LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters

Gerhard Wolber*,† and Thierry Langer‡

Inte:Ligand GmbH, Mariahilferstrasse 74B/11, A-1070 Vienna, Austria, and Computer Aided Molecular Design Group, Department of Pharmaceutical Chemistry, Institute of Pharmacy, University of Innsbruck, Innrain 52, A-6020 Innsbruck, Austria

Received April 3, 2004

From the historically grown archive of protein−ligand complexes in the Protein Data Bank small organic ligands are extracted and interpreted in terms of their chemical characteristics and features. Subsequently, pharmacophores representing ligand−receptor interaction are derived from each of these small molecules and its surrounding amino acids. Based on a defined set of only six types of chemical features and volume constraints, three-dimensional pharmacophore models are constructed, which are sufficiently selective to identify the described binding mode and are thus a useful tool for in-silico screening of large compound databases. The algorithms for ligand extraction and interpretation as well as the pharmacophore creation technique from the automatically interpreted data are presented and applied to a rhinovirus capsid complex as application example.

INTRODUCTION

Pharmacophore modeling has evolved to an important and successful method for drug discovery over the last few decades.[1] It has become essential for the description of molecules and complex biological systems such as proteins or nucleic acids. Besides others, the concept of describing pharmacon-drug interactions via pharmacophore models consisting of relevant chemical ligand features has become a well-accepted technique which can be considered as appropriate for the use in high-throughput virtual screening.[2,3]

The 3-D pharmacophore concept is based on specifically those kinds of interactions that have been observed in drug-receptor interaction: hydrogen bonding, charge transfer, electrostatic, and hydrophobic interactions. A 3-D pharmacophore (or pharmacophore model) is a set of interactions (chemical features or functionalities) aligned in three-dimensional space. This spatial arrangement of chemical features represents the essential interactions of small organic ligands with a macromolecular receptor. As a consequence, a pharmacophore model is limited to representing one single mode of action, i.e., representing the binding mode of ligands that bind to the same target.

There are two approaches for developing pharmacophore models: Either by analysis of the known X-ray or NMR structure of the receptor or starting from a set of ligands which are supposed to bind to the same area within the target. In the first case, relevant chemical features are intuitively derived from the known complex *(structure-based design)*, in the latter, the maximum common set of chemical features is searched for *(ligand-based design)*. The presented work concentrates on the creation of pharmacophore models from known three-dimensional complex data, thus on the structure-based design approach.

Structure-based drug design is normally tightly associated with docking, which in a first step flexibly aligns the ligand molecule into a rigid macromolecule environment and then estimates the tightness of the interaction by different scoring functions.

The docking methods used range from incremental construction approaches used in FlexX,[4−6] shape-based algorithms, as used by DOCK,[7] grid-based methods built upon the AMBER[8] force field as implemented in AUTODOCK[9] to genetic algorithms in GOLD[10,11] allowing partial protein flexibility. Systematic or random search techniques are used by Glide[12] and LigandFit.[13] A slightly different approach was presented in the computer program LUDI,[14] which searches for interaction centers in the protein and assembles potential new ligands by combining fragments from a three-dimensional structure library.

Docking takes all the information from a rigid protein environment and scores several possible interaction modes for different alignments. This approach has not been designed for screening large compound databases, as it is rather computing-intense. If used for such a task, however, massive parallelization is necessary, which implicates usage of multinode supercomputers or very large clusters of PCs. The screening of large molecular databases on a reasonable computing effort basis is a clear strength and common approach for both ligand-based and structure-based pharmacophores.[2] While ligand-based pharmacophores can be created automatically by overlapping different conformations of a defined ligand set binding to a single specific receptor, there is yet no fully automated method to create a pharmacophore from a known 3-D structure.

The aim of this work was to extract and interpret protein-bound small organic ligands from the Protein Data Bank (PDB)[15] concentrating on X-ray structure complexes, which

* Corresponding author phone: +43 6991 507 5000; fax: +43 1 8174955 1371; e-mail: wolber@inteligand.com.
† Inte:Ligand GmbH.
‡ University of Innsbruck.

10.1021/ci049885e CCC: $30.25
Published on Web 11/24/2004

have a sufficiently high resolution and quality. A PDB complex does not only contain experimental data but also a reasonable amount of interpretation done by the researcher carrying out the experiment. In most cases, structure determination efforts concentrate on the macromolecule, while in the present work ligand structures are to be elucidated, which, unfortunately, is to be done from incomplete data in the majority of the cases. Several projects have already concentrated on the problem of ligand perception: the probably best known approach is Hendlich's Relibase,[16] whose algorithm BALI[17] automatically assigns bond orders and hybridization states. The conversion results are publicly available via the World Wide Web, also providing comprehensive information on the complexes the ligands were extracted from. Unfortunately, the algorithms used are not described publicly in detail and include manual corrections of the results, so that they were not used in this work.

An article describing ligand interpretation algorithms in a more detailed way was published by R. Sayle, the author of the popular protein visualization program Rasmol.[18] This algorithm is in principle similar to Hendlich's BALI but additionally describes a collection of functional groups, which were substantially used in this work as described below.

The software presented in this paper will never tend to replace human intuition and ability of error recognition of a real chemist—it will, however, be able to recognize and correct common known errors and form a reasonable basis for further work in many cases. A close look at the PDB file for checking the plausibility of the macromolecule and the complexed ligand stays necessary to estimate applicability of the presented to historically grown PDB data. A possible verification tool for the protein is the program WHATCHECK[19] developed by Hooft and Vriend et al. It performs several useful rule-based plausibility checks on the macromolecule and therefore provides information on the quality of the whole PDB file.

