British Journal of Pharmacology (2007), 1–15
© 2007 Nature Publishing Group All rights reserved 0007–1188/07 $30.00
www.brjpharmacol.org

REVIEW

Chemogenomic approaches to rational drug design

D Rognan

Bioinformatics of the Drug, CNRS UMR 7175-LC1, Illkirch, France

Paradigms in drug design and discovery are changing at a significant pace. Concomitant to the sequencing of over 180 several genomes, the high-throughput miniaturization of chemical synthesis and biological evaluation of a multiple compounds on gene/protein expression and function opens the way to global drug-discovery approaches, no more focused on a single target but on an entire family of related proteins or on a full metabolic pathway. Chemogenomics is this emerging research field aimed at systematically studying the biological effect of a wide array of small molecular-weight ligands on a wide array of macromolecular targets. Since the quantity of existing data (compounds, targets and assays) and of produced information (gene/protein expression levels and binding constants) are too large for manual manipulation, information technologies play a crucial role in planning, analysing and predicting chemogenomic data. The present review will focus on predictive *in silico* chemogenomic approaches to foster rational drug design and derive information from the simultaneous biological evaluation of multiple compounds on multiple targets. State-of-the-art methods for navigating in either ligand or target space will be presented and concrete drug design applications will be mentioned.

British Journal of Pharmacology advance online publication, 29 May 2007; doi:10.1038/sj.bjp.0707307

Keywords: chemogenomics; ligand; target; virtual screening

Abbreviations: 1-D, one-dimensional; 2-D, two-dimensional; 3-D, three-dimensional; ADMET, absorption, distribution, metabolism, excretion and toxicity; GPCR, G protein-coupled receptor; NHR, nuclear hormone receptor; QSPR, quantitative structure–property relationships; QSAR, quantitative structure–activity relationships

Introduction

Until the recent sequencing of the human genome (Lander *et al.*, 2001; Venter *et al.*, 2001), drug discovery has long been a multidisciplinary effort to optimize ligands properties (potency, selectivity, pharmacokinetics) towards a single macromolecular target. It is estimated that, out of the 20–25 000 human genes supposed to encode for ca. 3000 druggable targets (Russ and Lampel, 2005), only a subset of that pharmacological space (ca. 800 proteins) has currently been investigated by the pharmaceutical industry (Paolini *et al.*, 2006). Remarkably, medicinal chemistry followed a parallel boost with the miniaturization and parallelization of compound synthesis, such that over 10 million non-redundant chemical structures covers the actual chemical space, out of which ca. 1000 have been approved as drugs. Therefore, only a small fraction of compounds describing the current chemical space has been tested on a fraction of the entire target space. Chemogenomics is the new interdisciplinary field, which attempts to fully match target and ligand space, and ultimately identify all ligands of all targets (Caron *et al.*, 2001). Various definitions of overlapping fields (chemical genetics, chemical genomics) have been proposed. We will herein consider a broad definition of chemogenomics encompassing chemoproteomics, namely the study of small-molecular-weight drug candidates on gene/protein function. From the definition of the field, one easily understands that chemogenomics will be at the interface of chemistry, biology and consequently informatics since data mining is required to extract reliable information. Furthermore, methodologies at the border of chemistry and biology (medicinal chemistry), chemistry and informatics (chemoinformatics), biology and informatics (bioinformatics) will also play a major role in bringing these major disciplines together. Chemogenomic approaches to drug discovery rely on at least three components, each necessitating hard experimental work: (1) a compound library, (2) a representative biological system (target library, single cell and whole organism), and (3) a reliable readout (for example, gene/protein expression, high-throughput binding or functional assay). By definition, analysing chemogenomic data is a never-ending learning process aimed at completing a two-dimensional (2-D) matrix, where targets/genes are usually reported as columns and compounds as rows, and where reported values are usually binding constants (K_i, IC_{50}) or functional effects (for example, EC_{50}). This matrix is sparse as far as all possible compounds have not been tested on all

Correspondence: Dr D Rognan, Bioinformatics of the Drug, CNRS UMR 7175-LC1, F-67400 Illkirch, France.
E-mail: didier.rognan@pharma.u-strasbg.fr

Received 15 January 2007; accepted 18 March 2007

possible genes/proteins. Predictive chemogenomics will thus attempt to fill existing holes by predicting compounds–genes/proteins relationships. *In silico* approaches to predict such data (target selectivity for various ligands and ligand selectivity for various targets) will span pure ligand-based approaches (comparison of known ligands to predict their most probable targets), pure target-based approaches (comparison of targets or ligand-binding sites to predict their most likely ligands) or ultimately target-ligand based approaches (using experimental and predicted binding affinity matrices).

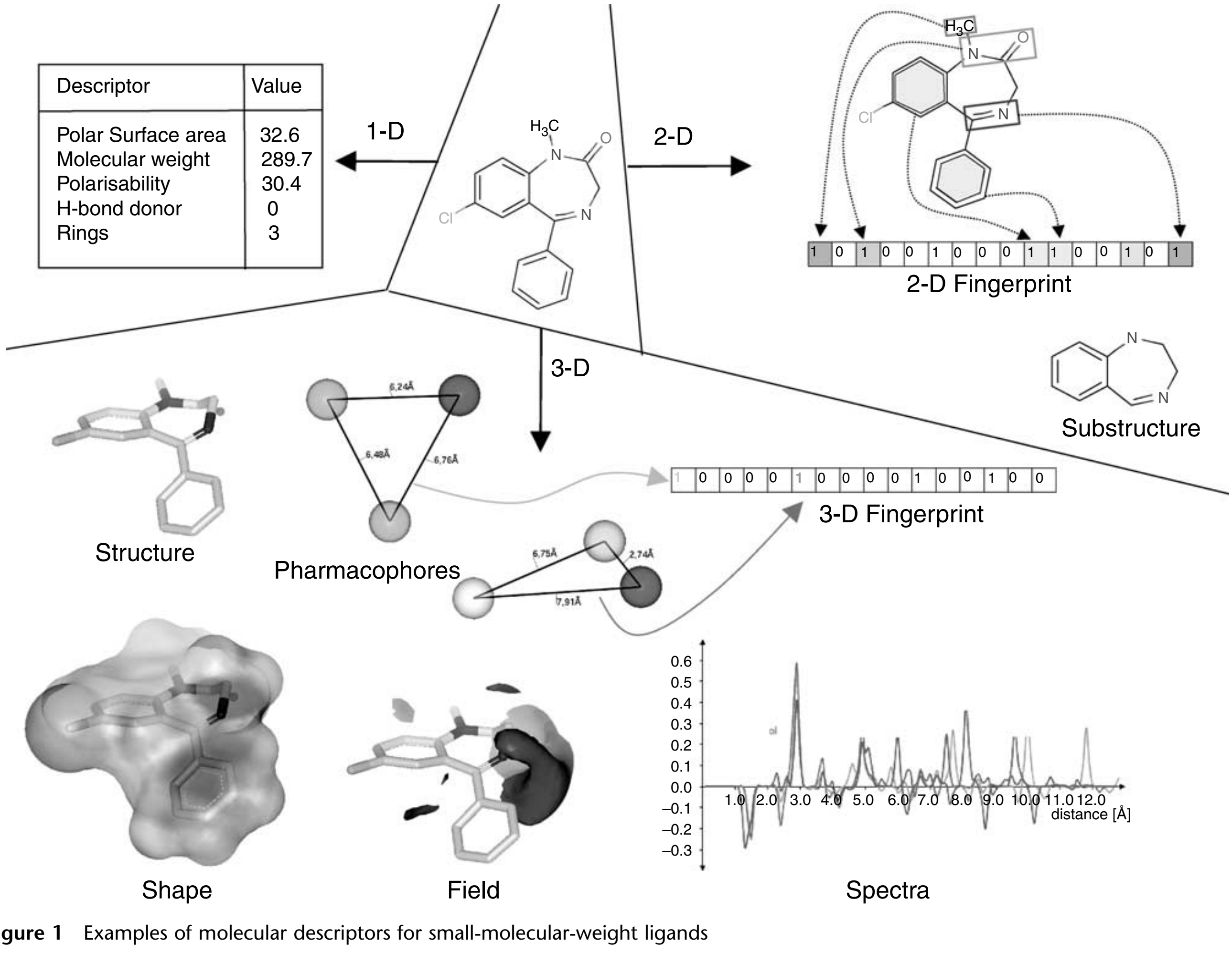
Description of ligand and target spaces

Basic assumptions of any chemogenomic-based approach are twofold: (i) compounds sharing some chemical similarity should also share targets and (ii) targets sharing similar ligands should share similar patterns (binding sites). Filling the full theoretical chemogenomic matrix thus implies that data on ‘unliganded’ targets should be gathered from the closest ‘liganded’ neighbouring targets, and that data on ‘untargeted’ ligands should be gathered from the closest ‘targeted’ ligands. The true question is how to measure distances between two ligands or two targets.

Ligand space

To efficiently navigate in ligand space, one first needs to describe the compound using appropriate properties (descriptors) and then to use a master equation to measure a distance between two compounds (similarity metric).

Descriptors are usually classified according to their dimensionality ranging from one dimensional (1-D) to three-dimensional (3-D) properties (Bender and Glen, 2004) (Figure 1 and Table 1) 1-D descriptors are easy and fast to compute. They describe global properties (for example, molecular weight, atom and bond counts), which can be derived from the chemical formulae and which are used in combination to predict absorption, distribution, metabolism, excretion and toxicity properties such as aqueous solubility (Votano *et al.*, 2004), 1-octanol-water partition coefficient (Clark, 2005), plasma protein binding or bioavailability (Wang *et al.*, 2006), but also to classify compounds (for example, drugs vs nondrugs (Sadowski and Kubinyi, 1998)) or ligands from various target families (Morphy, 2006) by linear or nonlinear quantitative structure–activity relationships/quantitative structure–property relationships (QSAR/QSPR) methods. To fasten comparisons, 1-D linear representations of compounds are often used. The most popular of this kind of simplified string is the ‘Simplified Molecular Input Line Entry System’ or SMILES (Weininger, 1988).



Descriptor	Value
Polar Surface area	32.6
Molecular weight	289.7
Polarisability	30.4
H-bond donor	0
Rings	3

Figure 1 Examples of molecular descriptors for small-molecular-weight ligands

Table 1 Ligand descriptors

Dimension	*Nature*	*Examples*
1-D	Global	Molecular weight, atom and bound counts (for example, number of H-bond donors, number of rings), polar surface area, polarizability, log *P*)
2-D	Topological	Topological and connectivity indices, fragments, substructures (for example, maximum common substructures), topological fingerprints (for example, structural keys)
3-D	Conformational	*n*-points pharmacophore, shape, field, spectra and fingerprints

Table 2 Structural classification of proteins

Dimension	*Classification scheme*	*Databases*
1-D	Sequence Patterns	UniProt (Wu *et al.*, 2006) and Pfam (Finn *et al.*, 2006) PRINTS (Attwood *et al.*, 2003) and PROSITE (Hulo *et al.*, 2006)
2-D	Secondary structure fold	SCOP (Casbon and Saqi, 2005) and CATH (Reeves *et al.*, 2006)
3-D	Atomic coordinates binding site	PDB (Berman *et al.*, 2000) and MODBASE (Pieper *et al.*, 2006) Binding MOAD (Hu *et al.*, 2005) and sc-PDB (Kellenberger *et al.*, 2006)

Abbreviations: MOAD, Mother of All Databases; UniProt, The Universal Protein Resource.

