

High-throughput structure-based pharmacophore modelling as a basis for successful parallel virtual screening

Theodora M. Steindl · Daniela Schuster ·
Gerhard Wolber · Christian Laggner ·
Thierry Langer

Received: 18 June 2006 / Accepted: 9 August 2006 / Published online: 29 September 2006
© Springer Science+Business Media B.V. 2006

Abstract In order to assess bioactivity profiles for small organic molecules we propose to use parallel pharmacophore-based virtual screening. Our aim is to provide a fast, reliable and scalable system that allows for rapid in silico activity profile prediction of virtual molecules. In this proof of principle study, carried out with the new structure-based pharmacophore modelling tool LigandScout and the high-performance database mining platform Catalyst, we present a model work for the application of parallel pharmacophore-based virtual screening on a set of 50 structure-based pharmacophore models built for various viral targets and 100 antiviral compounds. The latter were screened against all pharmacophore models in order to determine if their known biological targets could be correctly predicted via an enrichment of corresponding pharmacophores matching these ligands. The results demonstrate that the desired enrichment, i.e. a successful activity profiling, was achieved for approximately 90% of all input molecules. Additionally, we discuss descriptors for output validation, as well as various aspects influencing the analysis of the obtained activity profiles, and the effect of the searching mode utilized for screening. The results of the study presented here clearly indicate that pharmacophore-based parallel screening comprises a reliable in silico method

to predict the potential biological activities of a compound or a compound library by screening it against a series of pharmacophore queries.

Keywords Bioactivity profiling · Virtual screening · Pharmacophore modelling · LigandScout · Structure-based pharmacophores · Database mining · Parallel screening

Introduction

Considerable effort has been devoted in recent years to compressing the early phases of the pharmaceutical industry's drug discovery process. In particular, as combinatorial chemistry and high throughput screening considerably increased the numbers of new chemical entities to be studied, the expectations to find new bio-active molecules among these compounds were high at the beginning. However, neither HTS nor combinatorial chemistry yielded so far the expected success, and therefore in silico-based screening approaches emerged and largely evolved [1]. Virtual screening is established as one of the most important computational technique used to separate wanted from unwanted molecules within compound libraries, which has to be done as early as possible in order to reduce drug discovery costs. If the three-dimensional (3D) structure of the target is not known, pharmacophore models play an important role among the compound selection filters that have been constructed for retrieving bio-active compounds. If there is an experimentally determined high-resolution 3D structure of the target available, structure-based drug design can be performed which is normally tightly associated with

T. M. Steindl · G. Wolber · T. Langer (✉)
Inte:Ligand GmbH, Clemens Maria Hofbauer-Gasse 6, 2344
Maria Enzersdorf, Austria
e-mail: thierry.langer@uibk.ac.at

D. Schuster · C. Laggner · T. Langer
Institute of Pharmacy, Computer Aided Molecular Design
Group, and Centre of Molecular Biosciences, University of
Innsbruck, Innrain 52c, 6020 Innsbruck, Austria

docking. In this case, in a first step ligands are flexibly aligned into a rigid macromolecule environment and then the tightness of the interaction by different scoring functions is calculated. However, there are no scoring functions available that are equally applicable to any kind of targets, which makes docking and scoring again to a time-consuming task which cannot be automated for large number of targets [2]. Therefore, pharmacophore-based screening methods, which easily can be built into an automatic workflow environment, have gained again substantial interest [3, 4]. The biggest advantage in this field compared to docking and scoring is search speed which makes pharmacophore-based screening applicable to large-scale parallel screening (PS) [5].

Assuming that one has a multitude of pharmacophore models representing a variety of different pharmacological targets, it appears interesting to have a system to screen a molecule simultaneously against all these models. In such a case, one would obtain a hit list of matching pharmacophore models and could link them to the biological targets thereby enabling an *in silico* identification of macromolecular systems that will possibly be influenced by this ligand. That is exactly the aim that our PS approach seeks to achieve: a compound screened against a set of high-quality pharmacophore models will provide a hit list of mapping models, the so-called pharmacophoric profile. According to the targets encoded by these models, a pharmacological profile for the compound can be set up, however, restricting the prediction to the cases where reliable models exist. On one hand, this enables the characterization of the biological properties of new compounds by virtual screening experiments; on the other hand, the sphere of action for substances with already established biological activities could be enlarged. The latter represents an often highly successful concept in drug development. Of course if the PS system includes the appropriate models, predictions could also cover toxicity, side-effects, and metabolic pathways [6–11]. Another interesting aspect in PS is that the behaviour of a compound in the system can point towards promiscuity and therefore be a warning signal. Further applications of this approach might be fine-tuning of early results in high-throughput screening or ideas for the identification of targets hit by natural products that are therapeutically used because of long-time experience but not mechanistically characterized.

A system for PS therefore requires: (1) A large set of pharmacophore models including the availability of a fast and reliable tool for their automatic generation and (2) a high-speed screening platform to test one compound against a variety of models, which should

also allow for analysis, visualization and facile interpretability of the output data.

