$^{6.51}$, Phe/Tyr$^{7.43}$). A positively charged residue that probably interacts with the bioisosteric tetrazole and carboxylate groups should be located nearby the aromatic cluster. Hence, three basic residues (Lys$^{5.42}$, Arg$^{6.55}$, and Arg$^{7.35}$) fulfill this requirement. Last a polar side chain at position 6.52 (His/Gln) is conserved for the six investigated GPCRs and might H-bond to the acidic moiety of the privileged structure. We have then clearly identified the same important anchoring residues than Bondensgaard et al. (2004) by simply looking at sequence alignments of TM cavity-lining amino acids, without relying on any 3-D docking data. Searching our TM cavity database for additional GPCRs fulfilling the above-described requirements (Phe$^{6.44}$, Trp$^{6.48}$, Phe/Tyr/His$^{6.51}$, Phe/Tyr$^{7.43}$ and Lys$^{5.42}$ or Arg$^{6.55}$ or Arg$^{7.35}$ and His/Gln$^{6.52}$) permits us to extract 17 new GPCRs that might accommodate biphenyl-tetrazoles and biphenyl-carboxylic acids (Fig. 8). Among putative targets are 10 chemoattractant receptors (APJ, C3AR, C5AR, C5L2, CML1, FML1, FMLR, GP15, GP44, and GPR1), three brain-gut peptide receptors (MTLR, NTR1, and Q9GZQ4), two cationic phospholipid receptors (G2A, SPR1) and two peptide receptors (GALR, GALS). This target list encompasses

receptors recently identified by Bondensgaard et al. (2004) (e.g. APJ, NTR1). It also suggests totally new putative targets for the investigated privileged structure that might serve as a common scaffold for small-sized combinatorial libraries targeting the new receptors list.

3.2. 3-D screening

High-throughput docking of large chemical libraries (Halperin et al., 2002) has established as a promising tool for identifying new hits from protein 3-D structures coming mostly from X-ray diffraction data (Kitchen et al., 2004) but also from homology modeling (Evers and Klebe, 2004b). Finding out of a large library which ligands are likely to bind to a protein of interest is slowly turning to routine computational chemistry (Schoichet, 2004). Surprisingly, the opposite question is still an issue. Given a known ligand, is it possible to recover its most likely target(s)? Answering this question using the above-mentioned docking approach implies first the development of a collection of protein active sites (see Section 2), and then use of an inverse docking tool able to dock a single ligand to multiple macromolecules. Although inverse screening uses the same paradigm as ligand screening (predicting the most likely ligand-target interactions from molecular docking), docking a single ligand to a target library is more difficult to setup than classical docking of a ligand library to a single target. One should automate the generation of input files (3-D coordinates of the target or/and of the cognate binding site; docking configuration file) for a large array of heterogeneous targets, which is much more difficult than setting up a reliable set of coordinates for a ligand library. Notably, protein and binding site 3-D coordinates should be prepared automatically and should be rendered suitable for docking by removal of any additional molecule (solvent, ion, and co-factor) not essential for ligand binding. We have chosen the GOLD docking software (Verdonk et al., 2003) for two main reasons: (i) it is one of the most robust and accurate docking tool in our hands (Kellenberger et al., 2004); (ii) it only requires a single configuration file whose distribution over a target library is easy to process.

3.2.1. 3-D screening of the PDB: proof of concept

The first validation of inverse screening was to recover among 2 150 entries of the sc-PDB (release 1, February 2004) the true target(s) of either selective (e.g. biotin, 6-hydroxyl-1,6-dihydropurine ribonucleoside) or promiscuous ligands (e.g. 4-hydroxytamoxifen, methotrexate). Screening the sc-PDB database clearly allowed to unambiguously recover the true targets of the four investigated ligands (Paul et al., 2004). When screening our database for potential targets of biotin, 7 out of the 10 streptavidin entries present in the sc-PDB were ranked at the top eight positions with very good averaged fitness scores (Fig. 9). Interestingly, the three streptavidin copies with lower rankings (90th, 195th, 315th) correspond to either an active site