Most ligand descriptors range in the family of 2-D topological descriptors, where the connectivity table (list of atoms and bonds) is parsed to encode both atomic and bond properties. The most intuitive way to represent this kind of information is the 2-D sketch of the structure (Figure 1), which enables to browse a ligand library for compounds sharing a particular 2-D motif (fragment, substructure). Graph-based methods which transforms the 2-D structure into a molecular graph (atoms being the nodes) are relatively popular for substructure search and clustering chemical compounds into subfamilies (Raymond *et al.*, 2003), but present the noticeable disadvantage to be computationally slow. Much faster are fingerprint-based methods (Willett, 2006), where the occurrence of predefined structural events (atoms, fragments, rings, substructures and 2-D pharmacophores) are encoded into bit strings (sequence of '0' and '1' digits) called 'fingerprints' which ere easy to derive, handle and compare. Although receptor-ligand recognition is a 3-D event, 2-D fingerprints have been found repeatedly more appropriate true 3-D fingerprints for similarity searches (Sheridan and Kearsley, 2002). Latter descriptors encode conformation-specific properties (atomic coordinates, 3-D pharmacophores, shapes, potentials, fields, spectra; Table 1), and therefore usually necessitate a common alignment of molecules to be compared in the same 3-D Cartesian space (especially if grid-based fields or potentials have to be compared) and a relevant sampling of conformational space accessible to each ligand. To avoid the alignment step which may cause false positives in a virtual screen, 3-D information can be translated into a bit string, which stores the occurrence of all possible pharmacophore tuplets (doublets, triplets and quadruplets) with their corresponding features (for example, H-bond acceptor, positively ionisable atom, and so on) and interfeature distances. Hence, comparing bit strings is much easier than comparing structures. Most similarity searches prefer a binary representation of 2- or 3-D properties to derive simple similarity indices, the most popular being the Tanimoto coefficient (Equation 1)

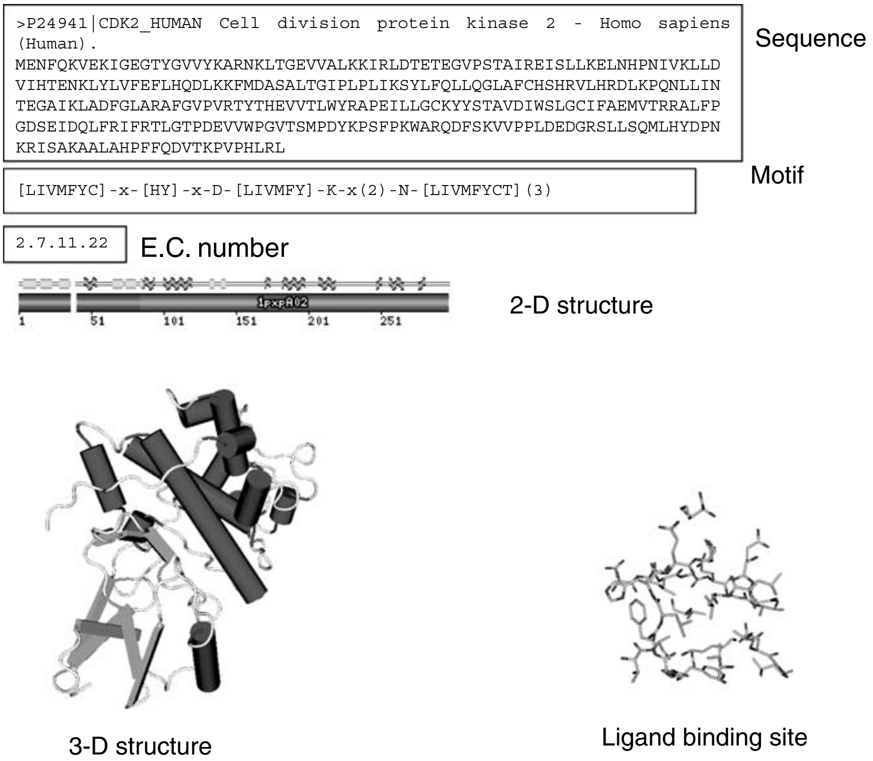
$$T_c = \frac{c}{a + b - c} \tag{1}$$

a is the count of bits on in compound A, b is the count of bits on in compound B and c is the count of the bits on in both compound A and B.

The Tanimoto coefficient will thus range from 0 for two completely dissimilar structures to 1 for two identical compounds.

Target space

Proteins are commonly classified according to their sequence and structure (Table 2). The full amino-acid sequence is the very first interesting information (Figure 2), which already enables a reliable clustering of targets by family (for example, G protein-coupled receptors (GPCRs) and kinases). However, sequence lengths may considerably vary within a protein family (for example, sequence lengths of human GPCRs range from 290 to 6200 residues), such that analysing similarities and differences first requires an alignment of amino-acid sequences which can be tricky in case of large insertions/deletions. Therefore, one may focus on specific motifs (Attwood *et al.*, 2003), which are a collection of continuous residues specific of a protein family (for example, DRY motif in TM III of rhodopsin-like GPCRs). To take into account the structural organization of the target, it can be of interest to look at the 2-D structure (mapping of α-helices, β-sheets, coils and random structures) and even better at the 3-D structure (atomic coordinates provided by X-ray diffraction, NMR or molecular modelling) and/or the corresponding fold. In chemogenomics-related approaches, one usually focuses on the ligand-binding site, where structural similarities among related targets are usually much higher than when considering the full 1-D sequence or 3-D structure.



```
>P24941|CDK2_HUMAN  Cell  division  protein  kinase  2  -  Homo  sapiens
(Human).
MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREISLLKELNHPNIVKLLD
VIHTENKLYLVFEFLHQDLKKFMDASALTGIPLPLIKSYLFQLLQGLAFCHSHRVLHRDLKPQNLLIN
TEGAIKLADFGLARAFGVPVRTYTHEVVTLWYRAPEILLGCKYYSTAVDIWSLGCIFAEMVTRRALFP
GDSEIDQLFRIFRTLGTPDEVVWPGVTSMPDYKPSFPKWARQDFSKVVPPLDEDGRSLLSQMLHYDPN
KRISAKAALAHPFFQDVTKPVPHLRL
```

```
[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3)
```

```
2.7.11.22
```

Figure 2 Various representations of a protein using 1-D to 3-D properties.

Targets may also be classified according to their pharmacological profile (binding affinity for a panel of ligands) which means according to the nature of ligands they recognize (Paolini *et al.*, 2006). Of course, there is a considerable overlap between sequence- and ligand-based classifications, since ligands generally bind to a subset of the protein universe. However, relationships across protein subfamilies are particularly interesting in drug design for predicting/modifying the pharmacological profile of a drug.

Target–ligand space

It is possible to directly navigate in the protein–ligand space by browsing full matrices in which either affinity or structural information is stored. Experimental evaluation of x compounds on y targets (for example, *in vitro* binding affinity assay) leads to a matrix of xy numbers (for example, IC_{50} values), which can be used to predict the affinity of a new compound to an existing target by multivariate linear regression (Kauvar *et al.*, 1995), measure a structure–activity relationships distance between two targets (Vieth *et al.*, 2004) and predict a global pharmacological profile (Krejsa *et al.*, 2003). A clear advantage of this approach is that it relies on true binding affinity values and that experimentally derived descriptors will usually outperform computed descriptors. A clear drawback is the enormous amount of data required to derive true information such that similar approaches are not realistic, for example, in an academic environment. Therefore, one might substitute experimental with predicted affinities derived from either docking or 3-D QSAR approaches (Matter and Schwab, 1999; Fukunishi *et al.*, 2006), although extrapolation will be limited here in a tiny protein space. Since binding free energy is extremely difficult to predict, replacing affinity by molecular interaction descriptors is possible. Of particular interest are structural interaction fingerprints (IFPs) (Singh *et al.*, 2006), which converts atomic coordinates of a protein–ligand complex into a bit string featuring for each residue of a binding site, the type of molecular interactions (for example, H-bond, aromatic interaction, hydrophobic contact) developed by a co-crystallized or docked ligand. Comparing a series of complexes between n ligands and a single protein or between one ligand and n-related proteins is then performed as for ligands by computing distances between 1-D IFPs (Figure 3).

Ligand-based chemogenomic approaches

Annotating ligand libraries

The basic paradigm underlying ligand-based chemogenomic approaches is that molecules sharing enough similarity to existing biologically annotated ligands have enhanced probability to share the same biological profile (Figure 4). It is therefore very important to annotate chemical libraries with biological information (targets, *in vitro* affinity data and ADMET properties). Over recent years, there has been a huge effort mainly from small biotech companies to compile such data by an exhaustive survey of literature and patent data (Table 3). Since chemogenomic approaches usually focus on target families, most of these archives are related to the most pharmaceutically important target families (GPCRs,

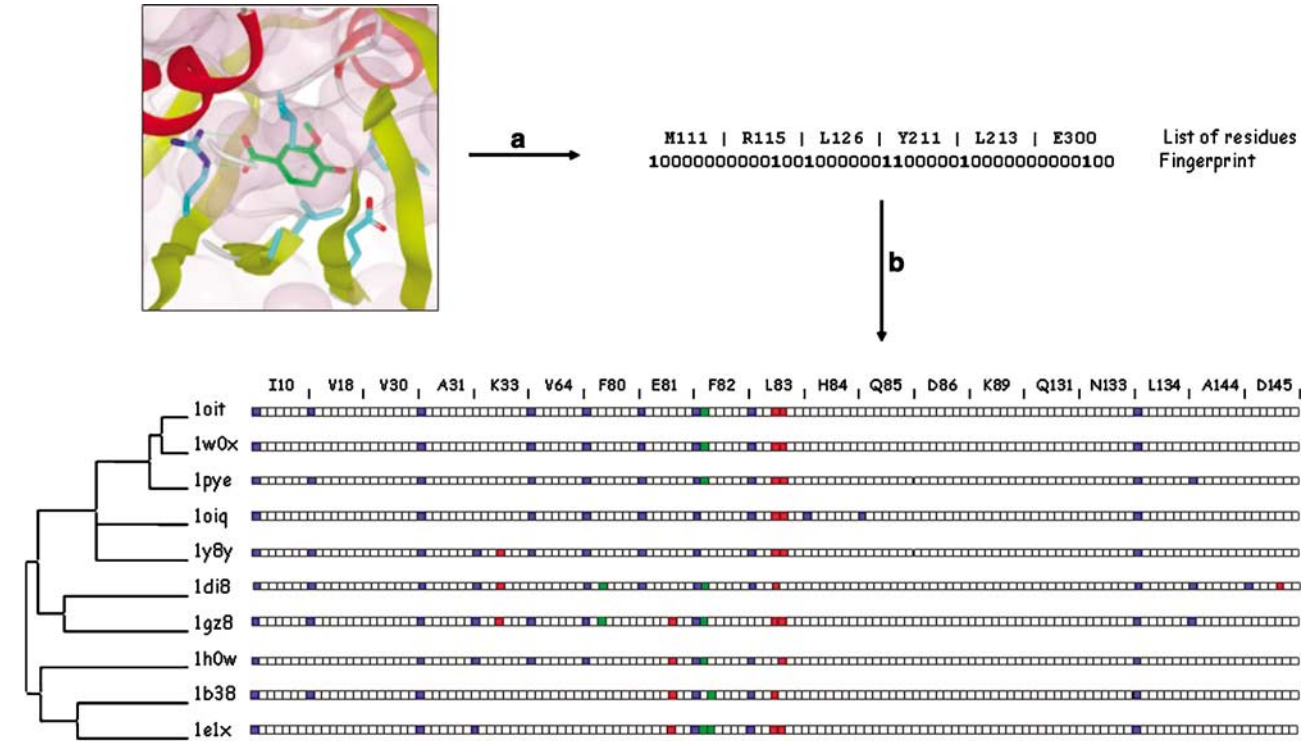


Figure 3 (**a**) Deriving and (**b**) comparing protein–ligand complexes by molecular interaction fingerprints. '0' and '1' digits are replaced by colour-coded squares for the ease of comparison (blue, hydrophobic interactions; green, aromatic interactions; red, hydrogen bonds).

kinases, nuclear hormone receptors (NHRs), proteases and phosphodiesterases).