Our aim is to provide such an automatic system for fast virtual activity profiling of compounds. The number of pharmacophore hypotheses in our system is constantly growing targeted at extensive coverage of all available drugable targets. The present application study is focussed on the validation of a fraction of this system. Thereby, the PS approach was tested on a test set of antiviral compounds screened manually with a set of structure-based pharmacophore models. The main question, which was addressed is how fast and reliable activity profiling would be possible using PS which includes how well the test compounds are attributed with known biological activities. A similar concept using 22 diverse targets and ligand-based virtual screening with models based on molecular similarity has been published recently [12]. The study design comprised 100 antiviral substances, which were screened against 50 pharmacophore hypotheses for several viral targets utilizing two different search algorithms. For analysis of the results we propose four descriptors that allow for quantifying the pharmacophoric profiles of the compounds and quote several examples to explain their meaning and interpretation. Although profile prediction is multi-factorial and up to a certain point depending on the user's preference and mindset, several guidelines can be derived. For most ligands the results clearly show an enrichment of correct models in the pharmacophore hit lists. Only in approximately 10% of the cases the data point at a false target and therefore give misleading profiles.

High-throughput pharmacophore modelling

For building the pharmacophore models used in this study, we applied our new program LigandScout [13], which is a software tool that allows to rapidly and transparently derive 3D chemical feature-based pharmacophores from structural data of macromolecule/ligand complexes in a fully automated and convenient way. LigandScout starts with a macromolecule/ligand complex and automatically detects bound ligands creating a standard residue hull around the non-standard residues. The advanced ligand bond interpretation is based on geometric interpretation as well as a matching algorithm to optimally distribute double bonds among sp² atoms [14]. The position of the ligand within the macromolecule is visualized using an animated protein–ligand handling that allows the user to zoom back into the protein without modifying the macromolecule at any time. From the protein–ligand

interaction pocket a pharmacophore is derived by identifying complementary interactions following extensive heuristics consisting of chemically and geometrically elaborated rules as described in [13]. Hydrogen bonds are represented as vectors optionally including projected points, aromatic PI-interactions are represented by planes, and lipophilic areas are represented as a set of spheres. Steric constraints in the form of inclusion volume spheres are added to make sure that lipophilic molecule parts are matched correctly in a virtual screening run. Once a pharmacophore is created, it can be aligned to imported or extracted molecules. Unlike other programs, the alignment is based on pharmacophoric points rather than on atomic contributions and thus better reflects the way the small molecule presents itself to the active site of the macromolecule. From several molecules or pharmacophores, a shared feature pharmacophore can be derived to determine common features, which then can be exported to several virtual screening platforms. LigandScout runs on all common operating systems and several successful application examples have been published [6, 15, 16].

Pharmacophore-based parallel screening: application example

Target proteins used in this study

Our study involved the selection of several ligand binding sites belonging to five viral target proteins, which were represented in the PS system by sets of chemical feature-based pharmacophore models. In order to be selected for our application example, a target had to fulfil certain criteria: since pharmacophore hypotheses were derived in a structure-based

approach, the existence of a sufficient number of complexes from the Protein Database (PDB) [17] was a requirement. Further, we focussed on proteins, whose inhibition offers therapeutic strategies in the combat of highly relevant viral diseases. These include human immunodeficiency virus (HIV) infection, influenza, common cold and hepatitis C [18]. To increase the applicability of our conclusions we aimed to provide diversity concerning the nature of the proteins as well as inhibitory mechanisms. The macromolecular targets used are listed in Table 1.

Pharmacophore model generation

For each of the selected target proteins, a set of ten pharmacophore models were generated based on receptor–inhibitor complexes from the PDB. The three allosteric sites of HCV polymerase were represented by three, five and two models, respectively. Structure-based pharmacophore model generation was performed with the software LigandScout [13] using the default settings and the standard workflow.

Although the built-in heuristics has been shown to work with high precision, thorough checking and processing of the LigandScout output is highly recommended: we therefore inspected all the extracted ligands and the proposed interaction patterns manually and compared them with information from original literature. LigandScout definition allows an atom or functional group to serve as root for multiple features, thus resembling the situation in the binding pocket. For instance, an amine might function as hydrogen bond donor and as positive ionizable feature, a hydroxyl group can face the appropriate interactions partners to accept and donate hydrogen bonds. Since such multiple features are not supported in the screening software platform Catalyst [19], which later was employed for

Table 1 Target proteins used for structure-based pharmacophore modelling

Target protein	Disease	Function	Mechanism of inhibition
HIV protease	HIV infection, AIDS	Cleavage of gag and gag-pol precursor polyproteins into mature, structural and functional viral proteins	Inhibition at active site
HIV reverse transcriptase	HIV infection, AIDS	Synthesis of a double-stranded DNA from virus RNA for integration into host chromosomal DNA and transcription to viral genomic and messenger RNA	Inhibition at allosteric site
Influenza virus neuraminidase	Influenza	Viral envelope glycoprotein involved in viral release, cleavage of sialic acid residues of new virus particles and host membranes	Inhibition at active site
Human rhinovirus coat protein	Common cold	Attachment to host cell receptor, viral entry and uncoating	Binding in hydrophobic pocket (capsid stabilization)
Hepatitis C virus RNA polymerase	Hepatitis C	Viral replication, transcription of genomic RNA	Inhibition at three different allosteric sites