The software developed in the course of the presented work will be able to elucidate pharmacophore information from known and yet unknown protein−ligand complexes without manual interaction or manual interpretation.

LIGAND PERCEPTION AND INTERPRETATION

Although the PDB contains a huge amount of valuable information, its data quality is questionable: The proposed software deals with data mining in protein complexes which have been continuously submitted for over 30 years, and the used file format was mainly created to describe proteins, never focusing on ligands or their detailed description.

To yield best results from data gathering, an interpretation algorithm needs to eliminate any possible way of data tampering caused by automated conversion. This is the reason interpretation had to be performed starting from the slightly out-dated original PDB file format, which was used to submit the biggest part of all ligands complexed in proteins.

Ligand perception and interpretation was performed in two steps: (i) the perception and correction of plausible molecular topology including ring perception and (ii) the interpretation and subsequent assignment of hybridization states and bond types from (often ambiguous) geometrical information.

Ligand Topology and Ring Perception. The first step in the interpretation of the ligands is the analysis of topological information contained in the molecular graph without taking note of the positions of the atoms. These calculations regularly can be performed at much lower computational costs than three-dimensional analysis.

All atom types and bond specifications were read as described in the PDB file specification.[20] With respect to the numerous deviations and errors in the PDB files, several corrections were implemented in order to retrieve plausible results: bonds are created only if their length exceeds 0.8 or falls below 2 Å, and atom names are guessed in ambiguous cases to find a plausible solution.[21]

However, only a part of the molecular graph information is contained in the file format: bond types and atom hybridization states are missing and will later be derived from the three-dimensional arrangement of connected atoms. Connectivity information, however, is already present and can be used to distinguish cyclic molecule parts from noncyclic ones. Especially for planar cycles, this separation is a prerequisite for geometry interpretation, because of their different geometry characteristics. Neighbor bonds to an sp^3 atom have a default tetrahedral bond angle of 109.5 degrees, which can be distinguished from an sp^2 atom with a default bond angle of 120 degrees, while a planar aromatic five-membered ring has a typical angle of 108 degrees, even though it contains sp^2 atoms.

Many algorithms dealing with ring perception have been reported.[22−26] As described in the summarizing review of Lynch et al. on ring perception,[27] there are many solutions for describing a ring set, as long as the description is consistent and reproducible. An enumeration of all possible smallest sets of smallest rings (SSSR) is suggested to meet this requirement.

In this work, reproducibility for complex ring combinations is not essential as long as all relevant ring atoms are covered. Therefore, an efficient algorithm for finding only one smallest set of smallest rings proposed by J. Figueras was implemented.[28] The reasons for this choice were the awareness of the relevant previous reports[25,26] in this article, a very concrete description of the presented algorithm and its high efficiency. It was shown that, for common cases, the breadth-first search (BFS) approach used by Figueras can be 2000 times faster than any common depth-first search (DFS) algorithm. The BFS implementation had to be adapted using reference lists instead of arrays in order to be applicable to ilib Java framework,[29] which were presented earlier and form the basis of LigandScout.

The result of the topological analysis as described above is a molecular graph containing information on atom connectivity and ring closures. At this point, no qualitative information on bond types or hybridization states apart from planar ring atoms is present. To retrieve this information with a reasonable confidence, the geometric arrangement of the atoms as well as semantic chemical patterns were investigated.

Interpretation I: Hybridization States Through Geometry Templates. Hydrogens are missing in PDB files in most cases and therefore cannot be used to determine the valence of bonds between heavy atoms. To assign correct bond values, atom hybridization states are to be determined from molecular geometry first.

The hybridization state of an atom is determined by the location of a central atom in relation to its direct neighbors. A common approach, which is also applied in the existing interpretation algorithms mentioned above, uses bond angles of at least two arbitrarily chosen neighbors in order to make a decision on the hybridization state of an atom. If a central atom and two arbitrary neighbors form an angle of 109.5 degree (plus/minus a defined tolerance), an sp^3 state is assigned, for 120 degrees sp^2, for 180 degrees (linear) sp. For the high-energy conformations that frequently occur in bound ligands, however, considering bond angles only often leads to misinterpretations due to ambiguous results: The typical bond angle of 109.5 degrees for sp^3 is too close to the typical 120 degrees for sp^2. To deal with this problem, a new model for estimating hybridization states was developed.

For each hybridization state there is a corresponding rigid geometrical body whose corner points have defined positions if at least three points are known. For an atom with at least two neighbors, this model defines absolute ideal positions for all neighbors relative to a central atom. Instead of looking at bond angles, for each atom a hypothetical geometrical body whose position is defined by two neighbors is created. The distance of all neighbors from their ideal position is subsequently summed up in order to form the absolute geometric deviation. The absolute geometric deviation d_a for n neighbors, a list of ideal points $I_{0...(n-1)}$, and a list of observed points $O_{0..(n-1)}$ can be expressed as follows:

$$d_a = \sum_{i=0}^{n} \sqrt{(I_i - O_i)^2} \tag{1}$$

The relative geometric deviation, which can be used to compare geometric deviations of different geometric bodies, can subsequently be defined as shown below.

$$d_r = \frac{d_a}{n} \tag{2}$$

$$d_a = \frac{\sum_{i=0}^{n} \sqrt{(I_i - O_i)^2}}{n} \tag{3}$$

The rigid geometric body is optimally aligned and can be used to calculate geometric deviations. The first two neighbors form the basis for aligning the geometric template, which may result in different results when more than two neighbors are present. For these cases, the two neighbors, which form the solution with the minimum absolute geometric deviation, are selected.