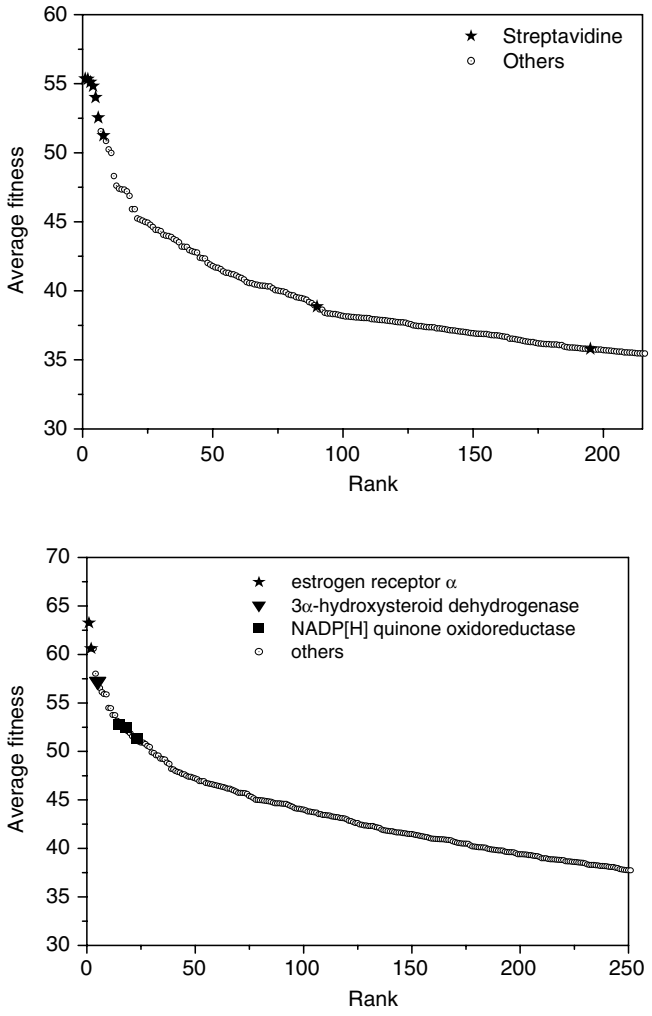

Fig. 9. Inverse screening of the sc-PDB database for finding the target of four small molecular weight ligands: top panel, biotin; bottom panel: 4-hydroxy tamoxifen. Filled stars indicate the different sc-PDB copies of the true target (top: streptavidin, bottom: estrogen receptor α). Filled triangles and squares indicate known secondary targets of 4-hydroxy tamoxifen (3α-hydroxysteroid dehydrogenase and NADP[H] quinone oxidoreductase, respectively). Targets are ranked by decreasing GOLD fitness scores averaged over 10 independent docking runs.

for which a key amino acid (Asp128) has been mutated (1swt) or alternative binding sites (peptide binding sites for 1vwr and 1rsu). Altogether, the proposed inverse screening protocol is able to unambiguously rank streptavidin as the most likely target for biotin with a percentage of coverage of 70% (7 out of 10) among the top 10 (0.5%) positions.

Likewise, the two sc-PDB entries of the estrogen receptor α were ranked at the top two positions when screening for the target of 4-hydroxy tamoxifen (Fig. 9) Interestingly, two other targets ((NADP[H] quinone oxidoreductase, 3α-hydroxysteroid dehydrogenase) at least ranked twice among the top 25 scorers, are known minor targets of this ligand. Therefore, inverse screening of target databases could also be viewed as a computational filter to roughly predict the selectivity profile of a given ligand and thus