DB screening, the LigandScout pharmacophores had therefore to be simplified when multiple features were detected. The models consist of the default Catalyst features except for a hydrogen bond acceptor whose definition was enlarged to also include fluorine atoms. Excluded volume spheres placed automatically within LigandScout and shape constraints added manually within Catalyst provide steric restrictions, where necessary, for sufficient selectivity of the models. For model validation, a fast flexible screening run within the Derwent World Drug Index (WDI) [20] containing more than 60 000 entries was performed. We found that 80% of the generated pharmacophore hypotheses (40 models) retrieve less than 500 compounds from this database. Only four of the models (8%) retrieved more than 1 000 hits. Aside from the total number of compounds also the occurrence of known active molecules in the hit lists was checked. According to these investigations the major part of the models can be attributed high selectivity. Only few models exhibited poorer performance. However, we decided to keep them in the pharmacophore set in order to study their impact on the results of the PS approach and to determine how activity profile accuracy is related to pharmacophore selectivity.

Antiviral compounds used as test set molecules

A total of 20 inhibitors for each of the five viral proteins were collected from PDB complexes, when available, and from various literature sources ensuring common binding modes. Example structures are shown in Chart 1. Inhibitors for HIV protease target the proteolytic site. Earlier compounds display large peptidic or smaller peptidomimetic characters, such as ritonavir (**1**) or amprenavir (**2**), while more recently developed substances comprise non-peptidic scaffolds, like the cyclic urea inhibitors, e.g. DMP 323 (**3**) [21]. For reverse transcriptase (RT) we herein address solely inhibitors acting on the non-nucleoside allosteric site of the enzyme including early structures like nevirapine (**4**) or efavirenz (**5**) and more recently established compounds like UC 781 (**6**) combining higher potency and resilience to drug resistance [21, 22]. Neuraminidase (NA) inhibitors used in our study cover initial transition state analogues, zanamivir (**7**)-like compounds which still carry the polar glycerol side chain of the sialic acid substrate, substances like oseltamivir acid (**8**), where this polar part is exchanged for lipophilic residues opening a hydrophobic pocket, as well as compounds with cyclohexyl or aromatic central parts, e.g. BANA 113 (**9**) [23]. For occupancy of the hydrophobic pocket within the human rhino virus

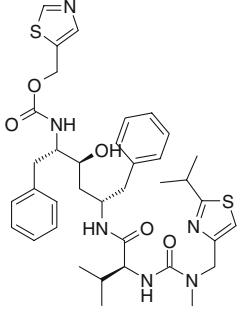
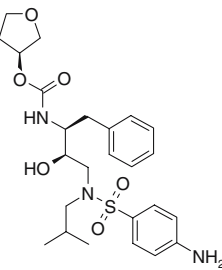
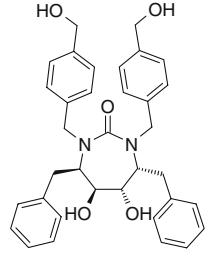
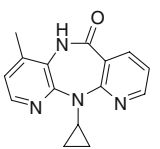
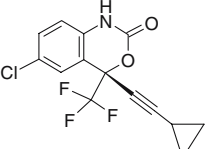
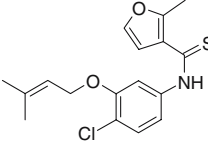
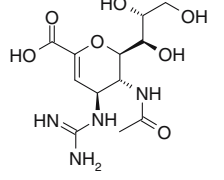
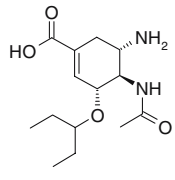
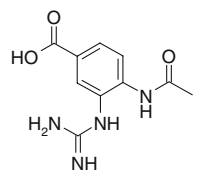
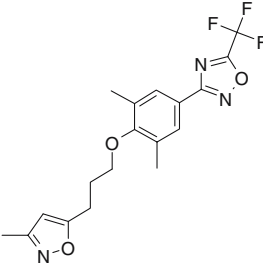
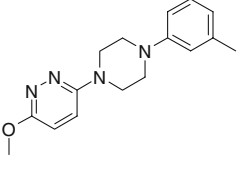
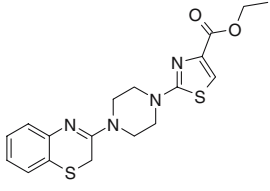
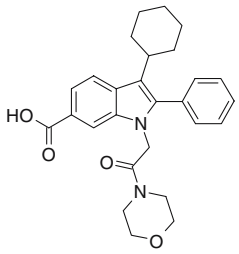
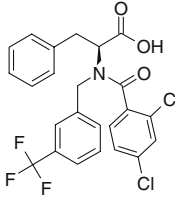
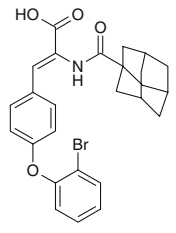
Chart 1 Examples of antiviral compounds from different structural classes used in PS study

(HRV) coat protein a long stretched topology and a predominantly hydrophobic character of the ligand is required. The so-called WIN compounds present the most prominent class of inhibitors with pleconaril (**10**) as figurehead. WIN compounds generally contain three (aromatic) rings: the central ring—a phenoxy moiety—is linked to an isoxazole by an aliphatic chain of various lengths and to the third ring via a single bond. But we also aimed to include substances with other structural composition, like R 61837 (**11**) or SDZ 880–061 (**12**) [24]. For the three HCV polymerase allosteric sites seven, five and eight inhibitors were selected, respectively.

