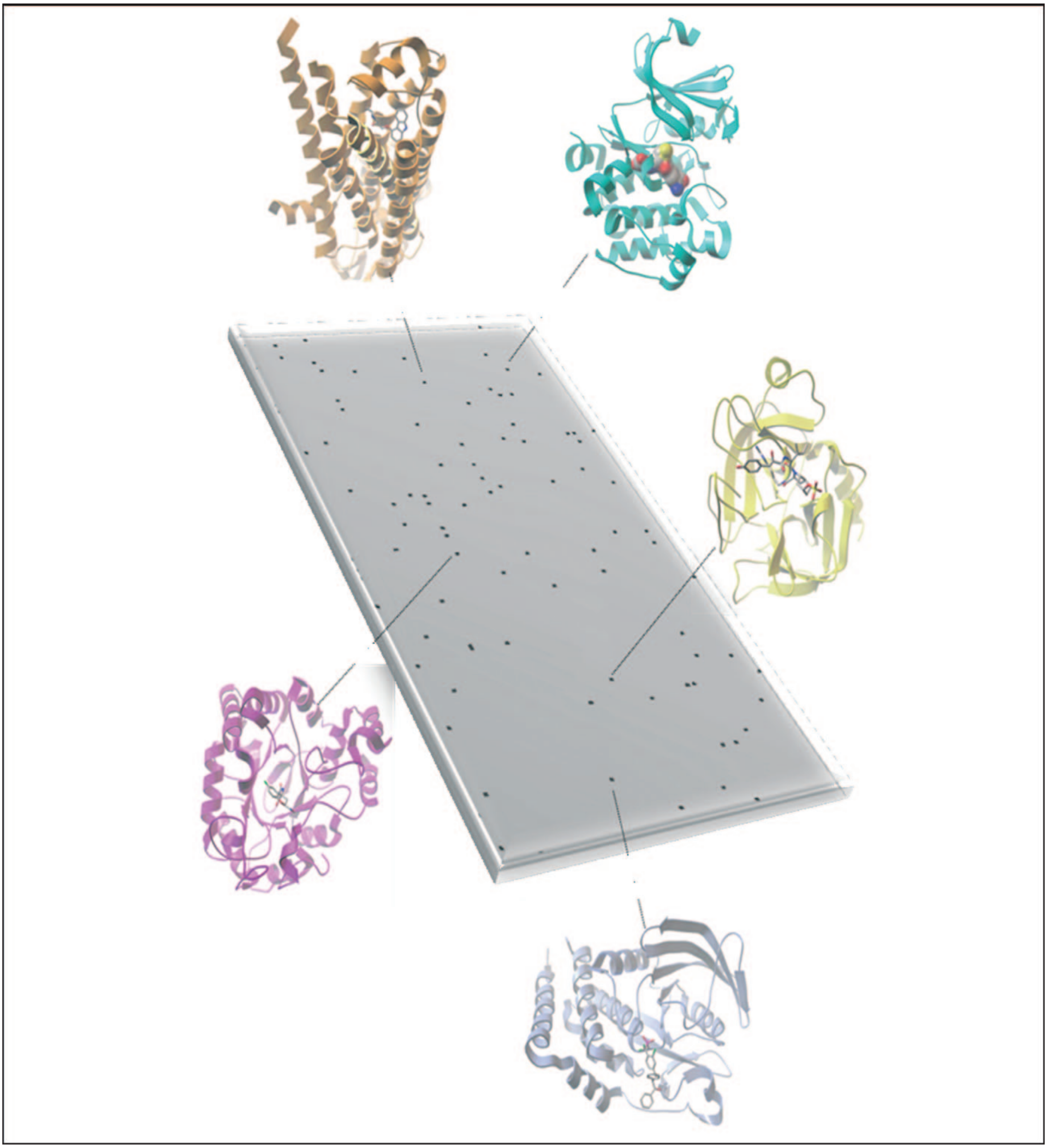
Review

DOI: 10.1002/minf.200900081

Structure-Based Approaches to Target Fishing and Ligand Profiling

Didier Rognan*[a]



Abstract: Chemogenomics is an emerging interdisciplinary field aiming at identifying all possible ligands of all possible targets. If one groups targets in columns and ligands in rows, chemogenomic approaches to drug discovery just fill the interaction matrix. Since experimental data do not suffice, several computational methods are currently actively developed to supplement time-consuming and costly experiments. They are either designed to fill rows and thus profile a ligand towards a heterogeneous set of targets (target profiling) or to fill columns and thus identify novel ligands for an existing target (standard virtual screening). At the interface of both strategies are now true chemogenomic computational methods filling well defined areas in the matrix. The present review will focus on (protein) structure-based approaches and illustrates major advances in this novel exciting field which is supposed to massively impact rational drug design in the next decade.

Keywords: Chemogenomics · Docking · Pharmacophore · Binding site · Fingerprint · Proteins · Computational chemistry

1 Introduction

Computational chemists have long been challenged by medicinal chemists and pharmacologists to predict the affinity of low molecular-weight compounds to a particular target, in order to guide hit-to-lead optimization programs. This demand has pushed, 10 years ago, molecular docking programs at the forefront of computational tools assisting drug discovery. After a few years of practice, it became clear that scoring functions used by docking programs were too inaccurate to predict binding free energies and precisely rank-order compounds by decreasing predicted affinity.[1]

Despite some disappointments in the computational chemistry community, docking tools and related structure-based computational approaches are nonetheless still useful in prioritizing design hypotheses and guiding experiments. If quantifying ligand binding is difficult, notably for a set of heterogeneous compounds, compound filtering and prioritization has for long proved its utility in virtual screening campaigns.[2] However, changing paradigms in rational drug design -making a binary yes-or-no answer instead or quantitative predictions- are pushing again structure-based algorithms under the spotlights.[3] Having learned from previous experiences, the question is no more to predict the affinity of a compound for a target but to profile a ligand against a wide collection of macromolecular targets and thus answer to two basic questions: (i) Which (novel) proteins should my compound bind to? (“Target fishing”) (ii) What could be the global pharmacological profile of this compound? (“Ligand profiling”)

Hence, comparing virtual target profiles seem to outperform standard chemical similarity measurements in assessing whether two ligands are ‘similar’ or not.[4] Ligand-centric approaches (Table 1) which are actively developed and used to predict the polypharmacological profile of bioactive compounds will not be surveyed here (for more details on the latter approach, the reader is invited to read recent reviews).[5,6]

The present report will only cover (protein) structure-based approaches to target fishing and ligand profiling. Four approaches with decreasing maturity level will be inspected: protein-ligand docking, structure-based pharmacophore searches, binding site similarity measurements, and protein-ligand fingerprints. After describing the underlying concepts, prospective studies will be reviewed and the relative advantages/drawbacks of each approach will be critically discussed.

2 Protein–Ligand Docking

2.1 Proof of Concept

The most straightforward computational approach to predict whether a ligand may bind to a macromolecular target is molecular docking, in other words finding the protein-bound conformation and location of the ligand form their respective three-dimensional (3-D) atomic coordinates. Since the pioneering work of Kuntz and co-workers,[16] a wide array of docking programs based on quite different physicochemical approximations and global optimization methods have been developed.[17] Docking has been widely used as a virtual screening engine to find novel ligands for pharmacologically-interesting targets[2] notably for protein kinases,[18] nuclear hormone receptors[19] and G protein-coupled receptors.[20] Quite strikingly, the opposite paradigm – finding novel targets for a pharmacologically-interesting ligand- has for long not been investigated. Basically, one only needs a database of ready-to-dock protein-ligand binding sites, a docking program, and a post-processing script for rank-ordering targets by decreasing docking score (Figure 1).

Automating the set-up of a large collection of heterogeneous binding sites was a difficult endeavor in the early 90s since the Protein Data Bank was not remediated[21] at that time. Numerous inconsistencies in various records (e.g. chemical description and nomenclature of monomer units,

[a] *D. Rognan*
Structural Chemogenomics, UMR 7200 CNRS-UdS
74 route du Rhin, F-67400 Illlkirch
phone: +33.3.68854235
fax: +33.3.68854310
**e-mail: rognan@unistra.fr*

Table 1. Ligand-based target fishing approaches.