A good example has been provided by Novartis scientists (Schuffenhauer *et al.*, 2003) who linked chemical space to target space by merging fields from separate chemical and biological databases to provide a unified and searchable chemogenomic database. On one hand, over 110 000 pharmaceutical ligands were gathered from the MDL Drug Data Report (Table 3). On the other hand, annotation of targets was based on existing classifications for enzymes and receptors. Linking MDDR 'activity keys' to the target classification scheme enabled the annotation of 53 000 compounds totalling 799 different activity keys and related targets. Since the target's sequence is linkable to the ligand, sequence-based similarity searches of ligands for protein homologues of liganded targets are therefore feasible. Annotated reference ligands for a particular GPCR were used as starting points to recover either new receptor ligands or ligands of receptors close to the reference GPCR. Interestingly, the efficiency of the virtual screening approach was dependent on the phylogenetic distance between the reference and the query targets. Another straightforward application of biologically annotated compound libraries is the design of target-directed combinatorial libraries (Savchuk *et al.*, 2004) focusing on chemotypes preferred by a family of targets.

Natural products also cover a very interesting chemical space of biological relevance because of the evolutionary pressure put on these compounds to bind, usually through highly specific mechanisms, to particular targets. The chemical space spanned by biologically annotated natural products was described recently as a structural and hierarchical scaffold tree (Koch *et al.*, 2005), which can be browsed to design natural product-oriented chemical libraries.

Biologically annotated compound libraries are a direct source of potentially new biological mechanisms to correct a phenotype. Root *et al.* (2003) designed a library of 2036 biologically active compounds covering 169 different biochemical mechanisms, which was shown to be structurally diverse and able to provide 85 hits in a cell viability and proliferation assay. Among the 85 hits, 27 were supposed to be active by new biochemical mechanisms.

Privileged structures

The term 'privileged structure' was first coined by Evans *et al.* (1988), who noticed the promiscuity of the 1,4-benzodiazepine scaffold for various targets (Figure 5). A privileged structure is defined as 'a substructure/scaffold exhibiting strong preferences for a particular area of the target space (for example, GPCRs) and suitable to orient the design of targeted compound libraries' (Klabunde and Hessler, 2002). In fact, a recent and deeper analysis of drug-like ligands show that privilege only appears upon a certain level of chemical functionalization of the scaffold (Schnur *et al.*, 2006). For example, the biphenyl substructure is not a privileged structure but a simple protein-binding motif, since it occurs in a wide array of protein ligands with no particular preference for a certain target family. However, extending the biphenyl motif to a 2-tetrazolo-biphenyl dramatically

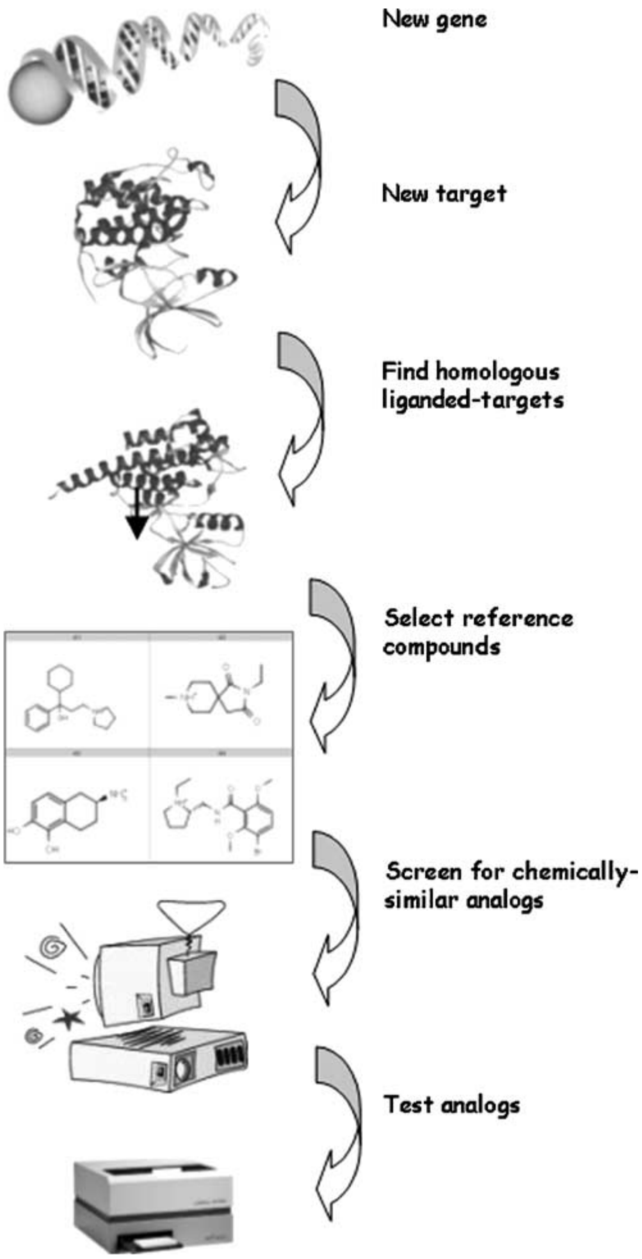

New gene
New target
Find homologous liganded-targets
Select reference compounds
Screen for chemically-similar analogs
Test analogs

Figure 4 Structure–activity relationship homology flowchart.

enhance the specificity of the latter substructure for GPCRs (Schnur *et al*., 2006). Remarkably, many substructures apparently have corresponding binding sites in unrelated target families (for example, GPCRs, kinases, ion channels, proteases, nuclear hormone receptors). Only a few of them (see an example in Figure 6) are really selective for a certain target family (Schnur *et al*., 2006). A main reason for this exquisite specificity is that specific binding sites for peculiar substructure have been conserved along the evolution of target subfamilies (Bondensgaard *et al*., 2004; Surgand *et al*., 2006). Family-specific privileged structures are of prime importance to design targeted libraries and enhance hit rates when a protein from the targeted family is screened experimentally. A nice application of designing targeted libraries was presented by Amgen (Xia *et al*., 2004). After training a machine-learning algorithm to distinguish true kinase inhibitors from non-kinase inhibitors, multiple chemotypes could be selected to design a kinase-targeted library, which yielded high enrichment in true inhibitors in subsequent kinase inhibition assays.

Ligand-based in silico *screening*

Main target families can be distinguished by a simple look at physicochemical properties (molecular weight, log*P*, polar surface area, H-bond donor and acceptor counts) of their cognate ligands (Morphy, 2006). One can thus easily imagine that more sophisticated descriptors can be used to predict a global target profile for any given compound, provided that targets to be predicted are sufficiently well described by existing ligands. Ligand-based *in silico* approaches to target fishing begin to appear in the literature (Cases *et al*., 2005; Bender *et al*., 2006; Bhavani *et al*., 2006; Mestres *et al*., 2006; Nettles *et al*., 2006; Nidhi *et al*., 2006; Steindl *et al*., 2006). They all share three basic components: (i) a set of reference compounds from which 2-D (scaffold, substructure, fingerprints) or 3-D descriptors (pharmacophore) are stored in a database, (ii) a screening procedure using either QSAR, machine learning (Bayesian classification, support vector machines) or pharmacophore searches and (iii) a screening collection to identify using above-described descriptors new molecules likely to share the same target or target profile than reference compounds (Figure 7).

Mestres *et al*. (Cases *et al*., 2005; Mestres *et al*., 2006) have annotated a library of molecules targeting NHRs. Using a hierarchical classification for 2000 ligands and 25 receptors, chemogenomic links bridging ligand to target space can be easily recovered to distinguish selective from promiscuous scaffolds. Using Shannon Entropy descriptors (SHED) based on the distribution of atom-centred feature pairs, any compound collection can be screened to identify hits presenting SHED distances to a reference NHR ligand beyond a defined threshold and therefore likely to share the same NHR profile.

Novartis successfully applied a machine-learning algorithm using Bayesian statistics (Xia *et al*., 2004) to predict target profiles from extended connectivity fingerprints of compounds from the biologically annotated Wombat database (Nidhi *et al*., 2006). For each activity class (target), a separate Bayesian model is trained to distinguish known actives from known inactives. Predicting the most likely targets of compounds in the test set is then operated by predicting the probability of each test compound to be a ligand of each of the targets. On average, the correct target was found 77% of the time when training with Wombat compounds and testing molecules from another dataset (MDDR) over 10 different activity classes (Nidhi *et al*., 2006). A significant improvement in the predictions is observed when considering, instead of a series of individual probabilities, the global profile of all training compounds in which all target-associated probabilities are concatenated into a 'Bayes affinity fingerprint' (Bender *et al*., 2006). Other 2- and 3-D descriptors have been assessed for the same application. 2-D descriptors were found to be more predictive with regard to correct target prediction than a pure 3-D pharmacophoric

Table 3 Biologically annotated compound libraries

Database	*Description*	*Website*
AurSCOPE	Target family-oriented knowledge database containing pharmacological and pharmacokinetical data for 160 000 GPCR ligands and 77 000 kinase inhibitors	http://www.aureus-pharma.com
Bioprint	Biological profile (*in vitro* and clinical data) of 2400 small-molecular-weight drugs and drug-like compounds	http://www.cerep.fr/
ChemBank	Storage of 50 000 compounds and related biological properties in 441 high-throughput screening and small molecule microarray assays	http://chembank.broad.harvard.edu/
ChemBioBase	Target centric ligand databases (GPCRs, kinases, PDE)	http://www.jubilantbiosys.com/
Kinase knowledge base	kinase structure–activity and chemical synthesis data	http://www.eidogen-sertanty.com/
MDL Drug Data Report	132 000 biologically relevant compounds and well-defined derivatives	http://www.mdli.com/
MedChem database	650 000 compounds with biological and pharmacological information	http://www.gvkbio.com
StARLITe	Highly curated target-compound SAR relationships	http://www.inpharmatica.co.uk/
Wombat	154 236 entries over 307 700 biological activities on 1320 unique targets	http://sunsetmolecular.com/