Compound structure models of these 100 inhibitors were generated by a standard procedure, starting from SMILES code, followed by 3D structure generation within the Catalyst software. In a subsequent step structure minimization was performed followed by conformational model generation using the following parameter settings: a maximum of 250 conformers, the best generation algorithm, and an energy threshold of 20 kcal above the calculated lowest energy conformation.

Parallel screening procedure and analysis of results

All 100 antiviral compounds were screened against the 50 structure-based pharmacophore models using both the fast and the best flexible search algorithm within Catalyst. The results were analysed in order to determine which molecule was retrieved as active by which models. In order to quantify the resulting data we decided to set up four validation descriptors that were calculated to express the quality of activity profile prediction: for each of the compounds a pharmacophore hit list was generated, i.e. a set of hypotheses by which this particular ligand is retrieved. A correct model represents the target for which the ligand is a known inhibitor. An incorrect model means that the pharmacophore hypothesis was built for another target. The first step was the calculation of the percentage of correct and incorrect models in the pharmacophore hit list. Furthermore, we were interested to what extent the available correct models were retrieved and which false target was most extensively identified. The relation between the last two parameters was found critical for activity profile prediction accuracy. Additionally, the pharmacophore models were validated in a similar manner inspecting the hit map vertically: the percent-

 <p>1</p>	 <p>2</p>	 <p>3</p>
Inhibitors of HIV Protease Active Site		
 <p>4</p>	 <p>5</p>	 <p>6</p>
Inhibitors of HIV Reverse Transcriptase Allosteric Site		
 <p>7</p>	 <p>8</p>	 <p>9</p>
Inhibitors of Influenza Virus Neuraminidase Active Site		
 <p>10</p>	 <p>11</p>	 <p>12</p>
Inhibitors of Hydrophobic Pocket in Human Rhinovirus Coat Protein		
 <p>13</p>	 <p>14</p>	 <p>15</p>
Inhibitors of Three Different Allosteric Sites of Hepatitis C Virus RNA Polymerase		

 Springer

Fig. 1 Hit matrix representing the results from PS approaches using the pharmacophore models and the Catalyst *fast flexible search* algorithm. *Green signal*: compound is found with model built from correct target; *Red signal*: compound is found with a model from another target; *Light green highlighting*: Correct activity profiling for a compound; *Red highlighting*: Incorrect activity profiling for a compound; *White*: Lack of profile predictability

age of known active versus inactive compounds in the obtained hit list plays a vital role for pharmacophore quality. Furthermore, it is of interest to determine the most frequently found false inhibitor class and the reasons, why its members are found. Finally, the switch from fast to best flexible search algorithm was analysed for identifying the impact on the selectivity of the PS system and the accuracy of activity profiling.

Results and discussion

The results of our PS experiments performed as described above are visualized in matrix hit maps (Figs. 1, 2). The pharmacophore models are represented as columns and the ligands as rows; green boxes indicate correct retrieval of a compound by a model for the correct target, while red boxes indicate retrieval of a compound with a model from another target, where the molecule is presumably inactive. A critical issue, however, in this context must be noted: in our system, we assume that one molecule displays activity only at one particular target and is inactive at all the others—an assumption which has not been tested and verified in biological assays. This kind of simplification is difficult to avoid because without it no analysis and validation of the activity profiling outcome would be possible [12]. Implicitly, it should be kept in mind that what is referred as an incorrect or false model or a misleading prediction later on, might in fact be correct and only indicate a new and so far unknown activity spectrum of a compound.

When analysing the hit matrix, a clear enrichment is observable: inspecting the maps horizontally, for many molecules preferentially green signals appear, that is models representing the correct target. However, in order to quantify the pharmacophoric profile obtained in the screening experiments for a compound, i.e. a list of pharmacophores mapping this compound, four descriptors for validation were chosen: primarily, the pharmacophore hit list was inspected for the fraction of hypotheses representing the correct target and those built for another target. Descriptor 1 (D1): percentage of correct models in profile list. Descriptor 2 (D2):

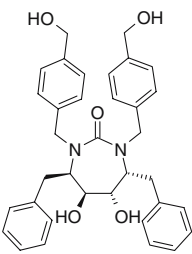
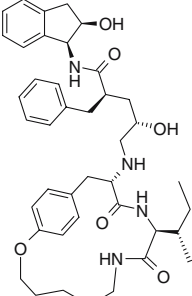
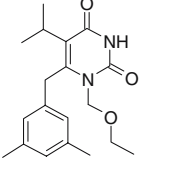
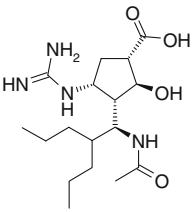
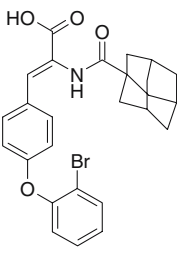
Fig. 2 Hit matrix representing the results from PS approaches using the pharmacophore models and the Catalyst *best flexible search* algorithm (for description of colour coding see Fig. 1)

percentage of incorrect models in profile list. Already this early step provides information for the assessment of profile prediction quality. The higher D1 is in comparison to D2, the better the prognosis for correct profiling. However, $D1 > D2$ does not necessarily mean that a correct prediction has been achieved. This would only be true for a system with an equal number of models for each target. Because of the fact that for some target proteins just very few PDB entries have been published, whilst others have been characterized and investigated in numerous complexes such an equalized situation will not exist in the system, requiring further description variables. Descriptor 3 (D3): percentage of models retrieved for the correct target. Furthermore, models representing other targets have to be investigated. Out of all wrongly identified targets the one with the best retrieval of the ligand—i.e. the highest percentage of mapping models—is the most interesting one, because this is the target, to which the activity profile might misleadingly point. Descriptor 4 (D4): percentage of models for one specific false target. Examples of several PS results showing the calculation of these descriptors and their impact for activity profile prediction are listed in Table 2.