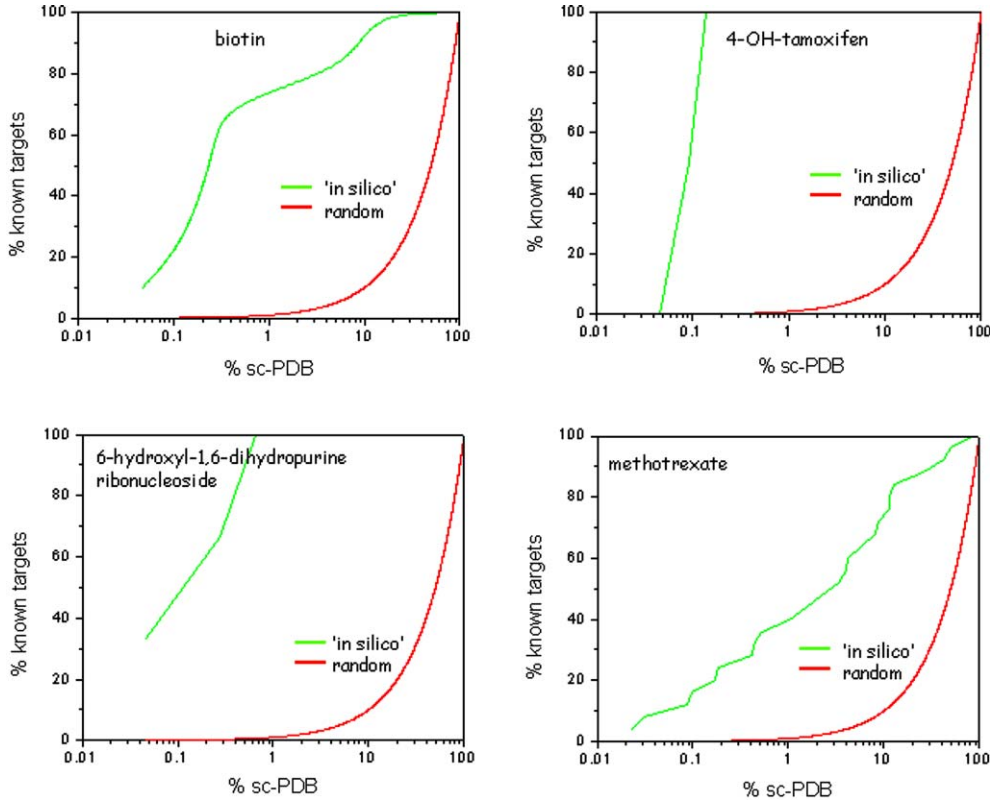

Fig. 10. Percentage of recovery of known targets as a function of the top scoring fraction found by inverse screening (green line) and random picking (red line). The percentage of coverage of known targets is the ratio in percentage between the number of true target entries recovered by inverse screening at a defined top scoring fraction and the total number of true target entries in the sc-PDB dataset.

putative side effects. When compared to random screening, a significant enrichment in the true target is observed among the top scorers (Fig. 10). Analyzing both the enrichment factor and the percentage of coverage of known targets indicates that the best compromise can be reached by selecting a very small fraction (0.5%) of the sc-PDB database. Even for the rather difficult case of methotrexate, selecting the top 2.6% scorers would allow to select 40% of all dihydrofolate reductase entries with a 15-fold enrichment with respect to random screening.

3.2.2. 3-D screening of the PDB: test case

Having validated the inverse screening approach for four unrelated ligands, a prospective screening was applied to the identification of putative targets for representative compounds of a scaffold-focused combinatorial library (Fig. 11). Release 1 of the sc-PDB (2148 entries) was screened to prioritize targets likely to accommodate five representative compounds from the library (Table 3). In the sc-PDB, a target is defined either as an enzyme from the PDB with a unique EC number, or a non-enzymatic protein with a unique name according to our previous

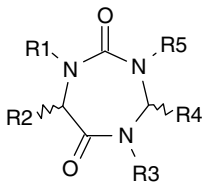

Fig. 11. The 1,3,5-triazepane-2,6-dione scaffold with five diversity points.

Table 3
Predicted targets for five compounds from a triazepanedione library

Target	E.C. number[a]	N[b]	Target rate (%)[c]				
			Cpd1	Cpd2	Cpd3	Cpd4	Cpd5
Aconitase	4.2.1.3	7	43		29	14	
DAAO[d]	1.4.3.3	2	50		50		
EST[e]	2.8.2.4	2	50	100			50
GT[f]	2.4.1–	2			100		50
HPRT[g]	2.4.2.8	6			33		17
MA[h]	3.4.11.18	5			20	100	
PLA2[i]	3.1.1.4	8			13		25
PNP[j]	2.4.2.1	6				17	83
TK[k]	2.7.1.21	5	80		20		