Figure 1 shows the projections for the tetrahedral geometry template in 3-D space, while Table 1 shows the calculated edge point position coordinates with respect to bond length r.

To align the overlapping p-orbitals of several neighboring sp^2 atoms, adjacent sp^2 planes are aligned in an additional single-rotation step.

Hybridization state recognition through geometry templates is superior to assigning bond angles for all cases where one atom has more than two neighbors. For the case of three neighboring atoms, not one single angle but the deviations of three dependent angles are regarded by simply comparing positions, which is especially important for interpreting error-prone data such as the ligand information in the PDB. For the case of two neighbors of a central atom, the method is practically equivalent to bond angle comparison bearing the advantage that bond length deviations are taken into account at no additional cost.

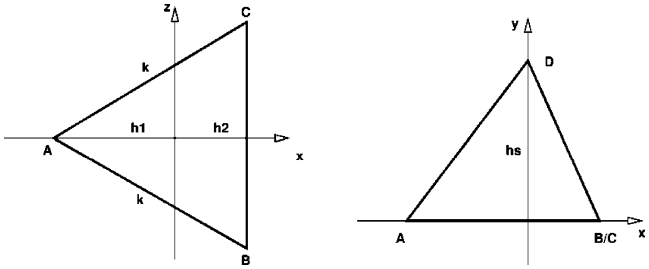


Figure 1. Tetrahedral geometry template projections.

Table 1. Edge Position Coordinates for the Tetrahedral Geometry Template with Respect to Bond Length r

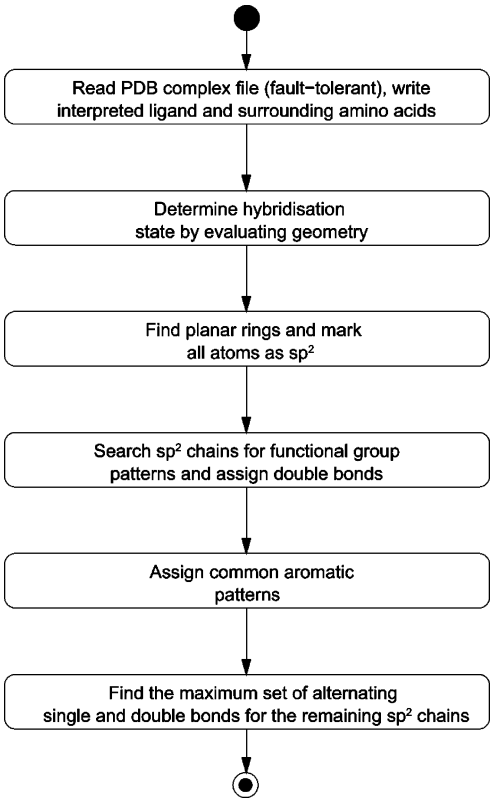
	Cartesian coordinates		
point label	X	Y	Z
A	$-r\sqrt{8}/3$	$-r/3$	0
B	$-r\sqrt{8}/6$	$-r/3$	$-\sqrt{6}r/3$
C	$-r\sqrt{8}/6$	$-r/3$	$-\sqrt{6}r/3$
D	0	r	0

Interpretation II: Common Functional Groups and Kekulé Patterns. Different to hybridization state assignment, the problem of double bond distribution to chains containing more than two sp^2 atoms cannot be solved by a similar straightforward approach like the assignment of geometry templates. Therefore, the next step after completion of hybridization state determination is the search for commonly occurring functional groups and bond patterns as it was proposed by R. Sayle.[21] Subsequently, commonly used aromatic Kekulé patterns are searched and assigned. Functional group patterns include the following: carboxylic acids and esters, thio acids and thioacetic esters, guanidine-like groups, acetamidine, acido groups, nitro groups, sulfonyl-groups, and phosphinoyl groups.

The user-extendable set of Kekulé patterns was implemented to include common cases of aromaticity. These currently include 1H-pyrrole and protoporphyrine. To implement the pattern recognition efficiently, the previously published *TopologyAnalyzer* algorithm[29] was modified to recognize subgraph isomorphism disregarding the bond type. The assignment of double bonds is directly done using the ligand/template compatibility graph. The Kekulé pattern set can be easily customized or reduced in order to favor or reject certain tautomers.

Interpretation III: Double Bonds in sp^2 Chains. Once functional group recognition and Kekulé pattern matching is accomplished, the remaining sp^2 chains are processed by a generic algorithm that finds the maximum number of remaining alternating double bonds for a molecule. From a given starting atom, it traverses a path of adjacent sp^2 atoms assigning double bonds if none of the conditions below is true: (i) A double bond adjacent to the processed atom already exists or (ii) a valence has to be exceeded. The solutions from this algorithms form a superset of all possible tautomers for the processed ligand. In the current imple-

Scheme 1. Steps Performed for Extraction and Interpretation of Ligands from the PDB



mentation the solutions with the maximum number of double bonds are selected per default. An overview of the workflow performed for each PDB complex containing ligands is shown in Scheme 1.