The accuracy of our activity profiling approach mainly depends on the last two descriptors D3 and D4. As described above it is not that important, how many correct versus incorrect models can be found in the pharmacophore hit list. The critical point is how the false models are distributed: if they are almost equally distributed amongst a variety of binding sites not showing considerable enrichment anywhere, PS will not be guided to a false target protein. However, caution is required for ligands found frequently and by a multitude of diverse pharmacophore queries, because this could point to a promiscuous and therefore unwanted compound. On the other hand, if there is a clear enrichment of the retrieved pharmacophore model for one particular false target, one might be misled to believe the compound would be active there. It then depends, whether the correct or one specific false target is better represented in the pharmacophore hit list, i.e. with a higher percentage of models. This question can be addressed determining the ratio between D3 and D4. A ratio higher or at least equal to one would point to the correct protein target and therefore gives beneficial profiling information. On the contrary, if the D3:D4 ration is lower

		HIV-protease	HIV-RT	NA	HRV	HCV1	HCV2	HCV3
		1met-1 1ajx-1 1ajv-1 1d4h-1 1g2k-1 1hvr-1 1hwr-1 1hvp-1 1dmp-1 1g35-1 1rt1-1-s 1ikv-1-s 1bqm-1 1hmv-1-s 1s1w-2 1rt7-1 1rt6-1-s 1ikx-1-s 1rt5-1 1c1c-1 2qwc-1-s 1iny-1-s 1inw-1-s 1ivd-1-s 2qwb-1-s 1bji-1 2qwf-1 2qwk-1 2qwe-1 2qwd-1 1r08-1-s 2m2-1 2m4-1 2rt1-1-s 2rt6-1 2rt7-1-s 2rt8-1-s 1ncr-2-s 2rs5-1-s 2brk-1 2brk-1-s 1nhu-1 1nhv-2-s 1yz-2 1yx-1 1os5-2-s 1z4u-1-s 1ywf-1-s	1rt1-mkc 1ikv-efz 1bqm-hby 1hmv-tbo 1s1w-uc1 1rt7-uc4 1rt6-uc3 1ikx-pnu 1rt5-uc2 1c1c-612 1c1b-gca 1ep4-s11 1hpb-aap 1s1v-lnk 1jlc-ftc 1jfl-nvp 1s6q-lpb 1s9e-adb 1tvr-tb9 1s9g-abz	2qwc-dan 1iny-eqp 1inw-axp 1ivd-st1 2qwb-sia 1bji-g21 2qwf-g20 2qwk-g39 2qwe-gna 2qwd-4am 2qwg-g28 1f8e-49a 1a4q-dpc 1f8d-9am BCX1812 BCX1923 BCX1898 BANA113 BANA106 182251-67-8 1r08-w42 2m2-w43 2rt4-w71 2rt1-w8r 2rt6-w35 2rs1-w84 2rt7-w33 2rs3-w59 1ncr-w11 2rs5-w56 2hwbWin56291 1qjyWin65099 1r09R61837 1qjyWin61209 1vthSDZ880-061 1hvhSDZ35-682 Win54221 Win56287 2hweWin54954 1qjyWin68934	2brl-poo 2brk-cmf 861966-42-9 861966-61-2 861965-95-9 861965-76-6 861965-88-0	1nhu-153 1nhv-154 1yz-jpc 1yx-ipc 1os5-nh1	1z4u-ph9 1ywf-ph7 639517-93-4 855301-46-1 860015-79-8 639518-06-2 PNU248809 855301-44-9	
HIV-protease inhibitors	1met-dmp 1ajx-ah1 1ajv-nmb 1d4h-beh 1g2k-nm1 1hvr-xk2 1hwr-216 1hvp-478 1dmp-450 1g35-ahf 1d4i-beg 1ody-lp1 1hrow-rit 1b6k-pi5 1d4k-pi8 1z1r-hbh 1d4j-msc 1qbu-846 1hvh-q82 1mtl-phm							
HIV RT inhibitors	1rt1-mkc 1ikv-efz 1bqm-hby 1hmv-tbo 1s1w-uc1 1rt7-uc4 1rt6-uc3 1ikx-pnu 1rt5-uc2 1c1c-612 1c1b-gca 1ep4-s11 1hpb-aap 1s1v-lnk 1jlc-ftc 1jfl-nvp 1s6q-lpb 1s9e-adb 1tvr-tb9 1s9g-abz							
NA inhibitors	2qwc-dan 1iny-eqp 1inw-axp 1ivd-st1 2qwb-sia 1bji-g21 2qwf-g20 2qwk-g39 2qwe-gna 2qwd-4am 2qwg-g28 1f8e-49a 1a4q-dpc 1f8d-9am BCX1812 BCX1923 BCX1898 BANA113 BANA106 182251-67-8							
HRV coat protein inhibitors	1r08-w42 2m2-w43 2rt4-w71 2rt1-w8r 2rt6-w35 2rs1-w84 2rt7-w33 2rs3-w59 1ncr-w11 2rs5-w56 2hwbWin56291 1qjyWin65099 1r09R61837 1qjyWin61209 1vthSDZ880-061 1hvhSDZ35-682 Win54221 Win56287 2hweWin54954 1qjyWin68934							
HCV Polymerase 1 inhibitors	2brl-poo 2brk-cmf 861966-42-9 861966-61-2 861965-95-9 861965-76-6 861965-88-0							
HCV Polymerase 2 inhibitors	1nhu-153 1nhv-154 1yz-jpc 1yx-ipc 1os5-nh1							
HCV Polymerase 3 inhibitors	1z4u-ph9 1ywf-ph7 639517-93-4 855301-46-1 860015-79-8 639518-06-2 PNU248809 855301-44-9							