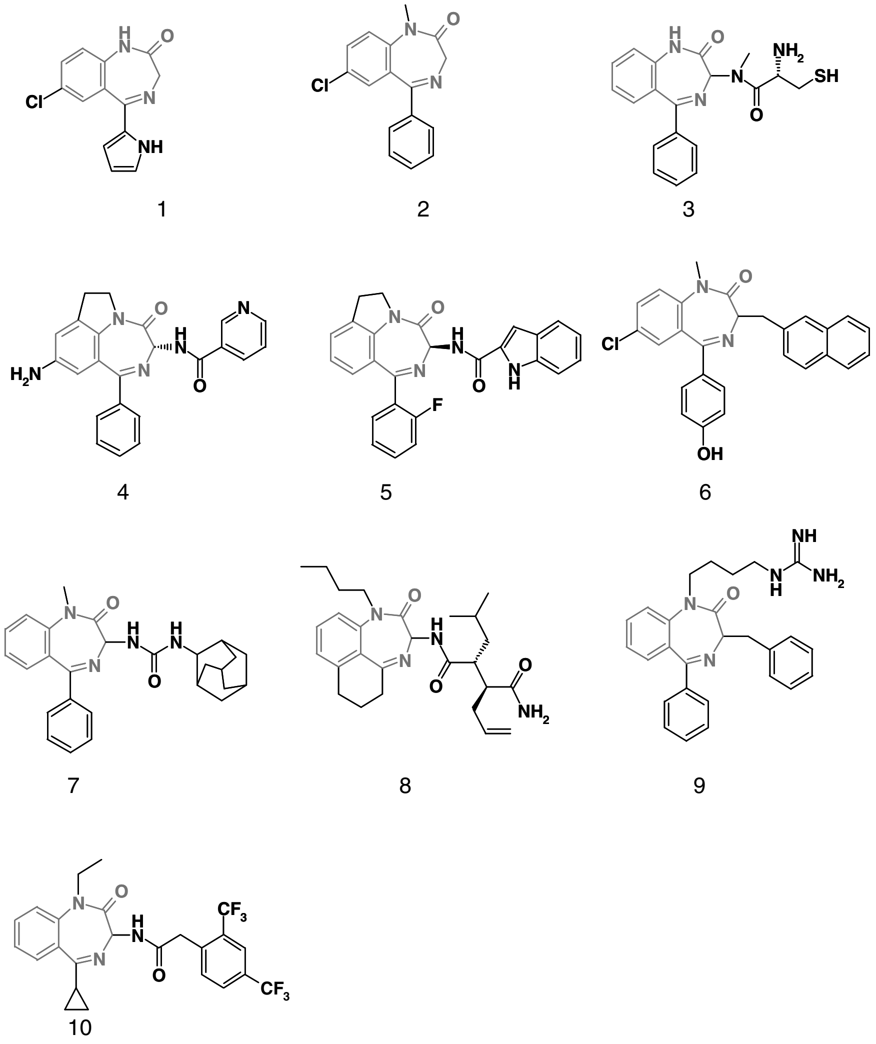


Figure 5 Permissivity of the 3H-1,4-benzodiazepin-2-one scaffold (in gray) across various targets. 1: Ro-5–3335, HIV-1 Tat inhibitor; 2: Diazepam, GABA-A receptor ligand; 3: 231023: farnesyltransferase inhibitor; 4: CI-1044, phosphodiesterase 4 inhibitor; 5: pranazepide, cholescystokinine (CCK) receptor antagonist; 6: BZ-423, F1F0 ATPase inhibitor, 7:171644, oxytocin receptor antagonist, 8: 309060: β-γ secretase inhibitor; 9: 278588: Stat5 agonist; 10: 276345: KV_s channel blocker.

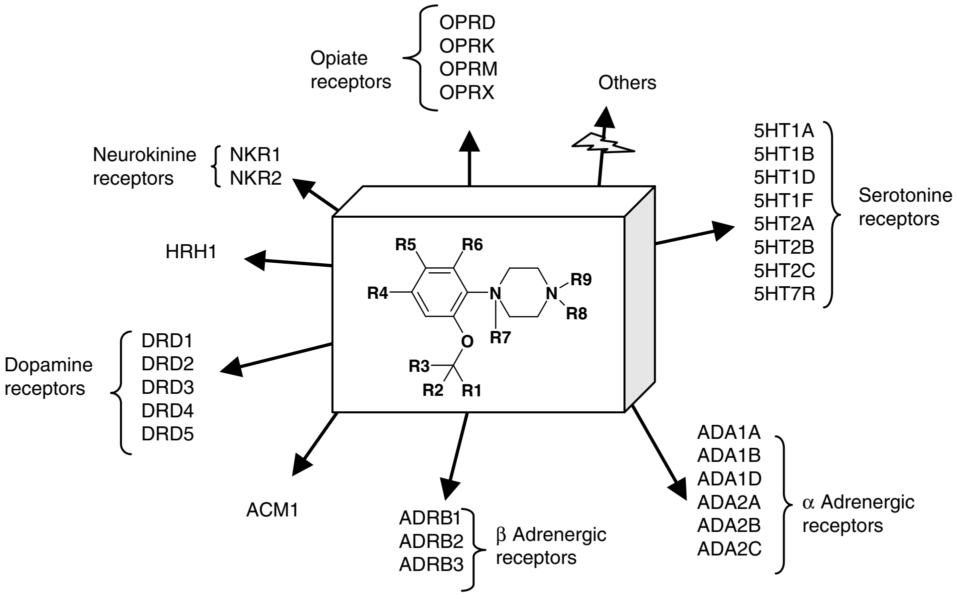
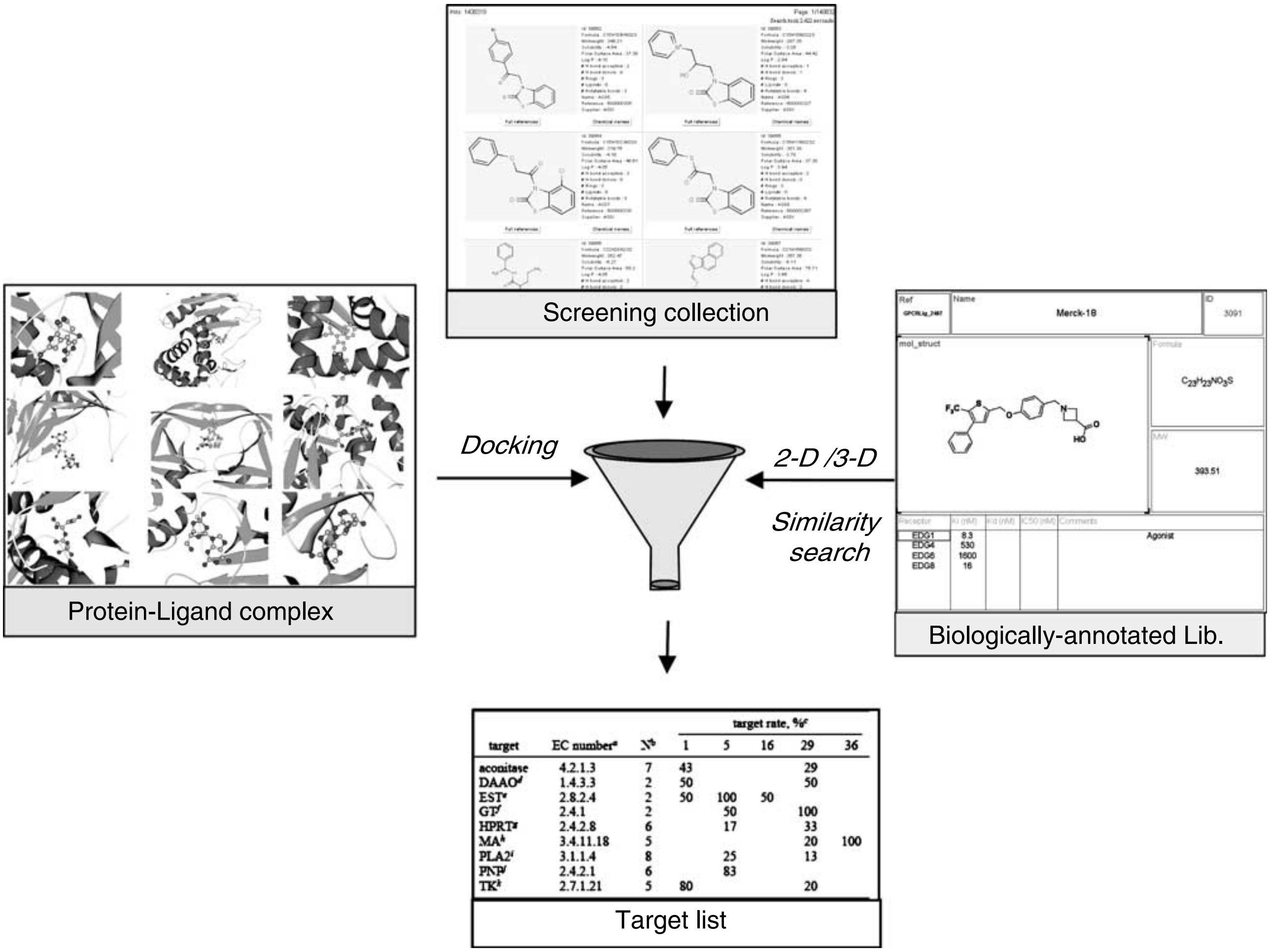


Figure 6 Human GPCRs targeted by the orthoalkoxy-*N*-phenylpiperazine privileged structure (http://bioinfo-pharma.u-strasbg.fr/hGPCRLig/). Remarkably, no other proteins ever co-crystallized with drug-like compounds are able to recognize this substructure (http://bioinfo-pharma.u-strasbg.fr/scPDB/).



Ref	Name	ID
GPCRLig_2487	Merck-18	3091

mol_struct

Formula: $C_{23}H_{23}NO_3S$

MW: 393.51

Receptor	Ki (nM)	Kd (nM)	IC50 (nM)	Comments
EDG1	8.3			Agonist
EDG4	530			
EDG6	1600			
EDG8	16			

			target rate, %[c]				
target	EC number[a]	N[b]	1	5	16	29	36
aconitase	4.2.1.3	7	43			29	
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		50		100	
HPRT[g]	2.4.2.8	6		17		33	
MA[h]	3.4.11.18	5				20	100
PLA2[i]	3.1.1.4	8		25		13	
PNP[j]	2.4.2.1	6		83			
TK[k]	2.7.1.21	5	80			20	

Figure 7 *In silico* target fishing approaches.

approach for test compounds structurally similar to those in the training set. For singletons (compounds exhibiting no strong similarity to molecules of the training set), the 3-D descriptor is more predictive.

In all these approaches, one must first automatically categorize compounds from the training set according to their molecular target without checking whether each compound really bind to its target, where (which binding site) and how (agonist or antagonist for receptor ligands). There is therefore a risk to train a machine-learning algorithm with incorrect data and to generate false rules. To overcome this drawback, more accurate but slower 3-D strategies are possible. Among them, a promising approach is to derive 3-D pharmacophores from protein–ligand complexes for which experimentally determined atomic coordinates and pharmacological activities exist (Steindl *et al.*, 2006). The target-annotated pharmacophore database can be browsed to identify target(s) of new compounds by a classical pharmacophore search. The advantage of the method relies on the higher quality of the reference dataset, but is nevertheless limited by the pharmacophore generation step and the still limited chemical diversity observed among protein data bank (PDB) ligands (Kellenberger *et al.*, 2006). For example, membrane receptors (for example, GPCRs, ion channels) cannot be predicted by this approach, crystallographic data being very sparse for these protein families, although homology model-based pharmacophores may be theoretically derived.