PHARMACOPHORE GENERATION AND CHEMICAL FEATURE PERCEPTION

The introduction of general chemical feature definitions such as charge transfers, lipophilic groups, or H-bond interactions has been an important step for describing the binding mode in a general way.[30] If abstract and general definitions are used, the resulting models become more universal at the cost of selectivity. Selectivity, nonetheless, is a major issue in pharmacophore validation, and therefore, in the common pharmacophore creation and validation process, too general feature descriptions are changed from reflecting universal chemical functionality to representing distinct functional groups. It has been a common approach to derive a model from distinct ligands in order to represent the specific mode of interaction as a chain of functional groups or exclusions thereof.[31]

By restricting general chemical feature definitions in the way described above, the number of standard well-known features (like H−bond donors, acceptors, ionizable groups, etc.) increase at the detriment of comparability. However, only comparable pharmacophores are universal and can represent a mode of action instead of a set of already existing ligands. Additionally, the automated processing of pharmacophores becomes much more transparent if chemical features stay comparable.

To describe the levels of universality and specifity of chemical features, a simple model was created to allow referral to these properties more easily. Table 2 shows a proposed classification of the abstraction layers of chemical feature constraints: A lower level corresponds to higher specifity and therefore lower universality. A location point defining the central position of the feature is included on every layer.

The software package Catalyst available from Accelrys Inc. provides several methods to derive pharmacophore models from a set of known biologically active molecules.[32] These pharmacophore models follow an approach by Green et al.[30] and include layer 3 and layer 4 features, which can be customized resulting in specific and thus incomparable layer 1 and 2 feature definitions.

A possible reason for creating features on the low universality levels one and two may be that the definitions of the higher levels are not sufficient to describe the features occurring in the training set.[31] In many cases, customization could have been avoided if the Catalyst software package[32] allowed improving the description of the chemical functionality (layer 3 and 4 information) in a universal way.

The pharmacophore creation tool created in this work should serve as a basis for the comparison of feature locations and properties. Therefore, a chemical feature set had to be created that is still universal but yet selective enough to reflect the specific ligand−receptor interaction for the entries in the Protein Data Bank.

Chemical Feature Definitions. The chemical feature definitions described in these sections are all categorized into hydrogen bond interactions, which were described as layer 3 features as well as charge interactions and lipophilic interactions, which represent level 4 features.

Hydrogen Bond Interactions and Analogues. Hydrogen bonding occurs when a covalently bound hydrogen with a positive partial charge interacts with another atom with a negative partial charge.[33] This typically happens when the partially positively charged hydrogen atom is positioned between partially negatively charged oxygen and nitrogen atoms but is also found in different situations as described later in this section. The investigation of hydrogen bonding in this work started with a model that only considered chemical functionality (as described in detail below) and the distance between the heavy atoms involved. The results were good and selective enough for high-quality models, but for entries in the PDB with a lower resolution, the distance had to be extended from 2.5 (H_BOND_MIN_DISTANCE) to 3.8 (H_BOND_MAX_DISTANCE) Å in order to include all plausible interactions. Nevertheless, for the purpose of maintaining a reasonable level of selectivity, additional geometric constraints were implemented based on the assumptions shown in Figure 2 for sp^2 donor atoms and Figure 3 for sp^3 donor atoms. The model assumes an ideal hydrogen bond angle of 180 degrees, and the hydrogen bond to be broken when the angle difference exceeds 34 degrees in both directions around the central position. This was derived from hydrogen bonding in water[34] and is reflected in the 146 degree angle shown in the figures.

In the case of an sp^2 atom being the donor atom, the position of the artificially added hydrogen atom can be used as a basis for the calculations since there is only one plausible possibility for determining the corresponding coordinates. If the angle formed by the two heavy atoms and the shared

Table 2. Abstraction Layers of Chemical Feature Constraints

layer	classification	universality	specifity
4	chemical functionality without geometric constraint, e.g. an H−bond acceptor without a projected point or a lipophilic group	+++	-
3	chemical functionality (H-bond acceptor, H-bond donor, positive ionizable, negative ionizable, hydrophobic) with geometric constraint, e.g. an H-bond acceptor vector including an acceptor point as well as a projected donor point; aromatic ring including a ring plane	++	+
2	molecular graph descriptor (atom, bond) without geometric constraint, e.g. a geometrically unconstrained phenol group	-	++
1	molecular graph descriptor (atom, bond) with geometric constraint, e.g. a phenol group facing a parallel benzenoid system within a distance of 2 to 4 Å	- -	+++

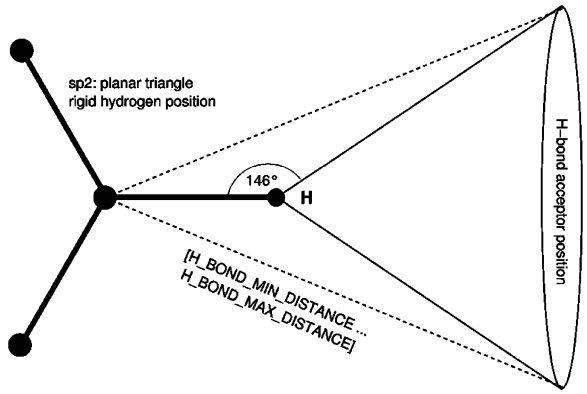


Figure 2. Geometric constraints for hydrogen bond interactions regarding sp^2 donor atoms.