Table 2 Examples for PS results for several inhibitors, descriptor calculation and activity profile prediction from the fast flexible search hit map

Inhibitor Structure	Green/red signals	Number of hit targets	D1	D2	D3	D4	Ratio D3 to D4	Profile prediction quality
 3	6/0	1	100	0	60	0	≥ 1	Good
 16	5/1	2	83	17	50	10	≥ 1	Good
 17	3/1	2	75	25	30	50	< 1	Uncertain
 18	7/3	3	30	70	30	60	< 1	Poor
 15	2/2	1	67	33	100	20	≥ 1	Good

than one, the activity profiling prediction will lead to a false target. The lower the ratio is, the lower the chances are that the correct locus of biological action can be identified.

This ratio was selected as the dominating parameter to determine the accuracy of prediction performance achieved in our screening approach and was therefore calculated for each compound in the hit map. First, the results obtained with the fast flexible search algorithm in Catalyst were inspected and gave well-convincing results: while a majority of inhibitors (89%) was predicted with correct activity profiles, in only 8% of the cases the ratio was lower than one and therefore prediction accuracy was doubtful. For three of the compounds no mapping hypothesis was found and therefore no clues concerning their biological activities were gained, which is a clue for a problem occurring in the conformational analysis procedure. For better visualization within the hit matrix, inhibitors in the first row are colour coded representing the outcome of this ratio calculation with light green highlighting characterizing $D3/D4 \geq 1$, red a ratio lower than one, and white indicating the lack of profile predictability.

In the first example, the HIV protease inhibitor **3** is identified by six pharmacophore queries. The signals in the hit map indicate 100% correct hit pharmacophores. Out of ten HIV protease pharmacophores in the system 60% are found. Activity profile prediction is correct, straight forward and easily interpretable. In the second example, the HIV protease inhibitor **16** is identified by six pharmacophore models. The signals in the hit map indicate 83 and 17% incorrect hit pharmacophores. Out of ten HIV protease pharmacophores in the system, 50% are found. The error signal represents one of ten models for RT resulting in a percentage of 10 for models found for one specific false target. Activity profile prediction is correct and easily interpretable because of the equalized situation between the two identified targets. In the third example, the HIV RT inhibitor **17** was identified by four pharmacophore models. Three signals in the hit map indicate 75 and 25% incorrect hit pharmacophores. Out of ten RT pharmacophores in the system 30% are found. The error signal represents one of two models for an HCV polymerase allosteric site resulting in a percentage of 50 for models found for one specific false target. Activity profile prediction is incorrect according to the D3/D4 ratio, however hard interpretable because of the not equalized number of models for the two identified targets. In example four, the influenza NA inhibitor **18** is identified by ten pharmacophore models. The signals in the hit matrix indicate 30 and 70% incorrect hit pharmacophores. Out of ten NA

pharmacophores in the system, 30% are found. The error signals represent two false targets: one comes from the HIV protease and six from the RT. In this equalized situation that makes RT the most extensively found specific false target with a percentage of 60. Activity profile prediction is incorrect according to the D3/D4 ratio, the outcome therefore clearly misleading to RT. In the fifth example, the HCV polymerase inhibitor **15** is identified by three pharmacophore models. The signals in the hit map indicate 67 and 33% incorrect hit pharmacophores. Both of the only two pharmacophores for this allosteric site in the system are found (100%). The error signal represents one of five models for another polymerase allosteric site resulting in a percentage of 20 for models found for one specific false site. Activity profile prediction is correct and easily interpretable because the two identified interaction sites both represent the same target.