Descriptor	Metric	Reference
MNA[a]	Probability	[7]
Similog keys	Tanimoto coefficient	[8]
SHED[b]	Euclidean distance	[9]
Surface points	Morphological similarity	[10]
ECFP[c]	Bayesian classifier	[11]
ECFP, MDL keys, FEFOPS[d]	Tanimoto coefficient	[12]
Daylight 2-D fingerprint	SEA[e] expectation value	[13, 14]
Bayes affinity fingerpint	Pearson correlation	[4]
ECFP, Molprint	Bayesian classifier	[15]

[a] Multilevel neighbourhood of atoms. [b] Shannon entropy descriptor. [c] Extended connectivity fingerprint. [d] feature point pharmacophores. [e] Similarity ensemble approach

description of organic molecules, biological annotations) had been accumulated over time due to the large number of contributors. Therefore, initial attempts of serial docking to several binding sites have been limited to a restricted set of very similar targets. Lamb et al. reported the docking of combinatorial libraries to three serine endopeptidases and were able to prioritize substituents for well-defined scaffolds.[22] Rockey et al. used AutoDock[23] to differentiate ADP from GDP receptors by docking both nucleotides to known receptors for both metabolites, but already customized a scoring function (electrostatic interaction energy of the purine ring) for rank ordering targets. Interestingly, the same authors provided a first proof-of-concept that predicting the specificity profile of three inhibitors to twenty protein kinases was achievable by docking.[24] Although a few false positives (good docking poses in non-inhibited kinases) were reported, no false negatives (no docking solution or poor scores for strongly inhibited kinases) were identified.

Serial screening at a larger scale awaited the development of protein-ligand binding sites collections. Several protein-ligand databases (Relibase,[25] LPDB,[26] ProLINT,[27] Ligbase[28]) derived from the PDB had already been described but none of them were directly usable to generate a collection of protein active sites customized to accommodate small molecular-weight 'drug-like' ligands. The first of such reports was published in 2001.[29] Using Dock[16] as docking engine, a set of 2700 protein cavities was searched for identifying known targets of 4-hydroxy tamoxifen and vitamin E. Docked poses were scored by a molecular mechanics-based scoring function and targets selected if the docking score was quantitatively comparable to that of the cognate ligand in the PDB entry. Out of relatively short-sized target lists (ca. 20 proteins) 6 and 2 true targets were recovered for both ligands, respectively.[29] A couple of years later, Paul et al.[30] reported a similar procedure using the Gold docking program[31] but with a set of 2100 protein-ligand binding sites (sc-PDB) defined at the atomic level from the position of bound ligands. The approach, although based on crude docking scores was able to recover the main targets of four bioactive ligands among the top 1% scoring entries. The major breakthrough was the definition of a multi-step flowchart for browsing standard PDB files and automatically detecting protein, ligand, co-factors and biologically irrelevant molecules. Using pharmacologically active ligands presented the advantage of removing a priori undruggable cavities (e.g. binding sites of detergents, ions) from the binding site library. The set-up of the sc-PDB database (http://bioinfo-pharma.u-strasbg.fr/scPDB) has benefited from several improvements over time[32,33] including enhanced biological annotation of protein targets, semi-automated correction of ligand atom types and bond orders, selection of druggable protein-ligand binding sites, definition of protein-ligand interaction fingerprints (IFPs),[34] and clustering of binding sites by physicochemical properties. It now registers 7079 entries including 2378 unique proteins and 3688 unique ligands. Recently, the potential drug target database (PDTD)[35] was developed along the same ideas from PDB targets of known drugs; it currently includes 1027 entries for 847 unique targets. Significant efforts to automate the entire docking process have been made to enable non-experts to use docking-based ligand profiling. A good illustration is the DOCK Blaster expert system,[36] which fully automates the docking preparation steps and relies on the USCF DOCK algorithm[37] to pose

Didier Rognan heads the Structural Chemogenomics Laboratory at the National Center for Scientific Research (CNRS) in Illkirch (France). He studied Pharmacy at the University of Rennes (France) and did a Ph.D. in Medicinal Chemistry in Strasbourg (France) under the supervision of Prof. C. G. Wermuth. After a post-doctoral fellowship at the University of Tübingen (Germany), he moved as an Assistant Professor at the Swiss Federal Institute of Technology (ETH Zürich) until October 2000. He was then appointed Research Director at the CNRS to build a new group in Illkirch (France). He is mainly interested in all aspects (method development, applications) of structure-based drug design, notably matching target with ligand space by means of chemogenomic computational methods.

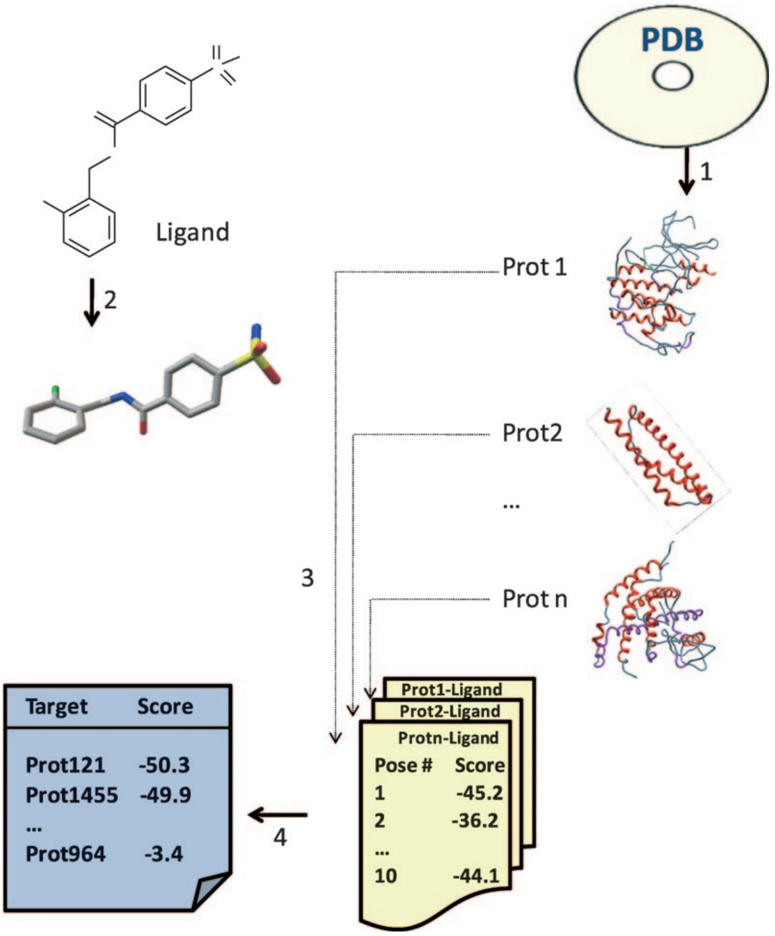


Figure 1. Inverse docking flowchart. 1) Retrieval of protein-ligand binding sites from the Protein Data Bank (PDB). 2) Converting a ligand 2-D sketch into a 3-D structure. 3) Serial docking of the ligand to all binding sites. 4) Post-processing the raw docking data to define a target list.

and score ligands. When applied to a set of 9050 PDB structures, DOCK Blaster could indeed redock 7755 ligands out of which 3056 had a good pose (rmsd $<$ 2.0 Å) and 1398 a good score in a retrospective virtual screen when seeded with decoys.[36]

2.2 Success Stories

One of the first application of inverse docking has been the identification of protein targets for natural products. Do et al. used an in-house developed Selnergy strategy[38,39] to propose targets for ε-viniferin, a cosmetic ingredient, and for meranzin, a component from a plant of the *Rutaceae* family. In both cases, a set of 400 manually-selected PDB targets were screened and ranked using the FlexX docking software.[40] Phosphodiesterase type 4 (PDE4) could be confirmed as a true target of ε-viniferin, and three additional proteins (COX1, COX2, PPARγ) could be identified as true targets of meranzin.

Identifying a target for a natural compound undoubtedly helps in optimizing its potency and selectivity. Screening a natural compound, discovered by a previous anti-*Helicobacter pylori* experimental screen, against the PDTD database with the TarFisDock approach[41] enabled the selection of 15 putative target candidates.[42] Only two of them had homologues in *H. pylori* and were consequently tested for *in vitro* binding. One of them, peptide deformylase, appeared to be a true target of the active compounds with low micromolar affinities. Co-crystallization studies confirmed the reverse docking hypothesis and thus constituted a good starting point for a lead optimization program.