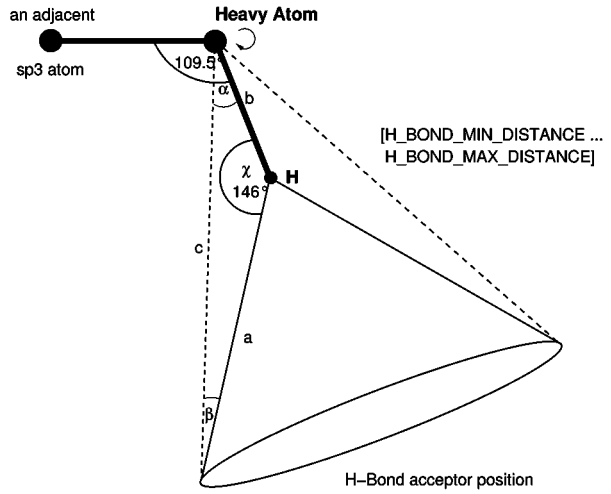


Figure 3. Geometric constraints for hydrogen bond interactions of sp^3 donor atoms.

hydrogen falls below 146 degrees, the H-bond is considered to be broken.

In the case of an sp^3 atom representing the donor atom the hydrogen may rotate freely, and the artificial hydrogen position cannot be taken into account. To formulate a constraint reflecting hydrogen bond plausibility, the first adjacent atom on the donor side was included in the consideration. As shown in Figure 2, the three donor atoms form the angle α, a is the distance from the (possibly hypothetical) hydrogen position to the acceptor heavy atom; the range of c must be in H_BOND_MIN_DISTANCE to H_BOND_MAX_DISTANCE, while b is the length of the bond between the heavy atom and the hydrogen. δ represents the tolerance angle that determines when the H−bond is considered to be broken. In the current default setup, δ is set to a value of 34 degrees.

$$\alpha = \delta - a \sin\left(\frac{b.\sin\chi}{c}\right) \tag{4}$$

If the angle formed between the adjacent atom and the two heavy atoms is within the range [109.5-α..109.5+α], the H-bond is considered to be plausible.

Hydrogen Bond Donors. Hydrogen bond donor atoms are recognized if they form part of one of the following functional groups: nonacidic hydroxyls (all OHs except sulfonic, sulfinic, carboxylic, phosphonic or phosphinic acids), thiols, acetylenic hydrogens, and NHs (except tetrazoles and trifluoromethyl sulfonamide hydrogens). If the heavy atom of a hydrogen bond acceptor within the surrounding macromolecule is found in the distance range of 2.5 to 3.8 Å of the marked atom and, additionally, the constraints described above are met, a hydrogen bond donor feature is created: It consists of a donor point on the ligand side and a projected point on the macromolecule side. A tolerance sphere of 1.5 Å is added to both points. The selected tolerance value is user-customizable; the default value of 1.5 Å, however, seems to be reasonable value regarding resolution and quality of the PDB structures.

Electrostatic Fluor-Hydrogen Bond Donor Interaction. The heavy atom on the acceptor side as described above can be replaced by a fluor atom. Although this kind of interaction is different from a chemical point of view, it is similar for the complementary group on the macromolecular side. Therefore, this feature was integrated into the group of hydrogen bond acceptors and analogues. The electrostatic interaction is represented as a vector that resembles the definition of the H-bond acceptor: The originating point is positioned on the fluor atom; the projected point is placed onto the heavy atom of the hydrogen bond donor on the macromolecule side.

If one of the criteria above is matched, the corresponding hydroxyl oxygen, thiol sulfur, acteylene carbon, or nitrogen is selected as the geometrically relevant heavy atom. The macromolecule is searched for a hydrogen bond acceptor atom within the distance range of 2.5 to 3.8 Å. If a suitable acceptor atom is found, a hydrogen bond donor feature is created with the donor point located on the position of the geometrically relevant atom on the ligand side and the projected point on the heavy atom of the acceptor point on the macromolecule side.

Hydrophobic Areas. In accordance with the concept of the Catalyst software package,[30,32] hydrophobic areas were

implemented in the form of spheres located in the center of hydrophobic atom chains, branches, or groups. First, a hydrophobicity scoring function pursuant to the Catalyst definition was implemented. As a next step, the algorithm checks if an ensemble of adjacent atoms is able to attain a sufficient overall hydrophobicity score. If this condition is met and a hydrophobic area in the macromolecule exists, a level 4 feature consisting of a sphere with a tolerance radius of 1.5 Å is added to the weighed center of these atoms. This implementation is conceptually compatible with the Catalyst default implementation.

A hydrophobic feature sphere is only added if a hydrophobic feature exists on the macromolecule side within a distance from 1 to 5 Å. The maximum distance was selected to be 5 Å because a larger gap would permit water molecules to be located in between.

Catalyst additionally checks the surface accessibility of the considered atoms. This test was omitted in order to increase efficiency, since the facing hydrophobic macromolecule area, which is the prerequisite for adding a hydrophobic features in this work, implicates surface accessibility. To reflect the steric constraints coupled with hydrophobic interactions, exclusion volume spheres are added at the coordinates of the lipophilic center on the macromolecule side.

Charge-Transfer Interactions. Positively ionizable areas (PI) are represented by atoms or groups of atoms that are likely to be protonated at a physiological pH. These are as follows: basic amines, basic secondary amidines, basic primary amidines, basic guanidines, and positive charges not adjacent to a negative charge.

Negatively ionizable areas (NI) are atoms or groups of atoms that are likely to be deprotonated at physiological pH, which matches trifluoromethyl sulfonamide hydrogens, sulfonic acids, phosphonic acids, sulfinic, carboxylic or phosphinic acids, tetrazoles, negative charges which are not neutralized by an adjacency to a positive charge.