In addition to the analysis of the ratio between D3 and D4 the result hit matrix was studied in order to investigate the activity profiles obtained for the screened ligands as well as the assessment of the performance of the pharmacophore models. This analysis revealed other important aspects influencing profile prediction that should therefore be taken into consideration:

Aspect 1. The first parameter that was studied was the influence of the search algorithm in the PS process. Therefore, we compared the outcome of *best* versus *fast flexible search* algorithms in Catalyst. Previous experiences from DB screening indicate that the *best flexible search* often retrieves a largely increased number of hits provoking the question if the results of the study would thereby become more unselective. In fact, in our PS case, the outcome was hardly altered: Still 88% of all inhibitors obtained a profile ratio higher or equal to one and were therefore predicted with correct activity profiles. For 12% of the compounds, misleading predictions were obtained. Although this seems to be a slight deterioration, the enrichment within the correct targets could be seen much more clearly, i.e. the intensity of correct signals was increased, which simplifies activity profiling vitally. In the course of quantifying this observation, an increase of D3 for *best* versus D3 for *fast flexible search* simultaneously with an unchanged value for D4 *best* compared to D4 *fast*, was defined to be an improvement in profile prediction for a compound when switching from the *fast* to the *best* screening modus. Conversely, $D3_{best}:D3_{fast} = 1$ and $D4_{best}:D4_{fast} > 1$ indicates deterioration. Looking at our total ligand set, 32 improvements face not even half as many (15) deteriorations. When inspecting only the 85 mol-

ecules with correct profiles (from *fast* and *best search*), 28 of them obtained better, eight of them poorer (nevertheless still correct) profile predictions with the *best* screening mode, whereas the remaining structures were not influenced. Incorrect or lacking profiles with *fast* or *best flexible search* have been observed for 15 compounds. When looking at the impact of switching from *fast* to *best* for these critical ligands, four profile prediction improvements and seven deteriorations can be seen. Furthermore, it can be seen that *best flexible search* returns activity profiles for all compounds under investigation. In this application case study, this fact can be interpreted as a way of fine tuning the approach without really compromising the overall selectivity. We therefore suggest *best flexible search* for screening approaches of DB subsets and for cases, where no or only very restricted activity profiles can be found for a compound, or where no distinct profile is achieved. *Best search* can always be helpful to affirm the outcome of *fast search*, however, this search mode might sometimes not be useful as starting procedure because of an extensive and difficult manageable data output.

Aspect 2. The quality/selectivity of the pharmacophore models in the hit list is a critical point to consider, when determining the activity profile of a compound and choosing the targets for biological testing. Retrieval of a compound with a highly selective model should obviously be attributed greater significance than retrieval with a less selective model. This means that on the one hand, information on model quality must be available and on the other hand, that preferentially selective models should serve as input for the system. Also an automatic integration of this parameter in profiling might be possible adding weight to targets represented by highly selective models in the pharmacophore hit list and reducing the significance of those identified by low-selectivity models. The impact of model quality on the obtained activity profile of a compound, however, varies depending on model quantity for a target. An unselective model solely representing a target will produce far more deteriorating results in PS than a single low-performing model amongst a set of selective models for a target. Therefore, rather the overall performance of a hypotheses pool is vital as could be seen in a test where the least selective model of the system, generated from an HIV RT-inhibitor complex, was left out in the hit map. For *fast flexible search* no alterations of the predicted profiles could be seen and the obtained ratio D3 : D4 being higher or lower than one remained the same for all molecules. *Best flexible search* revealed a profile prediction improvement for one NA inhibitor, deterioration for one RT inhibitor and alteration from

incorrect to not accomplishable profiling for one HRV coat protein targeting compound.

Aspect 3. An important aspect is the entire number of hypotheses by which a target is represented in the system, since it influences the significance of the calculated descriptor values. A high percentage of models from a specific target finding a particular ligand indicates activity at that target. If a large set of models exist for a target, this adds even more weight to that effect and offers additional security for activity profiling. For instance; 100% of 20 models finding a ligand weights more than 100% of two models. Since this effect is elusive and its actual impact hard to assess, it has not been included in the analysis process. In general, extensive retrieval of a compound for a target well represented in pharmacophore space can be interpreted as a firm indicator for biological activity there.

Aspect 4. When inspecting the pharmacophoric and later the pharmacological profile obtained for a compound, detailed investigation of the returned targets can prove very helpful. (1) It is critical to know, whether the targets are structurally or functionally related. Do they all belong to the same class of enzymes, for instance are they all proteases? A ligand widely recognized by models for protease targets can very reliably be forecasted a protease inhibitor. Thus, such a situation also raises the question of ligand selectivity, a task often difficult to handle with pharmacophore models. (2) In very advantageous cases the models lead to macromolecules involved in the treatment of the same diseases. The targets may for example belong to the same physiological pathway or cascade (like the renin-angiotensin-aldosterone system), or very rarely the models define diverse interaction sites at the same protein. The last example in Table 2 shows such a case: the HCV polymerase inhibitor **15** is predicted active at two different allosteric binding sites of this enzyme. This means that although the exact interaction site for this ligand cannot be foreseen so easily, the necessity for this is not that critical here, because testing against HCV and especially against the polymerase clearly seems advisable. (3) There is further the question of meaningfulness of a ligand profile with regard to drug development. Compounds for application in the anti-infective sector for example should be checked for the desired lack of interactions with mammalian enzymes. If the profile suggests that the compound might act as a promiscuous inhibitor or if the pharmacophore hit matrix contains many hits for models belonging to metabolizing proteins or antitargets, this might be a strong indication severe side

effects and probably failure in later phases of drug development.