Target-based chemogenomic approaches

Controlling the selectivity of ligands towards related targets from the same family is crucial information in early drug-discovery stages. There is therefore a growing interest in comparing all targets from the same family especially those for which there is enough structural data (X-ray or NMR structures) to enable a proteome-wide comparative modelling of targets of still unknown structure (for example, protein kinases). Target-based chemogenomic approaches can be classified in two categories depending on whether the amino-acid sequence or the 3-D structure of targets is compared.

Sequence-based comparisons

Sequence-based approaches are intended to be used for any kind of target family, provided that a multiple alignment of all targets to compare is reachable. They are generally used for target families where a lack of high-resolution structural data hampers target comparison. GPCRs constitute an ideal framework for sequence-based comparisons (Crossley, 2004; Frimurer *et al.*, 2005; Kratochwil *et al.*, 2005; Surgand *et al.*, 2006), because it is a very important target family for drug design and only one member of this family (bovine rhodopsin) has been crystallized to date (Palczewski *et al.*, 2000). After aligning all sequences, key residues supposed to map the binding site of most non-peptide ligands can be extracted and concatenated into an ungapped sequence of a few residues (Figure 8), which can be later used to derive a distance matrix based on sequence identity (Surgand *et al.*, 2006), sequence similarity (Kratochwil *et al.*, 2005) or physicochemical properties (Frimurer *et al.*, 2005). An exhaustive cavity-based clustering of 372 human GPCRs has recently been proposed using such a strategy (Surgand *et al.*, 2006). Interestingly, it reproduces perfectly the full sequence-based tree suggesting that only a few residues are really important when comparing targets across a family. This simplification enables a much simpler analysis of features (binding site regions), which are responsible for selective or permissive ligand binding by simply looking at residue conservation (Crossley, 2004; Surgand *et al.*, 2006). There are several potential applications of cavity-based trees in drug discovery. A simple one consists in target hopping, which means discovering receptor ligands for a particular receptor by considering first the known ligands of closely related receptors. For example, CRTH2 receptor antagonists could have been identified from existing angiotensin II type 1 receptor antagonists (Frimurer *et al.*, 2005), because both receptors were found close in the GPCR cavity-biased tree. In addition, the design of targeted libraries towards a particular area of the tree is facilitated by addressing those residues responsible for selectivity/promiscuity (Frimurer *et al.*, 2005; Kratochwil *et al.*, 2005).

Structure-based comparisons

Structure-based comparisons are only possible for target families where there are enough good structural templates (X-ray structures) to afford the homology modelling of other related targets. In general, only ligand-binding sites (Hu *et al.*, 2005; Kellenberger *et al.*, 2006) are compared, since the basic aim of such comparisons is to understand the selectivity/permissivity features of related targets of known ligands.

A first possible strategy is to compare computed molecular interaction fields from the cavities to compare (Naumann and Matter, 2002; Hoppe *et al.*, 2006; Pirard and Matter, 2006). Starting from a structural alignment of all targets, interaction energies generated by rolling several probe atoms (for example, sp3 carbon atom) at each point of 3-D grid encompassing the ligand-binding site are then concatenated into a MIF vector, which can be placed in a global matrix where rows describe targets and columns interaction energies at a given 3-D grid point (Figure 9) Comparing the MIFs and clustering the cognate targets can be done either by analysing the matrix by principal component analysis (Naumann and Matter, 2002; Pirard and Matter, 2006) or by calculating a MIF distance, which is later transformed in a target tree (Hoppe *et al.*, 2006). A clear issue with this approach is that the comparison is highly dependent on the structural alignment, the grid resolution and the choice of the probe atoms. Moreover, it cannot be applied to targets of different families. However, its has been successfully applied to protein kinases (Naumann and Matter, 2002; Hoppe *et al.*, 2006), serine proteases (Hoppe *et al.*, 2006), matrix metalloproteinases (Pirard and Matter, 2006) and nuclear hormone receptors (Hoppe *et al.*, 2006) to pinpoint cavity regions or subpockets explaining either selective or promiscuous ligand binding, and thus to guide the design of compound libraries towards the desirable selectivity pattern.

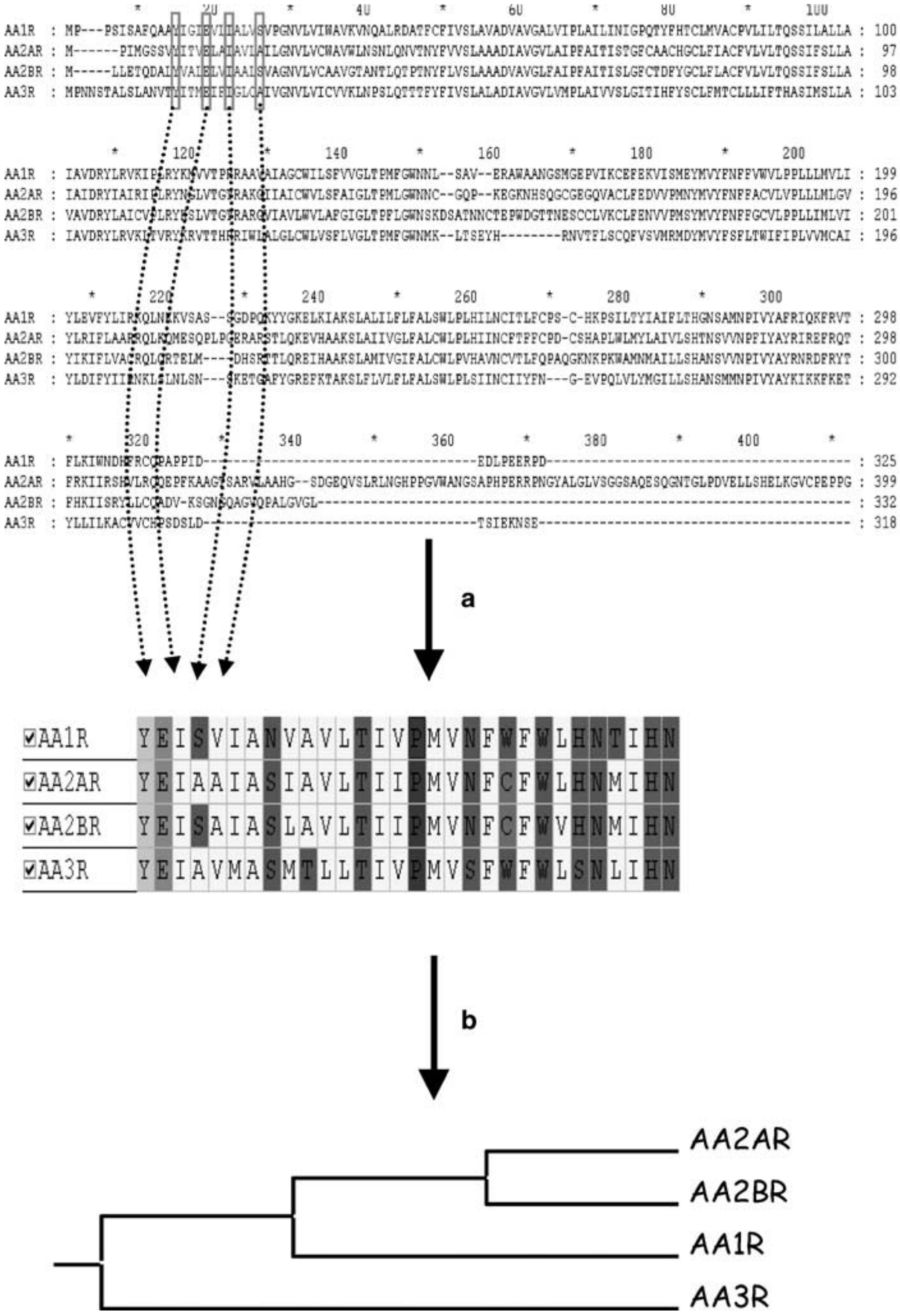

* 20 * 40 * 60 * 80 * 100
AA1R : MP---PSISAFQAAYIGIEVLIALVSVPGNVLVIWAVKVNQALRDATFCFIVSLAVADVAVGALVIPLAILINIGPQTYFHTCLMVACPVLILTQSSILALLA : 100
AA2AR : M------PIMGSSVYITVELAIAVLAILGNVLVCWAVWLNSNLQNVTNYFVVSLAAADIAVGVLAIPFAITISTGFCAACHGCLFIACFVLVLTQSSIFSLLA : 97
AA2BR : M-----LLETQDALYVALELVIAALSVAGNVLVCAAVGTANTLQTPTNYFLVSLAAADVAVGLFAIPFAITISLGFCTDFYGCLFLACFVLVLTQSSIFSLLA : 98
AA3R : MPNNSTALSLANVTYITMEIFIGLCAIVGNVLVICVVKLNPSLQTTTFYFIVSLALADIAVGVLVMPLAIVVSLGITIHFYSCLFMTCLLLIFTHASIMSLLA : 103
* 120 * 140 * 160 * 180 * 200
AA1R : IAVDRYLRVKIPLRYKMVVTPRRAAVAIAGCWILSFVVGLTPMFGWNNL--SAV--ERAWAANGSMGEPVIKCEFEKVISMEYMVYFNFFVWVLPPLLLMVLI : 199
AA2AR : IAIDRYIAIRIPLRYNGLVTGTRAKGIIAICWVLSFAIGLTPMLGWNNC--GQP--KEGKNHSQGCGEGQVACLFEDVVPMNYMVYFNFFACVLVPLLLMLGV : 196
AA2BR : VAVDRYLAICVPLRYKSLVTGTRARGVIAVLWVLAFGIGLTPFLGWNSKDSATNNCTEPWDGTTNESCCLVKCLFENVVPMSYMVYFNFFGCVLPPLLIMLVI : 201
AA3R : IAVDRYLRVKLTVRYKRVTTHRRIWLALGLCWLVSFLVGLTPMFGWNMK--LTSEYH--------RNVTFLSCQFVSVMRMDYMVYFSFLTWIFIPLVVMCAI : 196
* 220 * 240 * 260 * 280 * 300
AA1R : YLEVFYLIRKQLNKKVSAS--SGDPQKYYGKELKIAKSLALILFLFALSWLPLHILNCITLFCPS-C-HKPSILTYIAIFLTHGNSAMNPIVYAFRIQKFRVT : 298
AA2AR : YLRIFLAARRQLKQMESQPLPGERARSTLQKEVHAAKSLAIIVGLFALCWLPLHIINCFTFFCPD-CSHAPLWLMYLAIVLSHTNSVVNPFIYAYRIREFRQT : 298
AA2BR : YIKIFLVACRQLQRTELM----DHSRTTLQREIHAAKSLAMIVGIFALCWLPVHAVNCVTLFQPAQGKNKPKWAMNMAILLSHANSVVNPIVYAYRNRDFRYT : 300
AA3R : YLDIFYIIRNKLSLNLSN---SKETGAFYGREFKTAKSLFLVLFLFALSWLPLSIINCIIYFN---G-EVPQLVLYMGILLSHANSMMNPIVYAYKIKKFKET : 292
* 320 * 340 * 360 * 380 * 400 *
AA1R : FLKIWNDHFRCQPAPPID------------------------------------EDLPEERPD-- : 325
AA2AR : FRKIIRSHVLRQQEPFKAAGTSARVLAAHG--SDGEQVSLRLNGHPPGVWANGSAPHPERRPNGYALGLVSGGSAQESQGNTGLPDVELLSHELKGVCPEPPG : 399
AA2BR : FHKIISRYLLCQADV-KSGNGQAGVQPALGVGL-- : 332
AA3R : YLLILKACVVCHPSDSLD------------------------------------TSIEKNSE--- : 318
a
☑AA1R Y E I S V I A N V A V L T I V P M V N F W F W L H N T I H N
☑AA2AR Y E I A A I A S I A V L T I I P M V N F C F W L H N M I H N
☑AA2BR Y E I S A I A S L A V L T I I P M V N F C F W V H N M I H N
☑AA3R Y E I A V M A S M T L L T I V P M V S F W F W L S N L I H N
b
AA2AR
AA2BR
AA1R
AA3R

Figure 8 Sequence-based comparison of targets exemplified by human adenosine receptors (Surgand *et al.*, 2006). (**a**) Selection of key cavity-lining residues and (**b**) Clustering according to residue conservation.