[a] Enzyme commission number.
[b] Number of copies in the sc-PDB (release 1, February 2004).
[c] Target rate: Percentage of targets ranked in the top 2% scoring entries.
[d] D-amino-acid oxidase.
[e] Estrogen sulfotransferase.
[f] Lipopolysaccharide 3-alpha-galactosyltransferase.
[g] Hypoxanthine-guanine phosphoribosyltransferase.
[h] Methionine aminopeptidase.
[i] Phospholipase A2.
[j] Purine nucleoside phosphorylase.
[k] Thymidine kinase.

annotation of the database. Differences related to species, isoforms or mutations are thus not considered in our classification scheme. For each of the five investigated compounds, a target was selected if it fulfills any of the three following criteria: (i) 50% of target entries present in the sc-PDB were scored, according to the average GOLD fitness score, among the top 2% scoring entries, (ii) the average fitness score for all entries of the corresponding target was above 50; two entries of the same target were scored in the top 2% scoring entries.

Out of the nine targets fulfilling this selection procedure, five were finally selected for biological evaluation (ES, MA, PLA2, PNP, TK; Table 3). About 24 compounds enclosing the five representative used for inverse screening were tested for inhibition of the above-described five enzymes. Micromolar inhibitors from this small library could be found for three out of the five predicted entries (MA, PNP, PLA2). A detailed description of corresponding structures and inhibitory constants will be reported elsewhere.

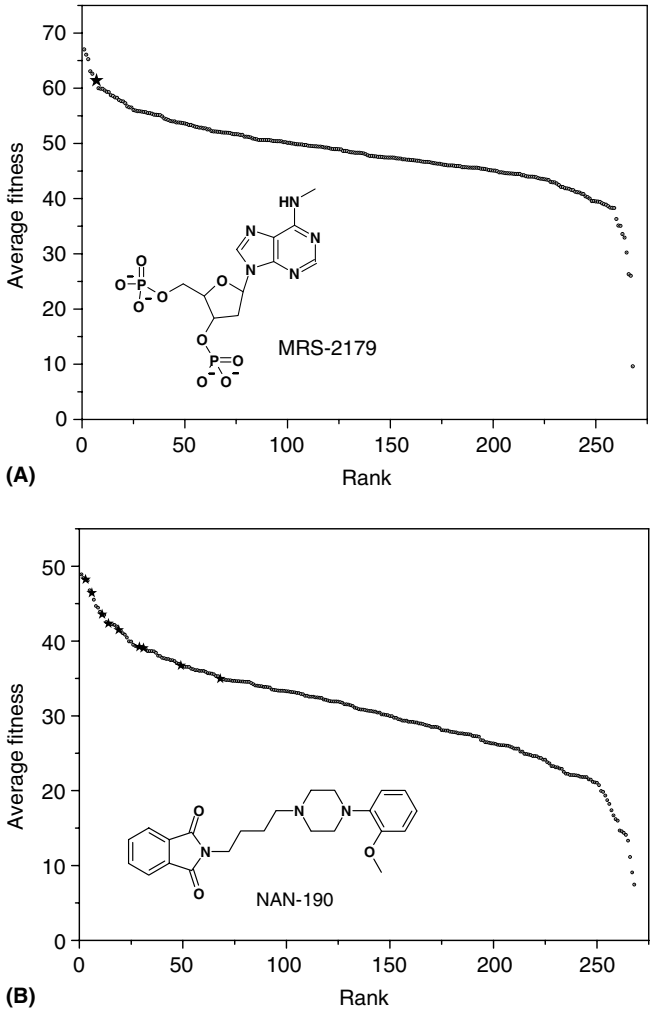

Fig. 12. Ranking of the true receptor(s) of a selective ligand (A: MRS-2179, $P2Y_1$ receptor antagonist) and of a promiscuous ligand (B: NAN-90, antagonist of the dopamine D_2 and D_3 receptors, serotonin 5-HT_{1A}, 5-HT_{1D}, 5-HT_{1F}, 5-HT_{2A}, 5-HT_{2C}, 5-HT_7 receptors, and adrenergic α_{1a} receptor). Known receptor(s) are indicated by filled stars. Targets are ranked by decreasing GOLD fitness scores averaged over 10 independent docking runs.