Aspect 5. In order to discard the possibility that conformer generation mode and conformer number might significantly influence the study outcome, exemplarily five inhibitors underwent another conformational model generation. Thereby, conformer generation type specification in Catalyst was set to fast creating a maximum 250 and 100 conformers, respectively. Since these parameters can be expected to have an effect primarily for very flexible structures, the inhibitor with the highest number of rotatable bonds for each of the five proteins was selected. For compounds with equal flexibility higher molecular weight was consulted for prioritization. The maximum number of rotatable bonds seen amongst HIV protease inhibitors was 33, for RT 11, for NA 17, for HRV coat protein 15 and for HCV polymerase 14. These ten conformational models of the five compounds were searched with all 50 pharmacophore models using fast as well as best flexible search and the hit pharmacophores were compared with the prior results (*best conformer generation* with a maximum of 250 conformers). Although the hereby obtained pharmacophoric profiles differed in some cases from previous signals, the pharmacological profile prediction quality—the ratio between D3 and D4—was generally not altered indicating that the effects of conformer number and mode of generation are relatively independent in this regard. Similar results are also confirmed in a study by Kirchmair et al. [25, 26]. The only exception was a highly flexible HIV protease inhibitor whose former correct prediction changed to an incorrect profile for *fast conformer generation* of a maximum of 100 conformers when applying the *fast flexible search* algorithm. Except for cases where large DBs are inserted into the PS system or where calculation time is a limiting factor, preferentially high-quality conformational models should be used for the input molecules. Thereby, potentially important pharmacophoric profiles are not missed because of an insufficient ligand representation.

Experimental settings

The entire study was carried out with Version 1.0 of the LigandScout software using an Athlon 1800 PC running MS Windows 2000 and with Catalyst Version 4.11 on a PC with a Pentium IV processor/2.8 GHz running Linux Fedora Core. If not mentioned otherwise in the text, default parameter settings of the programs were used.

Conclusions

The present study reports the results of the first applications example for a pharmacophore-based extensive PS approach aimed at the prediction of activity profiles. We used antiviral compounds that were searched with a large set of chemical-feature based pharmacophore models derived from publicly available ligand-target complexes. Successful activity profiling was achieved in the majority of the cases independent from the search algorithm used. Descriptors for the analysis of output data are discussed as well as the influence of factors like pharmacophore quality or total number of models for a target for the interpretation of an obtained activity profile. From the results obtained so far, up-scaling and automation of this approach seems easily feasible and thus will provide a system for fast virtual PS of compounds against a variety of targets and prediction of potential biological activities, which will offer new possibilities in the early phase of drug development.

References

1. Langer T, Krovat E-M (2003) Curr Opin Drug Discov Dev 6:370
2. Krovat E-M, Steindl T, Langer T (2005) Curr Comput-Aided Drug Des 1:93
3. Güner OF, Clement O, Kurogi Y (2005) Curr Med Chem 11:2991
4. Güner OF (2005) IDrugs 8:567
5. Langer T, Wolber G (2004) Pure Appl Chem 76:991
6. Schuster D, Langer T (2005) J Chem Inf Model 45:431
7. Sanguinetti MC, Mitcheson JS (2005) Trends Pharmacol Sci 26:119
8. Clement OO, Guener OF (2004) In: Testa B (ed) Proceedings of the 3rd pharmacokinetic profiling in drug research: biological, physicochemical, and computational strategies. Verlag Helvetica Chimica Acta, Zurich, pp381
9. Klabunde T, Evers A (2005) ChemBioChem 6:876
10. Norinder U (2005) QSAR Environ Res 16:1
11. Oloff S, Zhang S, Sukumar N, Breneman C, Tropsha A (2006) J Chem Inf Model 46:844–851
12. Cleves AE, Jain AN (2006) J Med Chem 49:2921
13. LigandScout 1.0 is available from Inte:Ligand GmbH, Vienna, Austria (<http://www.inteligand.com/ligandscout>)
14. Wolber G, Langer T (2005) J Chem Inf Model 45:160
15. Schuster D, Laggner C, Paluszczak A, Hartmann RW, Langer T (2006) J Chem Inf Model 46:1301
16. Krovat E-M, Frühwirth KH, Langer T (2005) J Chem Inf Model 45:146
17. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P (2000) Nucleic Acids Res 28:235
18. Zuckerman AJ, et al (2000) Principles and practice of clinical virology, 4th edn. Wiley, Chinchester New York Weinheim Brisbane Singapore Toronto
19. Catalyst Version 4.11 available from Accelrys Inc, San Diego, CA, USA

20. Derwent World Drug Index, available in Catalyst data format from Accelrys Inc, San Diego, CA, USA
21. De Clercq E (2004) *J Clin Virol* 30:115
22. Ren J, Nichols C, Bird L, Chamberlain P, Weaver K, et al (2001) *J Mol Biol* 312:795
23. Alymova IV, Taylor G, Portner A (2005) *Curr Drug Targets Infect Disord* 5:401
24. Hadfield AT, Diana GD, Rossmann MG (1999) *Proc Natl Acad Sci USA* 96:14730
25. Kirchmair J, Laggner C, Wolber G, Langer T (2005) *J Chem Inf Model* 45:422
26. Kirchmair J, Wolber G, Laggner C, Langer T (2006) *J Chem Inf Model* 46:1848