To avoid the previously reported structural alignment bias, 3-D atomic protein coordinates can be directly compared to measure a distance between two targets. Global structural alignment methods (Shindyalov and Bourne, 1998; Holm and Park, 2000; Standley *et al.*, 2005) usually count the number of structurally equivalent residues by comparing overlapping sequence fragments. Such methods, however, do not work very well for discontinuous sequences (active sites) and for proteins exhibiting different folds. A second approach is to identify pre-defined structural motifs or templates (for example, Ser–His–Asp catalytic triad in serine proteases) and align a query to a reference protein by matching templates (Artymiuk *et al.*, 1994; Wallace *et al.*, 1997). However, numerous proteins (for example, kinases, GPCRs, ion channels) may share a binding site for a unique ligand (ATP) without sharing any structural template similarity. Most recent approaches to generate structural alignment describe proteins by physicochemical properties at representative locations. Molecular surfaces can be easily discretized in either chemically labelled sparse points (Rosen *et al.*, 1998) or graphs (Kinoshita and Nakamura, 2003) and therefore aligned to maximize surface overlap

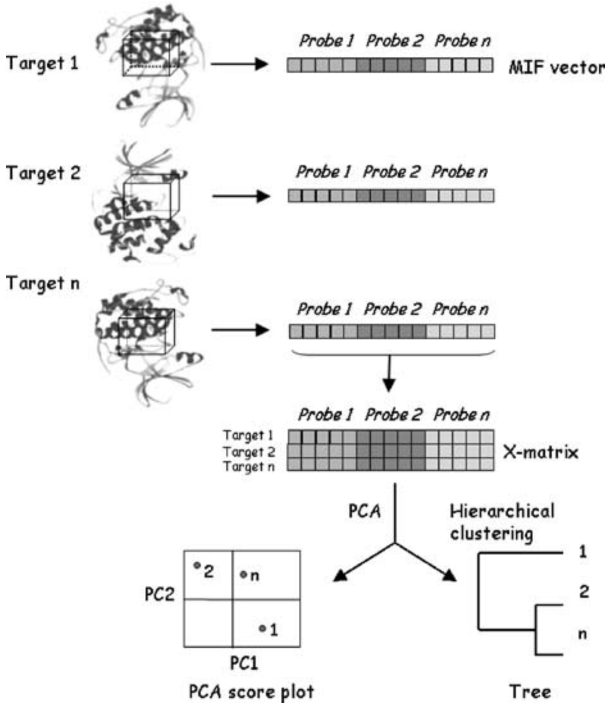


Figure 9 Molecular interaction field (MIF)-based clustering of targets.

with any reference. A database of protein surfaces (eF-site) has successfully been browsed to predict the function of a hypothetical archaeon protein (MJ0226) by detection of a mononucleotide binding site (Kinoshita and Nakamura, 2003). Surface-based comparisons are, however, relatively slow and thus incompatible with proteome-wide comparisons. Recent and faster methods (Schmitt *et al.*, 2002; Jambon *et al.*, 2003; Shulman-Peleg *et al.*, 2004; Powers *et al.*, 2006; Gold and Jackson, 2006) have been developed over the last 5 years. They all have in common to represent an active site of interest by pseudocenters (dummy atoms located along or close to every side chain of interest) encoding physicochemical properties (H-bonding capacity, aromaticity, hydrophobicity, charge) of their cognate residues, pseudocenters being linked together by edges and thus defining a molecular graph. Alignment is operated by detection of maximal common subgraphs (clique detection) (Gardiner *et al.*, 1997) or geometric hashing (Nussinov and Wolfson, 1991) from defined pseudocenters. Local similarity at ligand binding subpockets can thus be detected for proteins with totally different folds and catalytic activities. Predicted similar binding sites can even be linked together in a global network to better position a protein in the target space (Zhang and Grigorov, 2006).

A nice example of binding site similarities for distant proteins has been exemplified by Weber *et al.* (2004), who detected cross-reactivity of arylsulfonamide-based COX-2 inhibitors with human carbonic anhydrase (HCA) based on the similarity of COX-2 and HCA binding pockets. A problem with these matching techniques is that the computed similarity score (usually dependent on the number of atom/pseudocenter/triangle matches) is not always easy to interpret, notably for active sites of different dimensions, because large actives sites will have a tendency to present more matches than small ones even if the latter are more similar. Therefore, normalized distance metrics similar to those used for comparing ligands are needed. A promising approach is proposed by Surgand (2006), who discretizes an active site by a dimensionless 80-triangle sphere and projects, from cβ atoms of cavity-lining residues to the sphere centre, various topological and physicochemical descriptors. A distance between two active sites is thus simply computed by summing up the normalized differences in descriptor space between each triangle of the sphere. The method was able to recognize remote binding site similarities (Figure 10) between a GPCR (GPR30) and a NHR (estrogen receptor α) sharing 4-hydroxytamoxifen as high-affinity ligand (Revankar *et al.*, 2005).

The current speed of such comparisons enables the definition of all-against-all similarity matrices (Schmitt *et al.*, 2002; Shulman-Peleg *et al.*, 2004; Gold and Jackson, 2006), and opens the door to various applications: (i) functional analysis and classification of ligand binding sites, (ii) predicting potential ligands, and (iii) anticipating side effects caused by targeting a peculiar protein.

An alternative approach to compare ligand-binding sites is to evaluate the similarity of potential ligand binding envelopes for known X-ray structure of apo or holoproteins (An *et al.*, 2005). A first draft of the human pocketome, a collection of all possible ligand binding envelopes for a set of 943 crystallized human proteins, has been proposed recently (An *et al.*, 2005) and clustered by envelope similarity. Interestingly, the ligand envelope-based tree only partially matches alternative trees based on the amino-acid sequence of the target proteins or on bound-ligand similarities (An *et al.*, 2005).

Another recently proposed approach to compare proteins of the same family is to look at packing defects (Fernandez *et al.*, 2004) localized at the so-called 'dehydrons' (backbone heavy atoms with unsatisfied H-bonding partners), which are good indicators of protein capacity to interact with potential ligands and can be predicted form the amino-acid sequence. Packing distances between 32 PDB-reported kinases were shown to be almost identical to the pharmacological distance between these kinases estimated from an experimental affinity matrix derived for 17 inhibitors (Fernandez and Maddipati, 2006) and to efficiently guide the structure-based design of selective inhibitors for various enzymes specifically designed to target packing defects (Fernandez, 2005).

Target–ligand based chemogenomic approaches

Chemical annotation of target binding sites

Numerous biologically annotated chemical libraries can be browsed (Table 3) to link chemical to target spaces and focus ligand-based design to target families (Bender *et al.*, 2006; Nettles *et al.*, 2006; Nidhi *et al.*, 2006). However, as far as information about the binding site is missing, there is a potential risk to compare compounds sharing the same

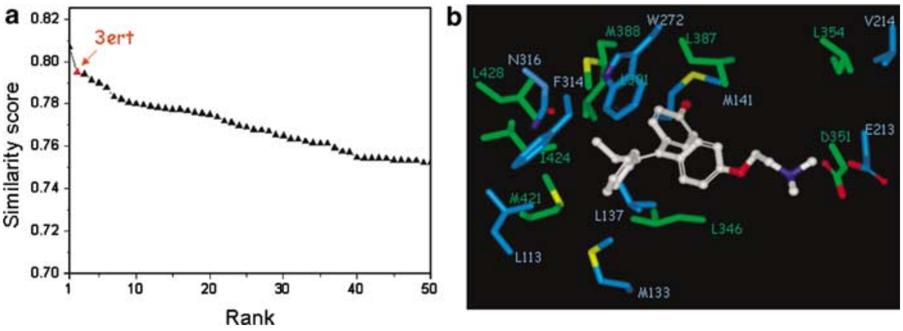


Figure 10 Screening a non-redundant subset of 1060 binding sites form the sc-PDB target library (Kellenberger *et al.*, 2006) for ligand binding sites similar to that of GPR30 for 4-hydroxy-tamoxifen. (**a**) Ranking sc-PDB entries by decreasing similarity (ranging from 1 to 0) to the GPR30 4-hydroxy-tamoxifen binding site (extrapolated from a consensus list of 30 residues delimiting a canonical non-peptide binding site for most GPCR ligands as proposed by Surgand *et al.* (2006). The 3ert sc-PDB entry (4-hydroxy-tamoxifen binding site in the estrogen receptor α) is ranked second among 1060 investigated binding site. (**b**) Predicted alignment of GPR30 (blue) to ER-α binding site (green) for 4-OH tamoxifen (white ball and sticks). Both binding sites present a water-accessible negatively charged residue and a buried hydrophobic region (similarity index of 0.79 according to the SiteAlign program (Surgand *et al.*, 2006)).