3.2.3. 3-D screening of the hGPCR library: proof-of-concept

Screening the collection of human GPCRs for identifying the receptors of known ligands is a quite demanding task regarding the current limited accuracy of GPCR models. We however tried to recover, from the GPCR target database, either the known receptor of a selective purinergic $P2Y_1$ antagonist (MRS-2179) or the known receptors of a promiscuous antagonist (NAN-190; Fig. 12) previously shown to bind to several monoamine receptors with nanomolar affinities (α_{1A}, D_2, D_3, 5-HT_{1A}, 5-HT_{1D}, 5-HT_{1F}, 5-HT_{2A}, 5-HT_{2C}, 5-HT_7). When screening the protein library for putative receptors of MRS-2179, the $P2Y_1$ receptor is indeed ranked among the top scorers (7th, Fig. 12A). Five out of the nine known targets of NAN-190, the second ligand investigated herein, are ranked in the top 25 positions, and seven out of nine in the top 31 positions (Fig. 12B). The worst-ranked true receptor (5-HT_{1A}) is ranked 68th. For both ligands, ca. 80% of GPCRs closely related to the true target(s) (P2Y receptors for MRS-2179; 5-HT receptors for NAN-190) usually clustered in the top 20% scorers. Thus, the current inverse screening procedure is more aimed at identifying the likely receptor subfamily (dopamine, serotonin, adenosine, etc.) than precisely mapping the individual preference for highly related GPCR subtypes. It could thus be used as a computational filter to study the most likely targets when addressing the selectivity profile of a given compound or trying to identify the yet unknown receptor of a molecule showing promising in vivo biological effects. Although the hGPCR database enclosed ground-state models suitable for docking antagonists and inverse agonists (Bissantz et al., 2003) we checked whether the same protocol could be applied to identify the

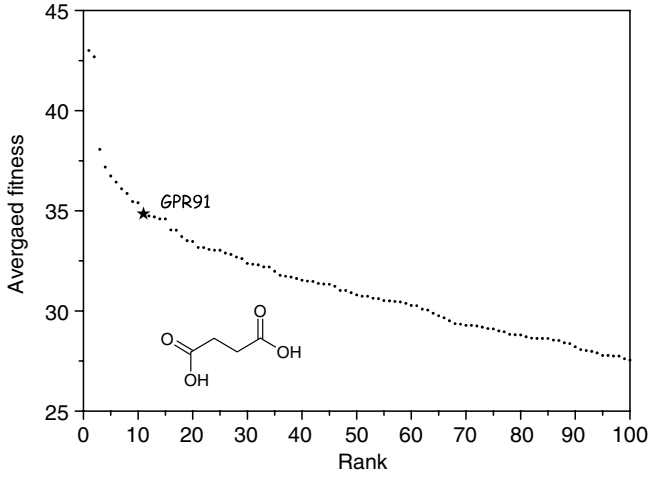

Fig. 13. Ranking of the true receptor (GPR91, filled star) of an endogenous ligand (succinic acid) by an inverse screening of a GPCR 3-D library. Targets are ranked by decreasing GOLD fitness scores averaged over 10 independent docking runs.

receptor of endogenous ligands. The hGPCR database was therefore screened to recover the receptor of succinic acid (Fig. 13), a recently identified ligand for the previously orphan GPR91 receptor (He et al., 2004). Although ground-state 3-D models were screened, the native receptor was surprisingly ranked among the top-scoring receptors (11th) in our inverse screening. Again, the true receptor was not ranked first but high enough in a shortlist that could be experimentally evaluated.

4. Conclusions

Virtual screening of target libraries offers new opportunities to prioritize a few targets for experimental evaluation by applying simple ligand-based or target-based queries. There is no reason that single ligand docking to a wide array of targets might not be as useful as classical docking of ligand libraries to a single protein, assuming comparable accuracies of input data. The increasing coverage of target space by the Protein Data Bank as well as the development of accurate comparative models describing entire protein families is likely to favor target screening in a near future. Pharmacophore-based and protein-based computational filters are nowadays used sequentially in virtual screening (Evers and Klabunde, 2005; Evers and Klebe, 2004b). One could imagine very similar scenarios for target screening, where interesting cavities would be first filtered by similarity measurements to a binding site of interest (Jambon et al., 2003; Weber et al., 2004), and then selected by ligand docking. Furthermore, orthogonal clustering of target families and of their ligands should soon provide precise chemogenomic information for selecting the most interesting compounds/scaffolds according to a predefined selectivity profile. Addressing simultaneously potency and selectivity in hit evaluation will undoubtedly affords added-value molecules in early drug discovery processes.