Selectivity for a gene product family is another successfully used application of inverse docking. Aronov et al.[43] docked, with the DOCK algorithm, a combinatorial library of phtalimide derivatives to 6 guanine phosphoribosyl transferases (GPRT) of known X-ray structures. Two compounds arising from a manual selection among the top 10% scorers for the Protozoan parasite *Giardia lamblia* GPRT were selected for their predicted selectivity over the human enzyme. After synthesis and in vitro evaluation, both analogues showed moderate micromolar affinities and selectivity for the parasite enzyme.

Determining a complete selectivity profile for a bioactive compound is experimentally not feasible, even for a protein family. One of the most exhaustive selectivity study reported to date is the profiling of 20 inhibitors on 119 out of the 516 protein kinases of the human kinome.[44] Docking-based serial screening is then a good approach to complement the known selectivity profile of existing drugs. Zahler et al. screened, with the GlamDock software,[45] the sc-PDB database for identifying novel protein kinase targets of indirubin derivatives.[46] In addition to known targets (CDK2, CDK5, GSK-3β) which were indeed recovered among the top 1% scorers, 6 additional kinases were prioritized. 3-phosphoinositide-dependant protein kinase 1 (PDK1) was indeed confirmed as a true target of one derivative (6BiO) exhibiting a 1.5 μM in vitro affinity.

A better control of the ligand selectivity profile is supposed to avoid launching drugs exhibiting side effects and serious adverse drug reactions (SADR). A nice illustration of the power of serial docking in elucidating molecular mechanisms of SADRs was recently brought by Yang et al.[47] A virtual chemical-protein interactome (CPI) was defined by docking 162 drugs known to cause SADR, to a set of 845 different proteins. Interestingly, drugs sharing similar SADRs were found to have similar CPI profiles. Notably, nine sulfonamide derivatives causing a toxic epidermal necrolysis (TEN) were sharing good docking scores to HLA-Cw*4, a protein encoded by an allele-specific TEN inducer.

Inverse docking may last be used to "deorphanize" a ligand which means finding putative targets for a novel chemotype. Muller et al. reported the identification of two secreted phospholipase A2 isoforms as targets of compounds sharing a 1,3,5-triazepan-2,6-dione scaffold by systematic docking of a small scaffold-focused combinatorial library to 2 100 sc-PDB structures.[48] A customized target selection flowchart using several filters (empirical docking score, target enrichment in the top 2% scorers) was mainly responsible for this success.

2.3 Pros and Cons

An obvious key asset of docking-based target fishing is that the putative protein-ligand coordinates are part of the output. Having both the target and the binding mode available presents noticeable advantages. The proposed binding mode can be used as a guide for a hit-to-lead optimization phase to improve ligand properties (potency, selectivity). The observed binding mode can also be compared to that of true ligands of the selected target to check whether or not key interactions are conserved. However, selecting a target by a structure-based approach does not necessarily mean that the corresponding ligand binding mode is correct, notably if standard empirical scoring functions have been used to prioritize targets. For example, the antipsychotic haloperidol was proposed to bind to the catalytic cleft of the HIV-1 protease.[49] In vitro binding experiments confirmed the prediction suggesting a novel lead for discovering antiviral compounds. However, the true binding mode, as revealed later on by X-ray diffraction, differed quite significantly from the predicted binding mode.[50] Fast scoring functions are notoriously inaccurate to predict binding free energies.[1,51] A high false negative target rate is thus to be expected from a pure docking-based approach. Notably, using a binary classification of targets (good, bad) based on a single docking score threshold is unlikely to perform well. Combining docking scores with a topological evaluation of docking poses was shown to improve the target selection process.[32] Alternative approaches to select docking poses (e.g. using protein-ligand interaction fingerprints[34]) permits to secure the selection of the right protein for the right reasons. Interestingly, protein flexibility (a major problem in protein-ligand docking)[17] is inherently taken into account as far as several sets of protein coordinates co-crystallized with various ligands are available in the binding site collection. Handling binding site flexibility is thus not mandatory at least for proteins present in multiple copies in the binding site library.

The major problem with a docking-based in silico target screening remains the preparation of a heterogeneous collection of binding cavities even if considerable progresses in data curation and harmonization have been brought to the Protein Data Bank.[21] Several steps (e.g. defining the position of polar hydrogen atoms, assigning a relevant tautomeric state, atom typing of cofactors) are not straightforward to automate. The influence of the binding site on the ligand ionization state is also difficult to anticipate. For example, an arylsulfonamide is neutral when bound to most protein cavities but can be deprotonated if a metal ion is present in the binding site (Figure 2). Modifying on-the-fly the protonation state of the ligand according to the binding site context would require a prior storage of all possible ionization states of both ligand and protein and is currently not available in most docking tools.

Above all, one should not forget that this target screening method is only applicable to proteins for which a 3-D structure is known, preferably in the apo (ligand-bound) form since docking to holo (ligand-free) structures is less successful.[52] The Protein Data Bank currently stores 60000 entries for a total of 35000 non redundant sequences.[53] The sc-PDB dataset of druggable protein-ligand binding sites registers 7000 entries for a total of 2378 different proteins.[54] The structural coverage of target space although significant is far from being complete. It notably excludes almost all membrane proteins (G protein-coupled receptors, ion channels, transporters) which are of outstanding importance in drug discovery. Noteworthy, recent advances in structural genomics let us anticipate a near complete coverage of target space in 15 years.[55] Docking-based target fishing can technically be applied to comparative 3-D models at the cost of a lower accuracy in target identification because of inherent docking failures.[52] It presents the advantage to considerably extend the frontiers of target space to ca. 8 million 3-D models covering ca. 2 mil-

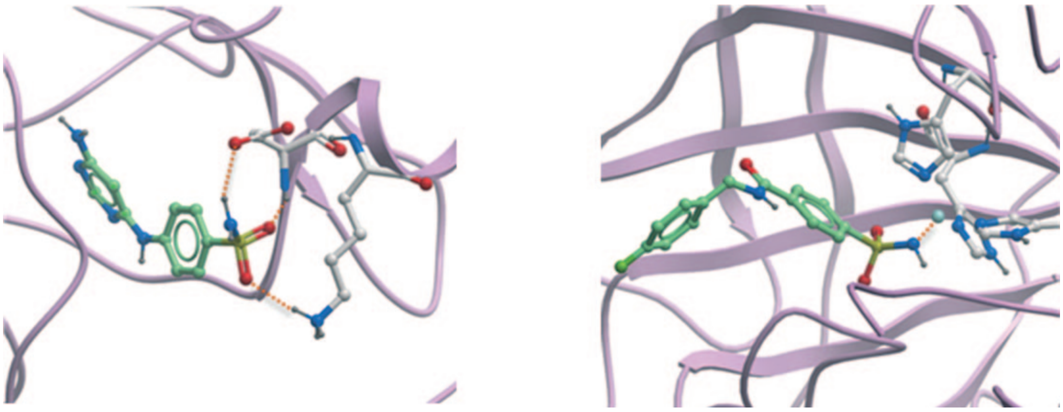


Figure 2. Modification of the ligand ionization state as a function of the binding site. An arylsulfonamide is neutral in most protein cavities (e.g. cyclin-dependant protein kinase 2, 1jsv PDB entry; left panel) but can be deprotonated in presence of a metal ion (e.g. carbonic anhydrase, PDB entry 1i9l; right panel). Hydrogen bonds and metal coordination involving the sulfonamide moiety are displayed as orange dotted lines.

lion sequences.[56,57] This target space is practically impossible to screen by docking because of the huge computing demand (1 docking/min on average). Docking is obviously the most computer demanding method out of all those presented here. With the rapid growth of affordable clusters, it can however be applied to profile a ligand against 10–20 000 proteins within a day. It is the decision of the user to include or not protein 3-D models in the target library and to choose the right docking strategy notably with regard to protein flexibility.[58]

3 Structure-Based Pharmacophore Screening

3.1. Proof of Concept

In 2006, Langer and co-workers introduced the concept of parallel screening[59] for profiling bioactive ligands against a collection of structure-based pharmacophores (Figure 3).