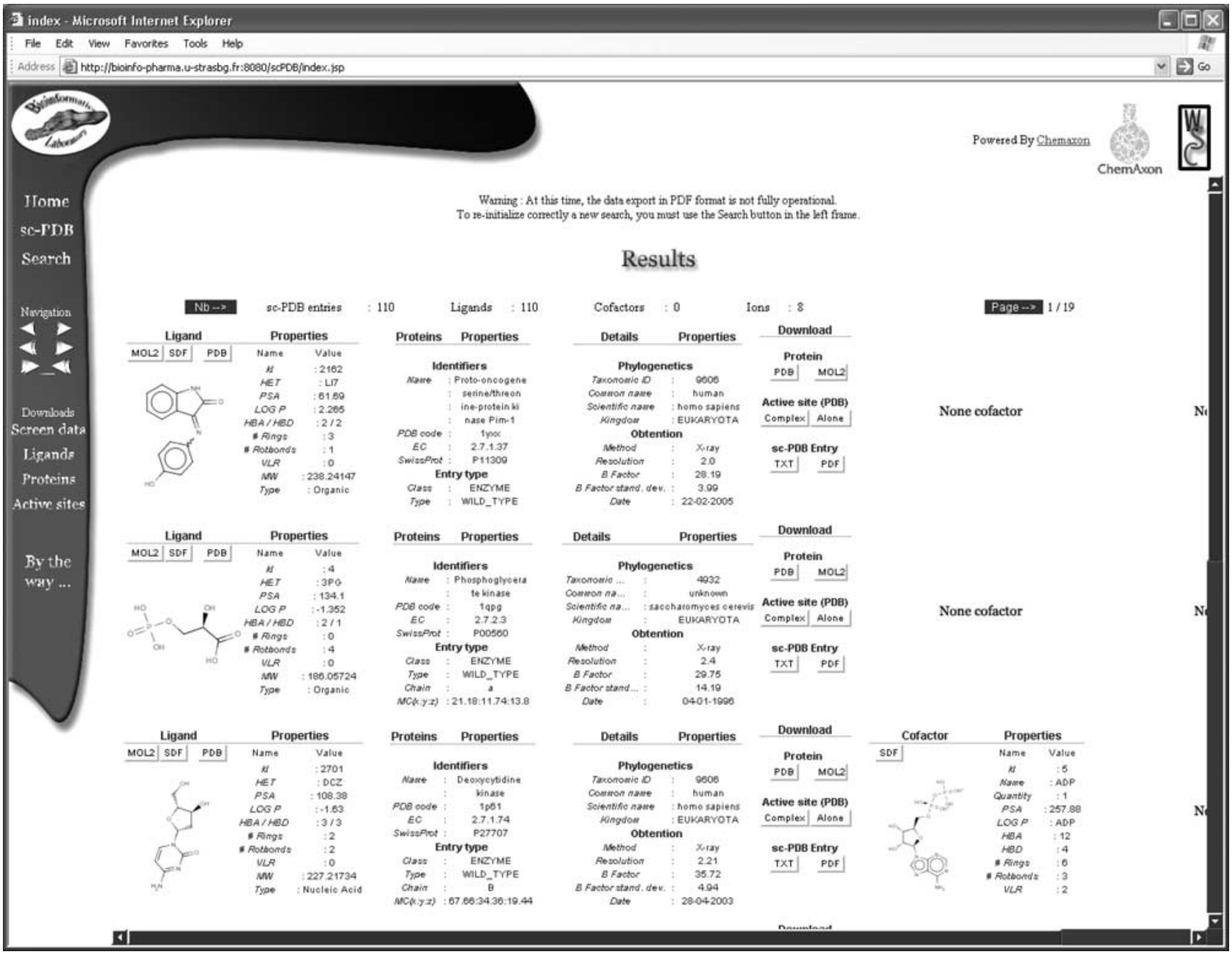
target but not the same binding site (for example, orthosteric and allosteric ligands). It is therefore important to rigorously annotate protein sequences and/or binding site by the chemotype of the ligands they can recognize. The SMID (Small Molecule Interaction Database) archive is an interesting initiative to annotate protein amino-acid sequences by domain-specific ligands (Snyder *et al.*, 2006). A total of 6300 ligands covering 230 000 experimentally observed domain/small molecule interactions have been stored in a relational database, which can be browsed to predict the most likely ligand of proteins of unknown 3-D structures by comparison of their domains to known protein structures using a reverse position-specific basic local alignment search (BLAST) procedures (Feldman *et al.*, 2006). Ligand-annotated binding sites from the PDB are annotated in several databases (Kellenberger *et al.*, 2006), but only two of them (Binding-MOAD, sc-PDB; Table 3) consider the ligand from a pharmacological point of view and are therefore of interest for chemogenomic approaches. Such databases can be used to prioritize either ligands or molecular scaffolds for designing targeted compound libraries covering a well-defined target space (Figure 11).

2-D searches

To browse and predict protein–ligand complexes, one needs to set up simple descriptors for both ligands and proteins from knowledge databases (Table 3) and concatenate them into a single protein–ligand description. The easiest way to encode this information is to start form experimental binding affinity matrices (Kauvar *et al.*, 1995; Krejsa *et al.*, 2003; Vieth *et al.*, 2004) and to define appropriate QSAR/QSPR models to predict the affinity of new compounds for registered targets or the full virtual profile by general neighbourhood behaviour modelling (Krejsa *et al.*, 2003). Another approach has recently been proposed for deorphanizing GPCRs in which a ligand fingerprint is merged to a sequence-based target fingerprint if a high-affinity complex ($pK_i > 7$) has been reported in the PDSP database (http://pdsp.med.unc.edu/). A machine-learning algorithm was trained from 5319 non-redundant known complexes and applied to a set of 1 911 415 virtual complexes (55 orphan receptors and 34 753 drug-like compounds from the NCI database) to predict the most likely associations (Bock and Gough, 2005). Out-of-sample validations (finding the receptors of a promiscuous ligand and the ligands of a single target) were in general agreement with literature data and some predictions still awaiting experimental validations have been made.

3-D searches

A straightforward way to predict putative targets of ligands is to dock each of the ligands of the compound library into each of the active site of the target library. This strategy has been validated by several groups and proved able to recover the known ligands of known targets and predict their off-targets and thus some potential side effects (Chen and Zhi, 2001; Paul *et al.*, 2004). Up to now, there is a single successful target fishing application described in the literature utilizing a docking approach (Muller *et al.*, 2006). Hence, inverse docking requires first a high-quality 3-D dataset of binding sites whose automated set-up is quite difficult, and second an accurate scoring function to properly rank targets. A problem is that energy-based scoring functions are not very good at quantifying very heterogeneous protein–ligand complexes by decreasing binding-free energies (Ferrara *et al.*, 2004) and that alternative ways of scoring are requested for efficient target selection. Among the most promising methods is the computation of IFPs between a protein and its ligand. Practically, IFPs are simple bit strings that convert 3-D information about protein–ligand interactions into simple 1-D bit vector representations (Figure 3) that can be quickly compared by the use of traditional metrics (for example, Tanimoto coefficient, Euclidian distance). Usage of IFPs have shown several promising features: (i) enhancing the quality of pose prediction in docking experiments (Deng *et al.*, 2004; Marcou and Rognan, 2006); (ii) clustering protein–ligand interactions for a panel of related inhibitors according to the diversity of their interaction with a target subfamily



Warning : At this time, the data export in PDF format is not fully operational.
To re-initialize correctly a new search, you must use the Search button in the left frame.

Results

sc-PDB entries : 110 Ligands : 110 Cofactors : 0 Ions : 8 1 / 19

Ligand — Properties

Name	Value
Id	: 2162
HET	: LI7
PSA	: 61.69
LOG P	: 2.265
HBA / HBD	: 2 / 2
# Rings	: 3
# Rotbonds	: 1
VLR	: 0
MW	: 238.24147
Type	: Organic

Proteins — Properties

Identifiers
Name : Proto-oncogene serine/threonine-protein kinase Pim-1
PDB code : 1yxx
EC : 2.7.1.37
SwissProt : P11309
Entry type
Class : ENZYME
Type : WILD_TYPE

Details — Properties

Phylogenetics
Taxonomic ID : 9606
Common name : human
Scientific name : homo sapiens
Kingdom : EUKARYOTA
Obtention
Method : X-ray
Resolution : 2.0
B Factor : 28.19
B Factor stand. dev. : 3.99
Date : 22-02-2005

Download
Protein
Active site (PDB)
sc-PDB Entry

Ligand — Properties

Name	Value
Id	: 4
HET	: 3PG
PSA	: 134.1
LOG P	: -1.352
HBA / HBD	: 2 / 1
# Rings	: 0
# Rotbonds	: 4
VLR	: 0
MW	: 186.05724
Type	: Organic

Proteins — Properties

Identifiers
Name : Phosphoglycerate kinase
PDB code : 1qpg
EC : 2.7.2.3
SwissProt : P00560
Entry type
Class : ENZYME
Type : WILD_TYPE
Chain : a
MC(x:y:z) : 21.18:11.74:13.8

Details — Properties

Phylogenetics
Taxonomic ... : 4932
Common na... : unknown
Scientific na... : saccharomyces cerevis
Kingdom : EUKARYOTA
Obtention
Method : X-ray
Resolution : 2.4
B Factor : 29.75
B Factor stand... : 14.19
Date : 04-01-1996

Download
Protein
Active site (PDB)
sc-PDB Entry

Ligand — Properties

Name	Value
Id	: 2701
HET	: DCZ
PSA	: 108.38
LOG P	: -1.63
HBA / HBD	: 3 / 3
# Rings	: 2
# Rotbonds	: 2
VLR	: 0
MW	: 227.21734
Type	: Nucleic Acid

Proteins — Properties

Identifiers
Name : Deoxycytidine kinase
PDB code : 1p61
EC : 2.7.1.74
SwissProt : P27707
Entry type
Class : ENZYME
Type : WILD_TYPE
Chain : B
MC(x:y:z) : 67.66:34.36:19.44

Details — Properties

Phylogenetics
Taxonomic ID : 9606
Common name : human
Scientific name : homo sapiens
Kingdom : EUKARYOTA
Obtention
Method : X-ray
Resolution : 2.21
B Factor : 35.72
B Factor stand. dev. : 4.94
Date : 28-04-2003

Download
Protein
Active site (PDB)
sc-PDB Entry

Cofactor — Properties

Name	Value
Id	: 5
Name	: ADP
Quantity	: 1
PSA	: 257.88
LOG P	: ADP
HBA	: 12
HBD	: 4
# Rings	: 6
# Rotbonds	: 3
VLR	: 2

Figure 11 Querying the sc-PDB chemogenomic database (http://bioinfo-pharma.u-strasbg.fr/scPDB) for rule-of-five compliant (Lipinski *et al.*, 2001) small molecular weight fragments (MW <300, clogP <3, H-bond donor count <3, H-bond acceptor count <6) co-crystallized with protein kinases.

(Chuaqui *et al.*, 2005; Marcou and Rognan, 2006); (iii) assisting target-biased library design (Deng *et al.*, 2006).

However, docking-independent 3-D methods may also constitute an interesting approach to predict protein–ligand complexes. A significant problem is to encode protein and ligand properties with similar descriptors such that one partner can be retrieved by using the second one as a query. A promising solution is proposed with the CoLiBRI (Complementary Ligands Based on Receptor Information) method (Oloff *et al.*, 2006) in which both ligand and active site atoms are described by a same vector of molecular descriptors derived from shape and electronic properties of isolated atoms. Therefore, it is possible to directly correlate chemical similarities between active site and their ligands by mapping patterns of active sites onto patterns of their complementary ligands. When applied to a test data set of 800 high-resolution PDB complexes, the complementary ligand was ranked among the top 1% of a large library in 90% of tested active sites. Accuracy dropped significantly for active sites very different from those in the test set but still usable as a prefiltering step for removing the most improbable ligands (Oloff *et al.*, 2006).

Concluding remarks

Chemogenomic approaches to rational drug discovery have been exploding in the last years as high-throughput data (structure, binding affinity and functional effects) become available for both targets and ligands of pharmaceutical interest. Numerous ways to link those data have been proposed focusing on either ligand or target neighbourhood. A clear data organization and storage is necessary to foster such applications and begins to emerge for the most interesting target families (kinases, GPCRs and NHRs). In a near feature, an earlier and better control of ligand selectivity can be anticipated by using chemogenomic data. This does not mean that more selective ligands are going to be designed, but simply that the observed selectivity profile of the compound will be compatible with a therapeutical usage. In addition, novel genomic targets could be better addressed after locating them in the target space and exploiting the associated chemical information.