Acknowledgements

I would like to thank several former and current collaborators of the Bioinformatics group (C. Bissantz, G. Bret, E. Kellenberger, A. Logean, P. Muller, N. Paul, and C. Schalon) for their invaluable work in the development of target libraries. Financial support of the French Ministry of Research and Technology, and of the Alsace-Lorraine Genopole is acknowledged as well as the allocation of computing resources at the Centre Informatique National de l'Enseignement supérieur (CINES, Montpellier, France).

References

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., Zygouri, C., 2003. PRINTS and its automatic supplement, prePRINTS. Nucl. Acids Res. 31, 400–402.

Bairoch, A., 2000. The ENZYME database in 2000. Nucl. Acids Res. 28, 304–305.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S., 2005. The Universal Protein Resource (UniProt). Nucl. Acids Res. 33, 154–159.

Bajorath, J., 2002. Integration of virtual and high-throughput screening. Nat. Rev. Drug. Discov. 11, 882–894.

Becker, O.M., Marantz, Y., Shacham, S., Inbal, B., Heifetz, A., Kalid, O., Bar-Haim, S., Warshaviak, D., Fichman, M., Noiman, S., 2004. G protein-coupled receptors: in silico drug discovery in 3D. Proc. Natl. Acad. Sci. USA 101, 11304–11309.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucl. Acids Res. 28, 235–242.

Bissantz, C., Bernard, P., Hibert, M., Rognan, D., 2003. Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? Proteins 50, 5–25.

Bissantz, C., Logean, A., Rognan, D., 2004. High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment three-dimensional model building and receptor library screening. J. Chem. Info. Comput. Sci. 44, 1162–1176.

Bondensgaard, K., Ankersen, M., Thogersen, H., Hansen, B.S., Wulff, B.S., et al., 2004. Recognition of privileged structures by G-protein coupled receptors. J. Med. Chem. 47, 888–899.

Evers, A., Klabunde, T., 2005. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. J. Med. Chem. 48, 1088–1097.

Evers, A., Klebe, G., 2004a. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. J. Med. Chem. 47, 5381–5392.

Evers, A., Klebe, G., 2004b. Ligand-supported homology modeling of g-protein-coupled receptor sites: models sufficient for successful virtual screening. Angew. Chem. Intl. Ed. Engl. 43, 248–251.

Fredriksson, R., Lagerstrom, M.C., Lundin, L.G., Schioth, H.B., 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol. Pharmacol. 63, 1256–1272.

Frimurer, T.M., Ulven, T., Elling, C.E., Gerlach, L.O., Kostenis, E., Hogberg, T., 2005. A physicogenetic method to assign ligand-binding relationships between 7TM receptors. Bioorg. Med. Chem. Lett. 15, 3707–3712.

Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A., Henrick, K., 2005. MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. Proteins 58, 190–199.

Halperin, I., Ma, B., Wolfson, H., Nussinov, R., 2002. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 47, 409–443.

He, W., Miao, F.J., Lin, D.C., Schwandner, R.T., Wang, Z., Gao, J., Chen, J.L., Tian, H., Ling, L., 2004. Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors. Nature 429, 188–193.

Hendlich, M., Bergner, A., Gunther, J., Klebe, G., 2003. Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. J. Mol. Biol. 326, 607–620.

Jambon, M., Imberty, A., Deleage, G., Geourjon, C., 2003. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 52, 137–145.

Ji, H., Leung, M., Zhang, Y., Catt, K.J., Sandberg, K., 1994. Differential structural requirements for specific binding of nonpeptide and peptide antagonists to the AT1 receptor. Identification of amino acid residues that determine binding of the antihypertensive drug losartan. J. Biol. Chem. 269, 16533–16536.

Kellenberger, E., Rodrigo, J., Muller, P., Rognan, D., 2004. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57, 225–242.

Kitajima, K., Ahmad, S., Selvaraj, S., Kubodera, H., Sunada, S., An, J., Sarai, A., 2002. Development of a protein–ligand interaction database, ProLINT, and its application to QSAR analysis. Genome Informat. 13, 498–499.

Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J., 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug. Discov. 3, 935–949.

Kramer, B., Rarey, M., Lengauer T, T., 1999. Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins 37, 228–241.

Laskowski, R.A., Chistyakov, V.V., Thornton, J.M., 2005. PDBsum: summaries and analyses of PDB structures. Nucl. Acids Res. D26, 221–222.

Lichtarge, O., Bourne, H., Cohen, F., 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257, 342–358.

Lipinski, C., Hopkins, A., 2004. Navigating chemical space for biology and medicine. Nature 432, 855–861.

Malherbe, P., Kratochvwil, N., Knoflach, F., Zenner, M.-T., Kew, J.N.C., Krattzeisen, C., Maerki, H.P., Adam, G., Mutel, V., 2003. Mutational analysis and molecular modeling of the allosteric binding site of a novel selective, noncompetitive antagonist of the metabotropic glutamate 1 receptor. J. Biol. Chem. 278, 8340–8347.

Nissink, J.W., Murray, C., Hartshorn, M., Verdonk, M.L., Cole, J.S., Taylor, R., 2002. A new test set for validating predictions of protein–ligand interaction. Proteins 49, 457–471.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Trong, I.L., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M., Miyano, M., 2000. Crystal structure of rhodopsin: a G protein-coupled receptor. Science 289, 739–745.

Paul, N., Bret, G., Kellenberger, E., Müller, P., Rognan, D., 2004. Recovering the true targets of specific ligands by virtual screening of the protein data bank. Proteins 54, 671–680.

Petrel, C., Kessler, A., Maslah, F., Dauban, P., Dood, R.H., Rognan, D., Ruat, M., 2003. Modeling and mutagenesis of the binding site of Calhex 231, a novel negative allosteric modulator of the extracellular Ca(2+)-sensing receptor. J. Biol. Chem. 278, 49487–49494.

Reiter, L.A., Koch, K., Piscopio, A.D., Showell, H.J., Alpert, R., et al., 1998. Trans-3-benzyl-4-hydroxy-7-chromanylbenzoic acid derivatives as antagonists of the leukotriene B4 (LTB4) receptor. Bioorg. Med. Chem. Lett. 8, 1781–1786.

Roche, O., Kiyama, R., Brooks III, C.L., 2001. Ligand–protein database: linking protein–ligand complex structures to binding data. J. Med. Chem. 44, 3592–3598.

Schoichet, B.K., 2004. Virtual screening of chemical libraries. Nature 432, 862–865.

Schwalbe, H., Wess, G., 2002. Dissecting G-protein-coupled receptors: structure, function, and ligand interaction. ChemBioChem 3, 915–919.

Stuart, C., Ilyin, V.A., Sali, A., 2002. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics 18, 200–201.

Smith, R.G., Cheng, K., Schoen, W.R., Pong, S.S., Hickey, G., et al., 1993. A non peptidyl growth hormone secretagogue. Science 260, 1640–1643.

Surgand, J.S., Rodrigo, J., Kellenberger, E., Rognan, D., 2006. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. Proteins 62, 509–538.

Varady, J., Wu, X., Fang, X., Min, J., Hu, Z., Levant, B., Wang, S., 2003. Molecular modeling of the three-dimensional structure of dopamine 3 (D3) subtype receptor: discovery of novel and potent D3 ligands through a hybrid pharmacophore- and structure-based database searching approach. J. Med. Chem. 46, 4377–4392.

Venter, J.C. et al., 2004. The sequence of the human genome. Science 291, 1304–1351.

Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., Taylor, R.D., 2003. Improved protein–ligand docking using GOLD. Proteins 52 (4), 609–623.

Weber, A., Casini, A., Heine, A., Kuhn, D., Supuran, C.T., Scozzafava, A., Klebe, G., 2004. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2 selective celecoxib: new pharmacological opportunities due to related binding site recognition. J. Med. Chem. 47, 550–557.

Wise, A., Jupe, S.C., Rees, S., 2004. The identification of ligands at Orphan G-Protein coupled receptors. Annu. Rev. Pharmacol. Toxicol. 44, 43–66.