Starting from protein-ligand complexes of known X-ray structure, pharmacophore hypotheses are generated by an automated pharmacophore perception algorithm (LigandScout)[60] and converted into Catalyst[61] pharmacophore queries. In the seminal validation paper,[59] 10 pharmacophore hypotheses including protein-based restraints (exclusion volume spheres and surfaces) were generated for each of 5 antiviral targets. A dataset of 100 antiviral compounds (20 for each target) was then screened with Catalyst against all 50 pharmacophore models and results analyzed in form of a heat map enabling to assess the sensitivity and specificity of each pharmacophore query (vertical inspection) and of every ligand profiling (horizontal inspection). For 90% of compounds, the profiling was correct i.e. the retrieval rate of true inhibitors by their corresponding target-based pharmacophore models was superior to the retrieval rate of true inhibitors by inappropriate target-based pharmacophore models. Interestingly, the conformational search parameters had a minor impact on the profiling accuracy of the method. Automation of the entire protocol was realized later on by embedding all steps into a Pipeline Pilot workflow[62] and evaluated on a set of 81 HIV-1 protease pharmacophore models. On average, true HIV-1 protease inhibitors were more often mapped to a larger set of pharmacophore models than other protease

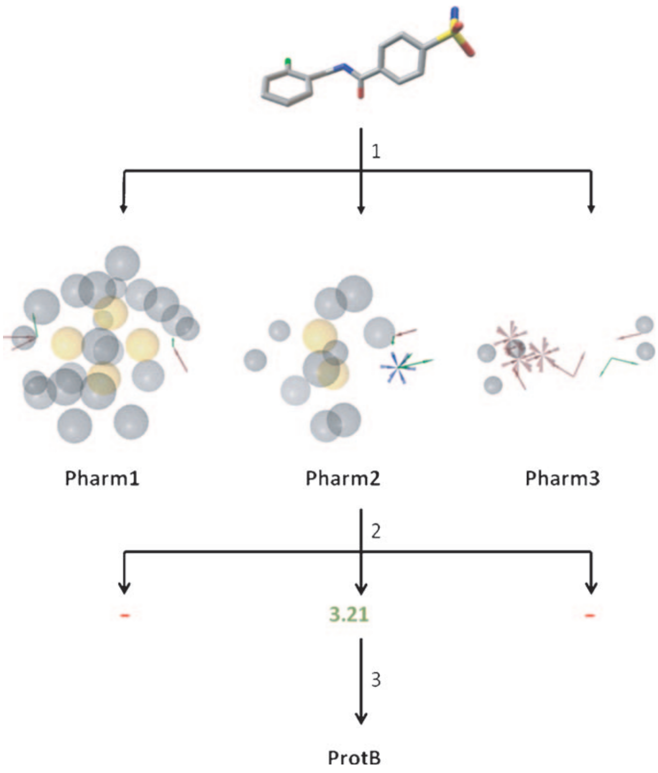


Figure 3. Pharmacophore-based parallel screening. A ligand is screened for matching predefined protein–ligand structure-based pharmacophoric features (step1). All matched pharmacophore models deduced from minimal fitness values (step2) result in a target hitlist and a target/pharmacological profile (step3).

inhibitors and far more than drug-like inactive and virtual compounds. About 30% of true inhibitors were incorrectly profiled because of inadequate conformational sampling. Larger scale profiling was described for a set of 321 PPAR agonists against a collection of 1 537 structure-based pharmacophore models from 181 unique targets.[63] Increasing the target space did not decrease the performance of the profiling method since the right target was ranked among the top 5 scorers for 71% of true actives.[63]

3.2 Success Story

Up to now, a single successful target assignment has been reported using pharmacophore-based ligand profiling methods. 16 secondary metabolites from the medicinal plant *Ruta graveolens* were screened against a dataset of 2208 pharmacophore models mixing both structure-based and ligand-based pharmacophores.[64] Experimental validation of virtual targets was focused on three proteins for both practical (bioassay availability) and rational reasons (pharmacophore match): acetylcholinesterase (AChE), human rhinovirus (HRV) coat protein and cannabinoid receptor type 2 (CB2). Despite several false positives were observed, binding of arborinine to AchE and HRV was confirmed in vitro, as well as binding of rutamarin to the CB2 receptor. Only a single false negative (6,7,8-trimethoxycoumarin for HRV) was observed in the in silico profiling method.

3.3 Pros and Cons

A first real advantage of pharmacophore profiling over docking-based methods is the pace (ca. 0.2 s/compound and hypothesis)[62] which enable the fast profiling of many molecules and even compound libraries when a specific target coverage/focus is desired. Targeted libraries[65] could thus be either designed or checked for a particular biological space. A second advantage lies in the fuzziness of pharmacophore hypotheses, explicitly inherent to the method since tolerances in matching pharmacophore features and various flavors of protein-based (exclusion spheres) and ligand-based restraints (ligand surface) are easy to add to pharmacophore hypotheses.[66] As for the previous docking-based approach, the more structure-based pharmacophore models generated from different protein-ligand complexes, the better handling of protein flexibility and multiple binding sites. Last, the method can be applied to any kind of pharmacophores notably to ligand-based pharmacophore hypotheses in order to enhance the target space coverage.

The parallel screening approach developed by Langer and co-workers is not usable for apo-proteins since pharmacophoric features are directly mapped on ligand atoms interacting with the co-crystallized protein. Ligand-free structure-based pharmacophore perception methods[67,68] exist but are likely to be much less accurate notably if they are generated in an automated manner. Homology models are also a priori excluded from the target list. Tailored-made pharmacophore hypotheses can still be generated from existing ligand-bound homology models for supplementing the existing collection. Unless a strong expertise, they are however supposed to be significantly less accurate than cocrystal-derived pharmacophores. The method is of course sensitive to crucial issues in pharmacophore perception. For example, the chemical diversity and representativeness of ligands used to generate pharmacophore collections will influence the in silico profiling. 'Populated' targets (e.g. HIV-1 protease, thrombin, cdk2) present in multiple copies in the PDB are a priori favored over 'rare' underrepresented targets. A normalized target score[63] is a good attempt to tackle this problem but more experience with the target ranking is required. A difficult parameter is to find the best balance between sensitivity and specificity of the generated pharmacophore hypotheses. The presence of too many features will enhance the specificity at the cost of a low sensitivity (recall of true actives). Conversely, an oversimplified representation will lead to a good sensitivity but a poor specificity. Likewise, pharmacophores dominated by hydrophobic features tend to be less specific, inclusion of at least one polar feature and exclusion volumes/surfaces is therefore recommended. Last but not the least, the definition and placement of pharmacophoric features[66] as well as the quality of the 3-D alignment[69] will directly influence the pharmacophore screen.

4 Comparing Protein–Ligand Binding Sites

4.1 Proof of Concept

A common assumption in chemogenomics is that similar receptors bind similar ligands.[70] Delineating remote binding site similarities for unrelated proteins is therefore a possible route for finding new targets for existing ligands (Figure 4). Assuming that similar ligands bind to similar cavities, function and ligands for a novel protein may be inferred from structurally similar liganded cavities. Since binding site similarities may be quite difficult to detect from amino acid sequences, 3-D computational methods for quantifying global or local similarities between protein cavities have been developed in the last decade.[71]

All described methods follow the same three-step flowchart. First, the structures of the two proteins to compare are parsed into meaningful 3-D coordinates in order to reduce the complexity of the pair-wise comparison. Typically, only key residues/atoms are considered and described by a limited number of points, which are labeled according to pharmacophoric, geometric and/or chemical properties of their neighborhood. Second, the two resulting patterns are structurally aligned using notably clique detection[72,73] and geometric hashing methods[74,75] to identify the maximum number of equivalent points. Last, a scoring function quantifies the number of aligned features in the form of