If one of these groups is found, a sphere with a tolerance radius of 1.5 Å representing a negative or positive ionizable feature is added provided that the opposite ionizable feature can be located on the macromolecule side within the distance of 1.5 to 5.6 Å. The distance range is fully customizable; the high default tolerance value was selected taking into account the PDB data quality and resolution. Small metal-ions are extracted together with the protein environment of the ligand, which allows the handling of metal-ion directed interactions in a straightforward way by creating negative and positive ionizable features on the ligand if a charge interaction occurs within the specific distance range relevant for the respective ionizable features.

Pharmacophore Creation. A rule-set has been introduced that investigates the interpreted protein−ligand complexes and creates universal, comparable, but yet specific pharmacophore models representing the described mode of action without requiring manual intervention.

The chemical features used in the pharmacophore generation algorithm all represent chemical functionality but not molecular topology or specific functional groups, i.e., pharmacophores containing layer 3 and 4 features only. This enables the resulting pharmacophores to find potentially active ligands with completely new structures. The generated models are in the major points compatible with pharmacophore hypotheses created with default values software package Catalyst.[32] Therefore, Catalyst can be used to further investigate the models, which have been created using the presented algorithms in a fully automated way.

APPLICATION EXAMPLES

Application Setup. The following section illustrates the functionality of the algorithm presented above by using automatically interpreted information and applying it to a screening process which was performed using the database management functionality of the software package Catalyst.[32]

As already mentioned, Catalyst provides a mature and universal pharmacophore concept combined with the possibility to search compound databases with 3-D pharmacophore patterns. Most of the enhancements and modifications of the chemical feature recognition that were added in this work can be reproduced by customizing chemical feature definitions in Catalyst.

However, a major difference is that Catalyst does not allow simultaneous multiple chemical feature mappings on one atom or group, e.g. a carboxylic acid cannot be mapped to a negatively ionizable feature and a hydrogen bond donor feature on the hydroxyl moiety at the same time, which might be important to account for the selectivity of a model. For these cases, unlike Catalyst, the algorithm presented in this paper maps all relevant chemical features to the corresponding atoms, thus reflecting all relevant interactions for one single binding mode. The examples were chosen in a way that Catalyst's limits do not play a role in the screening process; for other cases, it might be necessary to either reduce the model in case specific knowledge on the target is available, to perform chained screening experiments using several models, or to use a different screening platform.

The chemical feature definitions available in Catalyst were redefined by exactly reproducing the definitions presented earlier in this paper. The feature dictionary used for the application examples contains five types of different chemical features: lipophilic points (LIP), positive ionizable points (PI), negative ionizable points (NI), hydrogen bond donor vectors (HBD), and finally, hydrogen bond acceptor vectors extended by electrostatic interactions occurring between fluor atoms and hydrogen donors (HBA-F). Additionally, exclusion volume spheres were used as steric constraints as described above. Catalyst hypotheses were generated by interfacing the hypoedit tool, which is part of Catalyst and allows the creation of pharmacophores from a predefined feature dictionary.

Three databases were created: The first, 'PDB Ligands singleconf', contains one entry for each ligand that was converted from a current version of the PDB in its single bioactive conformation. The main purpose of creating this single conformer 3-D database was to check if LigandScout pharmacophores are capable of finding those molecules in the conformations the specific pharmacophores were created from. This validation step is important, because it ensures that the two different technologies (Catalyst and LigandScout) work together and that Catalyst is capable of using a LigandScout pharmacophore to identify the compound which was part of the pharmacophore creation process. Ligand conformations are conserved from the PDB in this database, which provides direct access to the underlying PDB complex

through the compound name for each hit. Therefore, the database contains 67 265 entries, including topologically identical molecules contained as different conformers in potentially different complexes. The large number of entries is caused by the creation of an entry for each ligand of each chain (e.g. two entries for a dimer). Very small molecules, such as counterions and waters, were removed; cofactors and carbohydrates, however, were left in the database in order to analyze the quality of the pharmacophore model, which should be capable of not responding to this type of compounds.

The second database, 'PDB Ligands multiconf', contains the same ligands, but the single active conformation retrieved from the PDB was discarded and replaced by a set of conformers generated by Catalyst (conformer generation method FAST, maxConfs=100). Searching the unfiltered multiconformational database should show if the pharmacophore model is capable of finding the same molecule after it has undergone an artificial conformation generation process, which is normally the case in a standard virtual screening process. Duplicate molecules were removed in the multiconformation database and only 6680 identical ligands remained; the difference to the number of three-letter codes in the PDB can be explained by the fact that three-letter codes define fragments which may be covalently bound to each other in a PDB HET group.

To determine the number of 'drug-like' molecules in these databases, the ligands were processed to form a third database, "PDB Ligands multiconf drug-like". A molecular weight constraint of minimum 250 and maximum 600 in combination with the 'Lipinski rule of 5' (max. 10 acceptors, max. 5 donors, max. logP 5) was applied.[35] Due to the lack of experimental logP values, the topological cLogP estimation algorithm of Wildman and Crippen[36] was used to filter the 6680 unique compounds from the multiconformational database. Because this sort of calculation is only approximate, the cLogP range was extended to a maximum of 6. Although this rough kind of filtering may be considered problematic, regarding this database should give an idea about the number of 'drug-like' molecules in the PDB and the sort of enrichment that a pharmacophore is able to provide against a background of pharmaceutically relevant compounds. From the 6680 unfiltered PDB ligands with removed duplicates, 2765 conforming to the simple drug-likeness criteria remained.