Conflict of interest

The author states no conflict of interest.

References

An J, Totrov M, Abagyan R (2005). Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* **4**: 752–761.

Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol* **243**: 327–344.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL *et al.* (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* **31**: 400–402.

Bender A, Glen RC (2004). Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem* **2**: 3204–3218.

Bender A, Jenkins JL, Glick M, Deng Z, Nettles JH, Davies JW (2006). 'Bayes affinity fingerprints' improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *J Chem In Model* **46**: 2445–2456.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H *et al.* (2000). The protein data bank. *Nucleic Acids Res* **28**: 235–242.

Bhavani S, Nagargadde A, Thawani A, Sridhar V, Chandra N (2006). Substructure-based support vector machine classifiers for prediction of adverse effects in diverse classes of drugs. *J Chem Inf Model* **46**: 2478–2486.

Bock JR, Gough DA (2005). Virtual screen for ligands of orphan G protein-coupled receptors. *J Chem Inf Model* **45**: 1402–1414.

Bondensgaard K, Ankersen M, Thogersen H, Hansen BS, Wulff BS, Bywater RP (2004). Recognition of privileged structures by G-protein coupled receptors. *J Med Chem* **47**: 888–899.

Caron PR, Mullican MD, Mashal RD, Wilson KP, Su MS, Murcko MA (2001). Chemogenomic approaches to drug discovery. *Curr Opin Chem Biol* **5**: 464–470.

Casbon J, Saqi MA (2005). S4: structure-based sequence alignments of SCOP superfamilies. *Nucleic Acids Res* **33**: D219–D222.

Cases M, Garcia-Serna R, Hettne K, Weeber M, van der Lei J, Boyer S *et al.* (2005). Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. *Curr Top Med Chem* **5**: 763–772.

Chen YZ, Zhi DG (2001). Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **43**: 217–226.

Chuaqui C, Deng Z, Singh J (2005). Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J Med Chem* **48**: 121–133.

Clark M (2005). Generalized fragment-substructure based property prediction method. *J Chem Inf Model* **45**: 30–38.

Crossley R (2004). The design of screening libraries targeted at G-protein coupled receptors. *Curr Top Med Chem* **4**: 581–588.

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* **47**: 337–344.

Deng Z, Chuaqui C, Singh J (2006). Knowledge-based design of target-focused libraries using protein–ligand interaction constraints. *J Med Chem* **49**: 490–500.

Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL *et al.* (1988). Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J Med Chem* **31**: 2235–2346.

Feldman HJ, Snyder KA, Ticoll A, Pintilie G, Hogue CW (2006). A complete small molecule dataset from the protein data bank. *FEBS Lett* **580**: 1649–1653.

Fernandez A (2005). Incomplete protein packing as a selectivity filter in drug design. *Structure* **13**: 1829–1836.

Fernandez A, Maddipati S (2006). A priori inference of cross reactivity for drug-targeted kinases. *J Med Chem* **49**: 3092–3100.

Fernandez A, Rogale K, Scott R, Scheraga HA (2004). Inhibitor design by wrapping packing defects in HIV-1 proteins. *Proc Natl Acad Sci USA* **101**: 11640–11645.

Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks III CL (2004). Assessing scoring functions for protein–ligand interactions. *J Med Chem* **47**: 3032–3047.

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T *et al.* (2006). Pfam: clans, web tools and services. *Nucleic Acids Res* **34**: D247–D251.

Frimurer TM, Ulven T, Elling CE, Gerlach LO, Kostenis E, Hogberg T (2005). A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorg Med Chem Lett* **15**: 3707–3712.

Fukunishi Y, Kubota S, Nakamura H (2006). Noise reduction method for molecular interaction energy: application to in silico drug screening and in silico target protein screening. *J Chem Inf Model* **46**: 2071–2084.

Gardiner EJ, Artymiuk PJ, Willett P (1997). Clique-detection algorithms for matching three-dimensional molecular structures. *J Mol Graph Model* **15**: 245–253.

Gold ND, Jackson RM (2006). Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. *J Mol Biol* **355**: 1112–1124.

Holm L, Park J (2000). DaliLite workbench for protein structure comparison. *Bioinformatics* **16**: 566–567.

Hoppe C, Steinbeck C, Wohlfahrt G (2006). Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials. *J Mol Graph Model* **24**: 328–340.

Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA (2005). Binding MOAD (Mother of All Databases). *Proteins* **60**: 333–340.

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS *et al.* (2006). The PROSITE database. *Nucleic Acids Res* **34**: D227–D230.

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.

Kauvar LM, Higgins DL, Villar HO, Sportsman JR, Engqvist-Goldstein A, Bukar R *et al.* (1995). Predicting ligand binding to proteins by affinity fingerprinting. *Chem Biol* **2**: 107–118.

Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D (2006). sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J Chem Inf Model* **46**: 717–727.

Kinoshita K, Nakamura H (2003). Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci* **12**: 1589–1595.

Klabunde T, Hessler G (2002). Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem* **3**: 928–944.

Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A *et al.* (2005). Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* **102**: 17272–17277.

Kratochwil NA, Malherbe P, Lindemann L, Ebeling M, Hoener MC, Muhlemann A *et al.* (2005). An automated system for the analysis of G protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application. *J Chem Inf Model* **45**: 1324–1336.

Krejsa CM, Horvath D, Rogalski SL, Penzotti JE, Mao B, Barbosa F *et al.* (2003). Predicting ADME properties and side effects: the BioPrint approach. *Curr Opin Drug Discov Devel* **6**: 470–480.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**: 3–26.

Marcou G, Rognan D (2006). Optimizing scaffold docking by use of molecular interaction fingerprints. *J Chem Info Model* **47**: 195–207.

Matter H, Schwab W (1999). Affinity and selectivity of matrix metalloproteinase inhibitors: a chemometrical study from the perspective of ligands and proteins. *J Med Chem* **42**: 4506–4523.

Mestres J, Martin-Couce L, Gregori-Puigjane E, Cases M, Boyer S (2006). Ligand-based approach to *in silico* pharmacology: nuclear receptor profiling. *J Chem Inf Model* **46**: 2725–2736.

Morphy R (2006). The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. *J Med Chem* **49**: 2969–2978.

Muller P, Lena G, Boilard E, Bezzine S, Lambeau G, Guichard G *et al.* (2006). *In silico* guided target identification of a scaffold-focused library: 1,3,5-triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J Med Chem* **49**: 6768–6778.

Naumann T, Matter H (2002). Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes. *J Med Chem* **45**: 2366–2378.

Nettles JH, Jenkins JL, Bender A, Davies JW, Glick M (2006). Bridging chemical and biological space: target fishing using 2D and 3D molecular descriptors. *J Med Chem* **49**: 6802–6810.

Nidhi, Glick M, Davies JW, Jenkins JL (2006). Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* **46**: 1124–1133.

Nussinov R, Wolfson HJ (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci USA* **88**: 10495–10499.

Oloff S, Zhang S, Sukumar N, Breneman C, Tropsha A (2006). Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). *J Chem Inf Model* **46**: 844–851.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA *et al.* (2000). Crystal structure of rhodopsin: a g protein-coupled receptor. *Science* **289**: 739–745.

Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006). Global mapping of pharmacological space. *Nat Biotechnol* **24**: 805–815.

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

Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A *et al.* (2006). MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **34**: D291–D295.

Pirard B, Matter H (2006). Matrix metalloproteinase target family landscape: a chemometrical approach to ligand selectivity based on protein binding site analysis. *J Med Chem* **49**: 51–69.

Powers R, Copeland JC, Germer K, Mercier KA, Ramanathan V, Revesz P (2006). Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **65**: 124–135.

Raymond JW, Blankley CJ, Willett P (2003). Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *J Mol Graph Model* **21**: 421–433.

Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006). Structural diversity of domain superfamilies in the CATH database. *J Mol Biol* **360**: 725–741.

Revankar CM, Cimino DF, Sklar LA, Arterburn JB, Prossnitz ER (2005). A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science* **307**: 1625–1630.

Root DE, Flaherty SP, Kelley BP, Stockwell BR (2003). Biological mechanism profiling using an annotated compound library. *Chem Biol* **10**: 881–892.

Rosen M, Lin SL, Wolfson H, Nussinov R (1998). Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng* **11**: 263–277.

Russ AP, Lampel S (2005). The druggable genome: an update. *Drug Discov Today* **10**: 1607–1610.

Sadowski J, Kubinyi H (1998). A scoring scheme for discriminating between drugs and nondrugs. *J Med Chem* **41**: 3325–3329.

Savchuk NP, Balakin KV, Tkachenko SE (2004). Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr Opin Chem Biol* **8**: 412–417.

Schmitt S, Kuhn D, Klebe G (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* **323**: 387–406.

Schnur DM, Hermsmeier MA, Tebben AJ (2006). Are target-family-privileged substructures truly privileged? *J Med Chem* **49**: 2000–2009.

Schuffenhauer A, Floersheim P, Acklin P, Jacoby E (2003). Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci* **43**: 391–405.

Sheridan RP, Kearsley SK (2002). Why do we need so many chemical similarity search methods? *Drug Discov Today* **7**: 903–911.

Shindyalov IN, Bourne PE (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**: 739–747.

Shulman-Peleg A, Nussinov R, Wolfson HJ (2004). Recognition of functional sites in protein structures. *J Mol Biol* **339**: 607–633.

Singh J, Deng Z, Narale G, Chuaqui C (2006). Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem Biol Drug Des* **67**: 5–12.

Snyder KA, Feldman HJ, Dumontier M, Salama JJ, Hogue CW (2006). Domain-based small molecule binding site annotation. *BMC Bioinformatics* **7**: 152.

Standley DM, Toh H, Nakamura H (2005). GASH: an improved algorithm for maximizing the number of equivalent residues between two protein structures. *BMC Bioinformatics* **6**: 221.

Steindl TM, Schuster D, Laggner C, Langer T (2006). Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J Chem Inf Model* **46**: 2146–2157.

Surgand JS (2006). *Développement de nouvelles méthodes bioinformatiques pour l'étude des récepteurs couplés aux protéines G*. Thèse de l'Université Louis Pasteur – Strasbourg I: France.

Surgand JS, 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.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG *et al.* (2001). The sequence of the human genome. *Science* **291**: 1304–1351.

Vieth M, Higgs RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H (2004). Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim Biophys Acta* **1697**: 243–257.

Votano JR, Parham M, Hall LH, Kier LB, Hall LM (2004). Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem Biodivers* **1**: 1829–1841.

Wallace AC, Borkakoti N, Thornton JM (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* **6**: 2308–2323.

Wang J, Krudy G, Xie XQ, Wu C, Holland G (2006). Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. *J Chem Inf Model* **46**: 2674–2683.

Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A *et al.* (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.

Weininger D (1988). SMILES 1. Introduction and encoding rules. *J Chem Inf Comput Sci* **28**: 31–36.

Willett P (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* **11**: 1046–1053.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **34**: D187–D191.

Xia X, Maliski EG, Gallant P, Rogers D (2004). Classification of kinase inhibitors using a Bayesian model. *J Med Chem* **47**: 4463–4470.

Zhang Z, Grigorov MG (2006). Similarity networks of protein binding sites. *Proteins* **62**: 470–478.