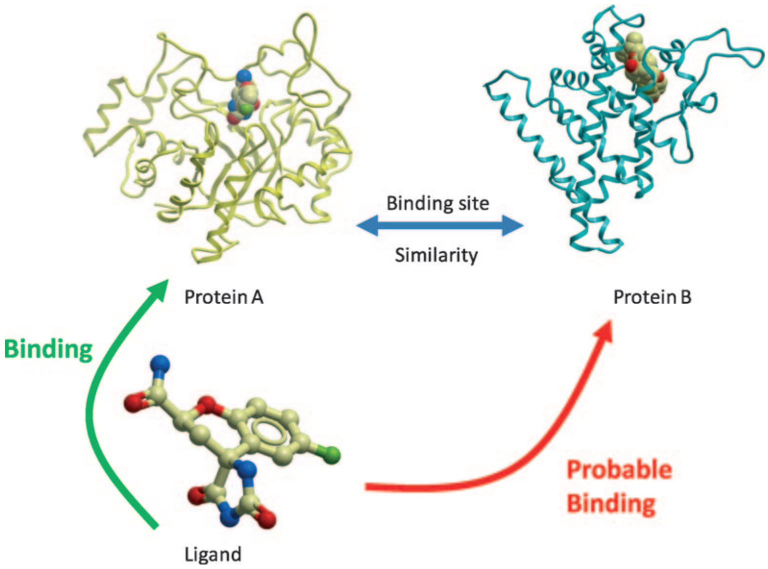


Figure 4. Remote binding site similarity as a tool for target identification. Unrelated proteins A and B sharing some similarity of their ligand-binding sites have a greater propensity to bind similar ligands.

rmsd, residue conservation or physicochemical property conservation.

Pure shape-based methods[76,77] have also been used to compute protein cavities and have led to exhaustive archives of protein-ligand binding sites (pocketomes)[76,78] that can be clustered or screened for envelope similarity to any query. One of the earliest cavity-guided explanations of unexpected ligand cross-reactivity was described by Weber et al.[79] Starting from the observation that many cyclooxygenase type-2 (COX2) inhibitors share with carbonic anhydrase (CA) inhibitors an arylsulfonamide moiety, known COX-2 inhibitors were tested for binding to various CA isoforms and revealed nanomolar binding affinities. A rational explanation to this cross-reactivity was obtained by comparing COX-2 inhibitor subpockets to a set of 9433 cavities with CavBase descriptors.[73] For two of three subcavities, CA subpockets were retrieved among the top-scoring entries. However, no global similarity could be detected between entire ligand-binding pockets of both enzymes.

4.2 Success Stories

Recently, the Soippa algorithm[80] was designed and applied to detect similarity between unrelated proteins. Starting from a graph representation of protein Cα atoms with a geometric potential and position-specific amino acid score assigned to each graph node,[81] a maximum common subgraph is detected to align two proteins. The SOIPPA method was successfully used to predict remote binding site similarities between the binding site of selective estrogen receptor modulators (SERMs) at the Erα receptor and the Sarcoplasmic Reticulum Ca2 ion channel ATPase protein (SERCA) transmembrane domain,[82] thus explaining kwown side effects of SERMs. Likewise, the NAD binding site of the Rossmann fold and the S-adenosyl-methionine (SAM)-binding site of SAM-methyltransferases were found similar and consequently permitted to predict the cross-reactivity of catechol-*O*-methyltransferase (COMT) inhibitors (entacapone, tolcapone) with the *M.tuberculosis* enoyl-acyl carrier protein reductase (InhA).[83]

Systematic pair-wise comparison of the staurosporine-binding site of the proto-oncogene Pim-1 kinase with 6,412 druggable protein-ligand binding sites[33] using the SiteAlign algorithm,[84] suggested that the ATP-binding site of synapsin I (an ATP-binding protein regulating neurotransmitter release in the synapse) may recognize the pan-kinase inhibitor staurosporine (Figure 5).[85]

Biochemical validation of this hypothesis was realized by competition experiments of staurosporine with ATP-γ^{35}S for binding to synapsin I. Staurosporine, as well as other more specific protein kinase inhibitors (roscovitine, quercetagetin), effectively bound to synapsin I with nanomolar affinities and promoted synapsin-induced F-actin bundling.[85] The selective Pim-1 kinase inhibitor quercetagetin was shown to be the most potent synapsin I binder (IC_{50} = 0.15 μM), in agreement with the predicted binding site similarities between synapsin I and various protein kinases.

Cavity comparisons can even leads to pocket deorphanization. The PocketPicker algorithm[77] was used to detect a cavity on the surface of a APOBEC3A structure, a protein able to inactivate retroviral genomes. Encoding the pocket as correlation vectors afforded to compare it to a set of 1300 ligand-binding sites from the PDBBind dataset.[86] Among top scoring entries were only nucleic acid-binding pockets.[87] Point mutation of the cavity-lining residues effectively led to mutants with a reduced antiviral activity. The pocket was shown to recognize the small 5.8S RNA[87] as a preliminary step to inactivate retroviral particles.

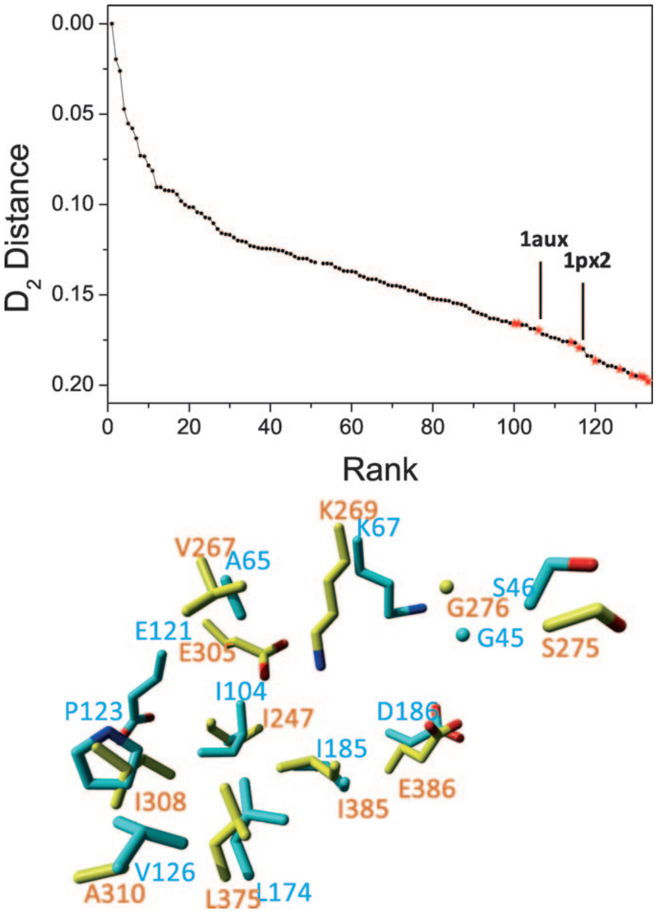


Figure 5. Similarity of the ATP-binding sites of Pim-1 kinase and of synapsin I. SiteAlign[84] virtual screening of 6 415 sc-PDB binding sites to the staurosporine binding site of human Pim-1 kinase (1yhs). Entries exhibiting similar binding site properties ($d2 < 0.20$; top panel) are ranked by decreasing d2 distance to the query. ATP-binding sites of protein kinases are displayed by dark circles, other binding sites by red stars. Two ATP-binding sites of synapsin I are labeled by their Protein Data Bank entry name (1aux, 1px2). The bottom panel shows 11 matching residues between the ATP-binding sites of Pim-1 (1yhs, cyan sticks) and synapsin I (yellow sticks). Residues are labeled according to the PDB residue numbering at their Cα atom.

4.3 Pros and Cons

Comparing protein-ligand binding sites is a fast method which presents the noticeable advantage to take into account protein space only. It avoids sampling the ligand conformational space and thus a putative incorrect definition of the ligand bioactive conformation. It can only be applied if the binding site comparison method is not too sensitive to variations of atomic coordinates. Noteworthy, the above-cited success stories relies on methods in which a simplified protein representation is used by mapping physicochemical properties to Cα protein atoms.[80,84]. This approach is still very sensitive to the quality of the protein-protein alignment utilized for scoring binding site similarities. In cases where only local and not global similarities can be detected between two unrelated protein cavities, this approach is likely to fail. Interestingly, alignment-free binding site comparison methods[88,89] have been recently reported to be as accurate as alignment-dependant site matching tools and suitable to detect local subpocket similarities. Although not necessary, the binding site reference to which all active sites are compared should be cocrystallized with a drug-like ligand to avoid ligand-induced fit phenomena. This is however a problem common to any structure-based approach. The inherent fuzziness embedded in some binding site comparison tools renders this approach less sensitive to moderate induced fits (up to 3.0 Å rmsd deviations)[84] than docking or pharmacophore searches.