Additionally, the Maybridge compound library (Version 2003 containing 59 194 compounds) was converted into multiconformational format using Catalyst and screened analyzing the resulting hit lists for their accordance to the Lipinski drug-likeness criteria as applied to the PDB ligand database. All hit lists were cut off at a molecular weight of 600 in order to omit very large compounds that only match a small fraction of the pharmacophore definition. The value of 600 was chosen for the two examples described below; it may need to be changed for pharmacophores targeting larger compounds.

Automated Overlaying of Pharmacophores. To find the relevant key features, an overlaying algorithm for several pharmacophore models for the presented examples was implemented: First, a compatibility graph of all feature point pairs was constructed, where vector base points and projected points were regarded as independent classes of features.

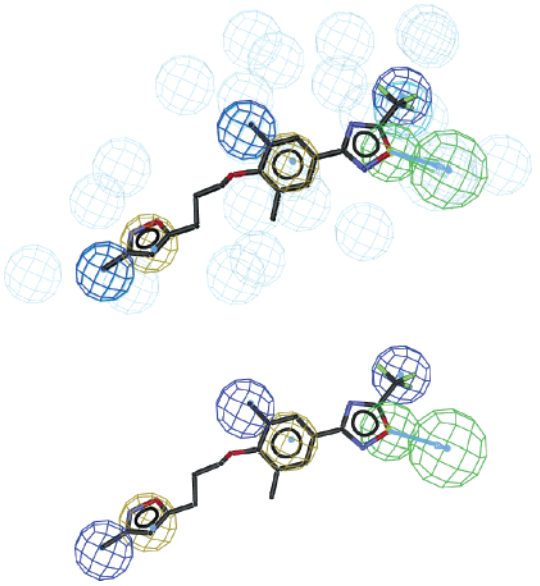


Figure 4. Merged pharmacophore from PDB records 1ncr, first3, and 1c8m with Pleconaril consisting of two lipophilic aromatic points (brown), three lipophilic points (blue), and one vectorized HBA-F (green); with and without excluded volume spheres.

Chemical feature pairs were defined to be compatible, if their distance is within the defined tolerance for both tolerance spheres. From the largest compatible subset of chemical feature compatibility pairs determined by maximum clique detection[29] two distinct common feature pharmacophore models were formed, which were subsequently aligned in 3-D space performing a single rotation using an analytical, efficient algorithm by Kabsch.[37] For the 3-D alignment algorithm, a weight of 1.0 was assigned to chemical features located on the ligand and a weight of 0.1 to excluded volume spheres; this was necessary to prioritize the alignment of chemical functionality rather than the alignment of exclusion spheres. From the resulting optimum 3-D alignment with the lowest RMS and the maximum number of chemical features a common feature model was derived by centering each feature point.

The two applications examples below were arbitrarily selected as application demonstration of the proposed algorithm. The discussion of a larger data set would exceed the scope of the presentation of the algorithm, which is the main focus of this paper and therefore will be shown elsewhere.

Example 1: Human Rhinovirus Serotype 16 Inhibitors. Human rhinoviruses (HRV) are members of the *Picornaviridae* family that are most frequently associated with viral infections causing symptoms of the common cold.[38] Pleconaril, a new capsid-binding agent, was recently published in the PDB by Viropharma.[39] Three PDB entries (1ncr, which is shown in Figure 4, 1nd3, and 1c8m) contain Pleconaril (PDB id: w11) bound to the protein hull of Rhinovirus subtype 16. From these three entries, three pharmacophores were automatically created (an example is shown in Figure 5), from which a common pharmacophore model was derived by applying the overlay algorithm described above. The final pharmacophore (see Table 3 and Figure 4) consists of 3 lipophilic points, 2 aromatic lipophilic points, 1 H-bond acceptor and 22 excluded volume spheres, which characterize the protein environment of the lipophilic points.