5 Protein–Ligand Fingerprints

Since target fishing and ligand profiling requires working in true protein-ligand chemogenomic space, it is wise to encode a true protein-ligand complex in a single descriptor. Protein-ligand fingerprints are vectors in which both information on the ligand and the protein (cavity) are encoded (Figure 6).

Three studies on fingerprinting GPCR-ligand pairs in chemogenomic applications have been described.[90–92] In a pioneering work, Bock et al.[90] used rather standard 2-D topological and atomic descriptors for ligands, physicochemical properties of amino acid sequences for receptors, and concatenate feature vectors for both the receptor and the ligand in a single fingerprint. A support vector machine (SVM) model was trained on 5319 receptor-ligand pairs from the PDSP Ki database[93] to predict the *Ki* of any ligand to any GPCR and used to propose novel ligands for orphan GPCRs. Unfortunately, none of these predictions have been validated up to now.

Recently, Jacob et al. proposed a similar approach[91] on 4051 pairs from the Glida database[94] with the noticeable difference that the tensor product between vectors describing ligands and proteins were used to better delineate correlations between ligand and target features. A support vector machine (SVM) classifier was used to train and predict out-of-sample pairs but no convincing external test cases could unfortunately be provided.

In the fingerprint proposed by Weill et al.[92] ligand properties have been represented by standard descriptors (MACCS keys,[95] SHED descriptors[96]), protein cavities are encoded by a fixed length bit string describing pharmacophoric properties of a fixed number of binding site residues. Several machine learning classification algorithms (SVM, Random Forest, Naive Bayes) were trained on two sets of roughly 200 000 receptor–ligand fingerprints with a different definition of inactive decoys. Cross-validated models show excellent precision (>0.9) in distinguishing true from false pairs with a particular preference for random forest models. In most cases, predicting ligands for

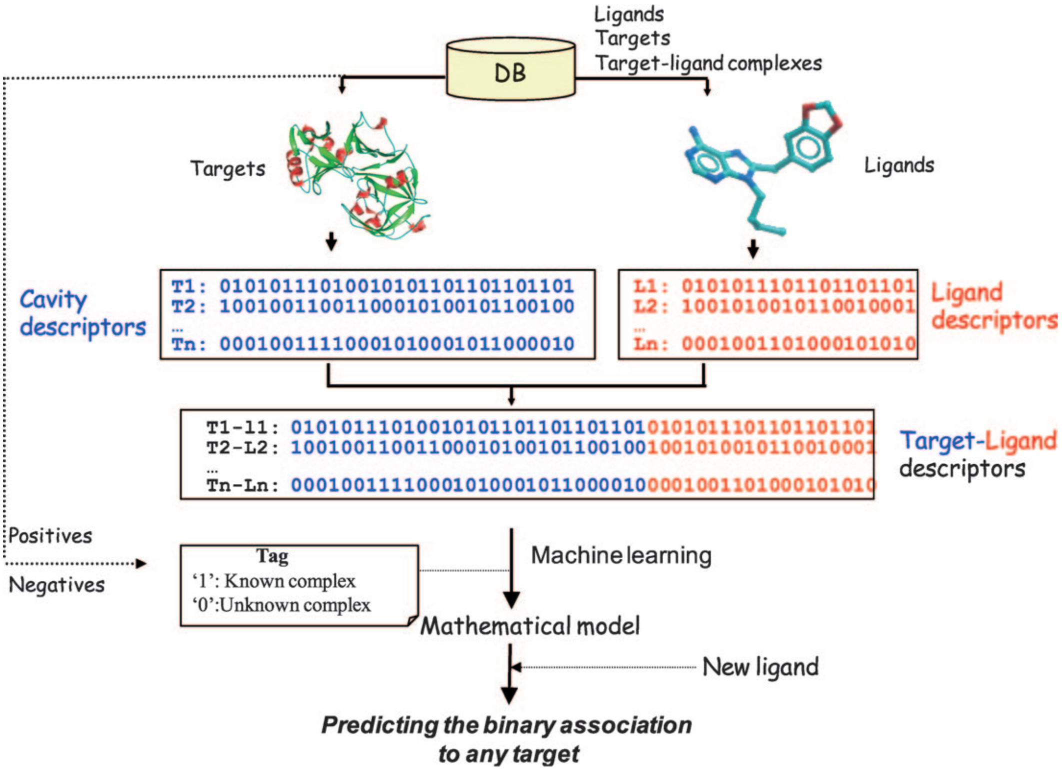

Ligands
Targets
Target-ligand complexes
DB
Targets
Ligands
Cavity descriptors
T1: 01010111010010101101101101101
T2: 10010011001100010100101100100
...
Tn: 00010011110001010001011000010
L1: 0101011101101101101
L2: 1001010010110010001
...
Ln: 0001001101000101010
Ligand descriptors
T1-l1: 010101110100101011011011011010101011101101101101
T2-L2: 100100110011000101001011001001001010010110010001
...
Tn-Ln: 000100111100010100010110000100001001101000101010
Target-Ligand descriptors
Positives
Negatives
Tag
'1': Known complex
'0':Unknown complex
Machine learning
Mathematical model
New ligand
Predicting the binary association to any target

Figure 6. Encoding and virtual screening of protein–ligand fingerprints.

a given receptor was easier than predicting receptors for a given ligand.

Although ligand profiling is possible, it probably requires a size-independent generic description of the ligand-binding site. Strömbergsson et al. proposed a generic amino acid sequence-based local descriptor for the protein.[97] Each residue of the protein is encoded by a set of nonoverlapping residue fragments describing its neighborhood. Ligands are classically described by a set of 27 diverse Dragon descriptors.[98] Starting from a set of 1 421 triplets (protein, ligand, p*Ki* inhibition constant) of known X-ray structure, the inhibition constant of 542 test complexes of unknown structures was predicted by SVM regression with a promising accuracy ($r^2 = 0.53$, root-mean square error of prediction of 1.50) with regard to the diversity of the test set.

A recent work form Bajorath and colleagues[99] suggests that simplified strategies for designing target-ligand kernels should be used since varying the complexity of the target kernel does not influence much the identification of ligands for virtually deorphanized targets. However, no true 3-D cavity descriptor has yet been reported as target kernel.

Noteworthy, protein–ligand fingerprints usually outperforms the corresponding ligand fingerprints in mining the target–ligand space.[92] Since they can be applied to a much larger number of receptors (e.g. orphan targets) than ligand-based fingerprints, protein-ligand fingerprints represent a novel and promising way to directly screen protein-ligand pairs in chemogenomic applications. Whether predictions are qualitative (binary association) or quantitative (p*Ki*), no information is derived about the putative binding mode of the protein-ligand under consideration. This is a considerable difference to the three other approaches reviewed herein but not necessarily a drawback. Hence, ligand profiling does not require outputting structural information on protein–ligand complexes. A target list (as short and specific as possible) should only be available to experimental validation. Only prospective applications will tell whether usage of protein-ligand fingerprints really represents a breakthrough with respect to either pure structure-based or ligand-based methods. Clearly, some expertise is required in the frequently underestimated training set elaboration phase. A deep knowledge of existing ligands is necessary. Biologically annotated ligand databases (e.g. MDDR,[95] Wombat,[100] ChEMBLdb,[101] GOSTAR,[102] PDSP,[93] PubChem BioAssay[103]) are invaluable sources but still need to be carefully used. Diversity of protein-ligand complexes is more important than just numbers. In addition, one has to carefully set-up decoys (unproductive protein-ligand binary associations) and to calibrate the ratio of active to inactive complexes. Beside the DUD dataset[104] which has been exclusively designed for docking purpose, no standard and robust protocol has yet been described for achieving this simple task in QSAR applications. Last, it should not be forgotten that target space is discontinuous. It is therefore not surprising that local models (for a particular biological space) usually outperform global models in predicting true protein-ligand associations.[92]

6 Summary

Although not in frontline of in-silico-based target assignment methods, structure-based approaches are getting increasing importance for filling the gap between chemical and biological spaces. Paradoxically target-centric methods have been less used than ligand-based approaches in the past although they have apparently generated more successful reports of novel target annotation for existing ligands. Considerable knowledge and expertise is however required for each of the structure-based methods used. It usually relies on years of accumulated experience and careful data mining. Getting the best of both ligand and protein worlds is expected to occur with the usage of newly emerging protein–ligand fingerprints that can be easily defined and browsed at an incomparable throughput. Next advances in structural coverage of the proteome[55] will undoubtedly boost this tendency.

Acknowledgements

I sincerely acknowledge many former and present lab members (*G. Marcou*, *J. Meslamani*, *P. Muller*, *N. Paul*, *C. Schalon*, *J. S. Surgand*, *N. Weill*,) for their invaluable work. I also thank *E. Kellenberger* and *T. Langer* for a critical reading of the manuscript.

References

[1] P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, C. L. Brooks, , 3rd, *J. Med. Chem.* **2004**, *47*, 3032 – 47.
[2] H. Kubinyi, in *Computer Applications in Pharmaceutical Research and Development*, (Ed: S. Ekins), Wiley Interscience, New York, **2006**, pp. 377 – 424.
[3] D. Rognan, *Br. J. Pharmacol.* **2007**, *152*, 38 – 52.
[4] A. Bender, J. L. Jenkins, M. Glick, Z. Deng, J. H. Nettles, J. W. Davies, *J. Chem. Inf. Model.* **2006**, *46*, 2445 – 56.
[5] J. Bajorath, *Curr. Opin. Chem. Biol.* **2008**, *12*, 352 – 8.
[6] S. Ekins, J. Mestres, B. Testa, *Br. J. Pharmacol.* **2007**, *152*, 9 – 20.
[7] V. V. Poroikov, D. A. Filimonov, Y. V. Borodina, A. A. Lagunin, A. Kos, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349 – 55.
[8] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391 – 405.
[9] J. Mestres, L. Martin-Couce, E. Gregori-Puigjane, M. Cases, S. Boyer, *J. Chem. Inf. Model.* **2006**, *46*, 2725 – 36.
[10] A. E. Cleves, A. N. Jain, *J. Med. Chem.* **2006**, *49*, 2921 – 38.
[11] Nidhi, M. Glick, J. W. Davies, J. L. Jenkins, *J. Chem. Inf. Model.* **2006**, *46*, 1124 – 33.
[12] J. H. Nettles, J. L. Jenkins, A. Bender, Z. Deng, J. W. Davies, M. Glick, *J. Med. Chem.* **2006**, *49*, 6802 – 10.
[13] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet, *Nat. Biotechnol.* **2007**, *25*, 197 – 206.
[14] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet, B. L. Roth, *Nature* **2009**, *462*, 175 – 81.
[15] F. Nigsch, A. Bender, J. L. Jenkins, J. B. Mitchell, *J. Chem. Inf. Model.* **2008**, *48*, 2313 – 25.
[16] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. E. Ferrin, *J. Mol. Biol.* **1982**, *161*, 269 – 88.
[17] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, C. R. Corbeil, *Br J. Pharmacol.* **2008**, *153 Suppl 1*, S7 – 26.
[18] I. Muegge, I. J. Enyedy, *Curr. Med. Chem.* **2004**, *11*, 693 – 707.
[19] M. Schapira, R. Abagyan, M. Totrov, *J. Med. Chem.* **2003**, *46*, 3045 – 3059.
[20] C. de Graaf, D. Rognan, *Curr. Pharm. Des.* **2009**, *15*, 4026 – 4048.
[21] K. Henrick, Z. Feng, W. F. Bluhm, D. Dimitropoulos, J. F. Doreleijers, S. Dutta, J. L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C. L. Lawson, J. L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E. L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin, H. M. Berman, *Nucleic Acids Res.* **2008**, *36*, D426 – 33.
[22] M. L. Lamb, K. W. Burdick, S. Toba, M. M. Young, A. G. Skillman, X. Zou, J. R. Arnold, I. D. Kuntz, *Proteins* **2001**, *42*, 296 – 318.
[23] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem.* **1998**, *19*, 1639 – 1662.
[24] W. M. Rockey, A. H. Elcock, *Proteins* **2002**, *48*, 664 – 71.
[25] M. Hendlich, A. Bergner, J. Gunther, G. Klebe, *J. Mol. Biol.* **2003**, *326*, 607 – 20.
[26] O. Roche, R. Kiyama, C. L. Brooks, , 3rd, *J. Med. Chem.* **2001**, *44*, 3592 – 8.
[27] K. Kitajima, S. Ahmad, S. Selvaraj, H. Kubodera, S. Sunada, J. An, A. Sarai, *Genome Inf.* **2002**, *13*, 498 – 499.
[28] A. C. Stuart, V. A. Ilyin, A. Sali, *Bioinformatics* **2002**, *18*, 200 – 1.
[29] Y. Z. Chen, D. G. Zhi, *Proteins* **2001**, *43*, 217 – 26.
[30] N. Paul, E. Kellenberger, G. Bret, P. Muller, D. Rognan, *Proteins* **2004**, *54*, 671 – 80.
[31] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727 – 48.
[32] E. Kellenberger, N. Foata, D. Rognan, *J. Chem. Inf. Model.* **2008**, *48*, 1014 – 25.
[33] E. Kellenberger, P. Muller, C. Schalon, G. Bret, N. Foata, D. Rognan, *J. Chem. Inf. Model.* **2006**, *46*, 717 – 27.
[34] G. Marcou, D. Rognan, *J. Chem. Inf. Model.* **2007**, *47*, 195 – 207.
[35] Z. Gao, H. Li, H. Zhang, X. Liu, L. Kang, X. Luo, W. Zhu, K. Chen, X. Wang, H. Jiang, *BMC Bioinformatics* **2008**, *9*, 104.
[36] J. J. Irwin, B. K. Shoichet, M. M. Mysinger, N. Huang, F. Colizzi, P. Wassam, Y. Cao, *J. Med. Chem.* **2009**, *52*, 5712 – 20.
[37] D. M. Lorber, B. K. Shoichet, *Protein Sci.* **1998**, *7*, 938 – 50.
[38] Q. T. Do, C. Lamy, I. Renimel, N. Sauvan, P. Andre, F. Himbert, L. Morin-Allory, P. Bernard, *Planta Med.* **2007**, *73*, 1235 – 40.
[39] Q. T. Do, I. Renimel, P. Andre, C. Lugnier, C. D. Muller, P. Bernard, *Curr. Drug Discov. Technol.* **2005**, *2*, 161 – 7.
[40] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470 – 89.
[41] H. Li, Z. Gao, L. Kang, H. Zhang, K. Yang, K. Yu, X. Luo, W. Zhu, K. Chen, J. Shen, X. Wang, H. Jiang, *Nucleic Acids Res.* **2006**, *34*, W219 – 24.
[42] J. Cai, C. Han, T. Hu, J. Zhang, D. Wu, F. Wang, Y. Liu, J. Ding, K. Chen, J. Yue, X. Shen, H. Jiang, *Protein Sci.* **2006**, *15*, 2071 – 81.

[43] A. M. Aronov, N. R. Munagala, I. D. Kuntz, C. C. Wang, *Antimicrob Agents Chemother.* **2001**, *45*, 2571–6.

[44] M. A. Fabian, W. H. Biggs, , 3rd, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. A. Carter, P. Ciceri, P. T. Edeen, M. Floyd, J. M. Ford, M. Galvin, J. L. Gerlach, R. M. Grotzfeld, S. Herrgard, D. E. Insko, M. A. Insko, A. G. Lai, J. M. Lelias, S. A. Mehta, Z. V. Milanov, A. M. Velasco, L. M. Wodicka, H. K. Patel, P. P. Zarrinkar, D. J. Lockhart, *Nat. Biotechnol.* **2005**, *23*, 329–36.

[45] S. Tietze, J. Apostolakis, *J. Chem. Inf. Model.* **2007**, *47*, 1657–72.

[46] S. Zahler, S. Tietze, F. Totzke, M. Kubbutat, L. Meijer, A. M. Vollmar, J. Apostolakis, *Chem. Biol.* **2007**, *14*, 1207–14.

[47] L. Yang, J. Chen, L. He, *PLoS Comput. Biol.* **2009**, *5*, e1000441.

[48] P. Muller, G. Lena, E. Boilard, S. Bezzine, G. Lambeau, G. Guichard, D. Rognan, *J. Med. Chem.* **2006**, *49*, 6768–78.

[49] R. L. DesJarlais, G. L. Seibel, I. D. Kuntz, P. S. Furth, J. C. Alvarez, P. R. Ortiz de Montellano, D. L. DeCamp, L. M. Babe, C. S. Craik, *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 6644–8.

[50] E. Rutenber, E. B. Fauman, R. J. Keenan, S. Fong, P. S. Furth, P. R. Ortiz de Montellano, E. Meng, I. D. Kuntz, D. L. DeCamp, R. Salto, et al., *J. Biol. Chem.* **1993**, *268*, 15343–6.

[51] G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, M. S. Head, *J. Med. Chem.* **2006**, *49*, 5912–31.

[52] S. L. McGovern, B. K. Shoichet, *J. Med. Chem.* **2003**, *46*, 2895–907.

[53] http://www.rcsb.org/pdb/static.do?p=general information/pdb statistics/index.html

[54] http://bioinfo-pharma.u-strasbg.fr/scPDB

[55] R. Nair, J. Liu, T. T. Soong, T. B. Acton, J. K. Everett, A. Kouranov, A. Fiser, A. Godzik, L. Jaroszewski, C. Orengo, G. T. Montelione, B. Rost, *J. Struct. Funct. Genomics* **2009**, *10*, 181–91.

[56] http://www.proteinmodelportal.org/?

[57] U. Pieper, N. Eswar, B. M. Webb, D. Eramian, L. Kelly, D. T. Barkan, H. Carter, P. Mankoo, R. Karchin, M. A. Marti-Renom, F. P. Davis, A. Sali, *Nucleic Acids Res.* **2009**, *37*, D347–54.

[58] C. B. Rao, J. Subramanian, S. D. Sharma, *Drug Discov. Today* **2009**, *14*, 394–400.

[59] T. M. Steindl, D. Schuster, C. Laggner, T. Langer, *J. Chem. Inf. Model.* **2006**, *46*, 2146–57.

[60] G. Wolber, T. Langer, *J. Chem. Inf. Model.* **2005**, *45*, 160–9.

[61] Accelrys, Inc., San Diego, CA 92131, USA.

[62] T. M. Steindl, D. Schuster, C. Laggner, K. Chuang, R. D. Hoffmann, T. Langer, *J. Chem. Inf. Model.* **2007**, *47*, 563–71.

[63] P. Markt, D. Schuster, J. Kirchmair, C. Laggner, T. Langer, *J. Comput. Aided Mol. Des.* **2007**, *21*, 575–90.

[64] J. M. Rollinger, D. Schuster, B. Danzl, S. Schwaiger, P. Markt, M. Schmidtke, J. Gertsch, S. Raduner, G. Wolber, T. Langer, H. Stuppner, *Planta Med.* **2009**, *75*, 195–204.

[65] I. Akritopoulou-Zanze, P. J. Hajduk, *Drug Discov. Today* **2009**, *14*, 291–7.

[66] G. Wolber, T. Seidel, F. Bendix, T. Langer, *Drug Discov. Today* **2008**, *13*, 23–9.

[67] C. Barillari, G. Marcou, D. Rognan, *J. Chem. Inf. Model.* **2008**, *48*, 1396–410.

[68] J. Chen, L. Lai, *J. Chem. Inf. Model.* **2006**, *46*, 2684–91.

[69] A. R. Leach, V. J. Gillet, R. A. Lewis, R. Taylor, *J. Med. Chem.* **2010**, *53*, 539–58.

[70] T. Klabunde, *Br. J. Pharmacol.* **2007**, *152*, 5–7.

[71] E. Kellenberger, C. Schalon, D. Rognan, *Curr. Comput. Aided Drug Des.* **2008**, *4*, 209–220.

[72] K. Kinoshita, J. Furui, H. Nakamura, *J. Struct. Funct. Genomics* **2002**, *2*, 9–22.

[73] S. Schmitt, D. Kuhn, G. Klebe, *J. Mol. Biol.* **2002**, *323*, 387–406.

[74] N. D. Gold, R. M. Jackson, *J. Mol. Biol.* **2006**, *355*, 1112–24.

[75] A. Shulman-Peleg, R. Nussinov, H. J. Wolfson, *J. Mol. Biol.* **2004**, *339*, 607–33.

[76] J. An, M. Totrov, R. Abagyan, *Mol. Cell Proteomics* **2005**, *4*, 752–61.

[77] M. Weisel, E. Proschak, G. Schneider, *Chem. Cent J.* **2007**, *1*, 7.

[78] M. Weisel, E. Proschak, J. M. Kriegl, G. Schneider, *Proteomics* **2009**, *9*, 451–9.

[79] A. Weber, A. Casini, A. Heine, D. Kuhn, C. T. Supuran, A. Scozzafava, G. Klebe, *J. Med. Chem.* **2004**, *47*, 550–7.

[80] L. Xie, P. E. Bourne, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5441–6.

[81] L. Xie, P. E. Bourne, *BMC Bioinformatics* **2007**, *8 Suppl 4*, S9.

[82] L. Xie, J. Wang, P. E. Bourne, *PLoS Comput. Biol.* **2007**, *3*, e217.

[83] S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, P. E. Bourne, *PLoS Comput. Biol.* **2009**, *5*, e1000423.

[84] C. Schalon, J. S. Surgand, E. Kellenberger, D. Rognan, *Proteins* **2008**, *71*, 1755–78.

[85] E. de Franchi, C. Schalon, M. Messa, F. Onofri, F. Benfenati, D. Rognan, manuscript in preparation.

[86] R. Wang, X. Fang, Y. Lu, S. Wang, *J. Med. Chem.* **2004**, *47*, 2977–80.

[87] B. Stauch, H. Hofmann, M. Perkovic, M. Weisel, F. Kopietz, K. Cichutek, C. Munk, G. Schneider, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12079–84.

[88] N. Weill, D. Rognan, *J. Chem. Inf. Model.* **2010**, *50*, 123–35.

[89] K. Yeturu, N. Chandra, *BMC Bioinformatics* **2008**, *9*, 543.

[90] J. R. Bock, D. A. Gough, *J. Chem. Inf. Model.* **2005**, *45*, 1402–14.

[91] L. Jacob, B. Hoffmann, V. Stoven, J. P. Vert, *BMC Bioinformatics* **2008**, *9*, 363.

[92] N. Weill, D. Rognan, *J. Chem. Inf. Model.* **2009**, *49*, 1049–62.

[93] http://pdsp.med.unc.edu/.

[94] Y. Okuno, A. Tamon, H. Yabuuchi, S. Niijima, Y. Minowa, K. Tonomura, R. Kunimoto, C. Feng, *Nucleic Acids Res.* **2008**, *36*, D907–12.

[95] Symyx Technologies, Inc. , CA, Santa Clara.

[96] E. Gregori-Puigjane, J. Mestres, *J. Chem. Inf. Model.* **2006**, *46*, 1615–22.

[97] H. Strombergsson, P. Daniluk, A. Kryshtafovych, K. Fidelis, J. E. Wikberg, G. J. Kleywegt, T. R. Hvidsten, *J. Chem. Inf. Model.* **2008**, *48*, 2278–88.

[98] TALETE srl, 20124 Milano, Italy.

[99] A. M. Wassermann, H. Geppert, J. Bajorath, *J. Chem. Inf. Model.* **2009**, *49*, 2155–67.

[100] http://www.sunsetmolecular.com/

[101] http://www.ebi.ac.uk/chembldb/index.php

[102] http://www.gostardb.com/

[103] http://pubchem.ncbi.nlm.nih.gov/

[104] N. Huang, B. K. Shoichet, J. J. Irwin, *J. Med. Chem.* **2006**, *49*, 6789–801.

Received: December 10, 2009
Accepted: February 3, 2010
Published online: March 5, 2010