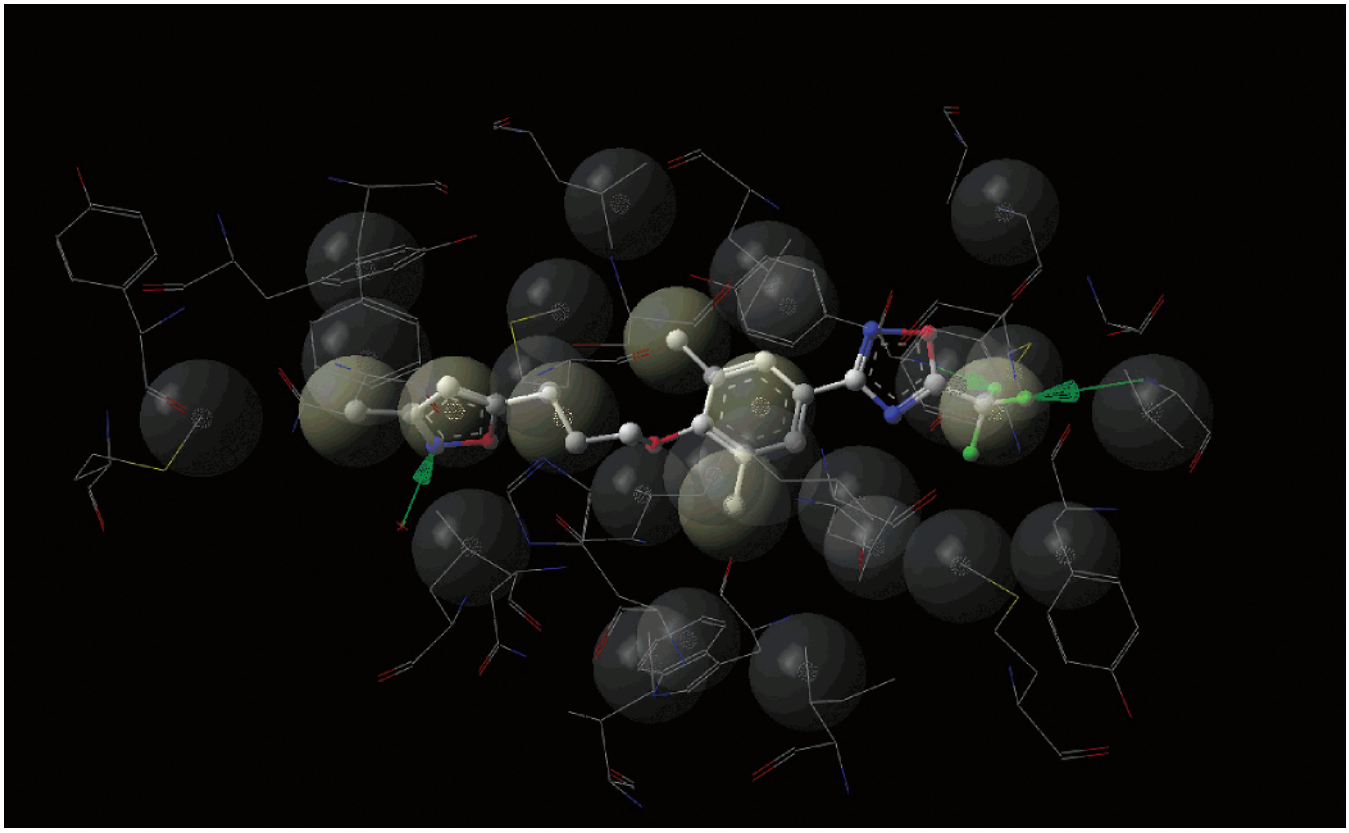


Figure 5. One of three source pharmacophores from PDB entry 1ncr aligned with Pleconaril and the rhinovirus protein hull.

Table 3. Chemical Feature Coordinates and Tolerance Radii (Å) Found by LigandScout in the Combined HRV Pharmacophore Model

		Cartesian coordinates		
feature type	radius	X	Y	Z
lipophilic	1.5	4.80	−4.85	−1.07
lipophilic	1.5	−8.72	5.50	0.99
lipophilic	1.5	0.71	2.08	0.62
HBA-F	1.5	4.84	−5.60	−0.36
(projected point)	1.5	2.96	−7.32	1.50
lipophilic aromatic	1.5	−6.62	4.04	0.41
lipophilic aromatic	1.5	5.04	−0.68	−0.03

Table 4. Virtual Screening Results for HRV Subtype 16

database	# hits	drug-like hits	Pleconaril hits
PDB singleConf	8	8	5
PDB multiConf	1	1	1
Maybridge	67	48	0

The results of the screening experiments are shown in Table 4: The search in the single- and the multiconformational database exactly yielded four distinct hits: Pleconaril (PDB id: w11), WIN61209 (PDB id: w01), WIN68934 (w02), and WIN65099 (w03), which are all reported to be Rhinovirus 16 hull protein binding agents.[40] Remarkably, the Pleconaril hits from the single-conformational database also included conformations that were extracted from HRV subtype 14 complexes but no other HRV subtype 14 binding agents, of which several exist in the PDB. This could be a hint that Pleconaril binds to subtype 14 in a similar conformation as to subtype 16, but there are different chemical features that stabilize this conformation. Searching the Maybridge database yielded 67 hits, of which 49 confirmed to the drug-likeness criteria described above.

Example 2: BCR-ABL Tyrosin Kinase Inhibitors. Chronic myelogenous leukemia (CML) is caused by a single genetic defect, which produces a fusion protein BCR-ABL, for which the Novartis group developed the potent inhibitor STI-517 (also known as CPG 57148B or Imatinib) with an IC_{50} of 25 nM for ABL and BCR-ABL and nearly no activity for other protein kinases, except the inhibition of two closely related Type III RPTKs, platelet-derived growth factor and c-Kit.[41−43] STI-571 has been approved in the United States for treatment of CML under the trade-name Gleevec. Remarkably, crystal structures of a close analogue of STI-571 revealed that STI-571 binds to the inactive form of ABL tyrosin kinase, stabilizes it, and thus prevents activation.[44] This binding mode was targeted in the second application example of LigandScout.

Three relevant PDB entries for this investigation were identified: 1fpu, 1iep, and 1opj. From these three entries, six pharmacophores were created from three complexes (all three records contained two different chains with a ligand each) and two different ligand molecules, STI-571 (PDB id: sti) and its variant N-[4-methyl-3-[[4-(3-pyridinyl)-2-pyridinyl]amino]phenyl]-3-pyridinecarboxamide (PDB id: prc) used in 1fpu. In a straightforward approach all six pharmacophore models were merged together into one single hypothesis using the clique detection algorithm together with Kabsch alignment described above.

The resulting pharmacophore model is shown in Table 5 and Figure 6 and contains four lipophilic aromatic areas, two acceptors, and eight excluded volume spheres. The same screening experiments as already carried out with the first example showed the results depicted in Table 6: The pharmacophore was able to identify all Gleevec entries from the singleconf PDB database and missed none. From the

Table 5. Chemical Feature Coordinates and Tolerance Radii (Å) Found by LigandScout in the Combined BCR-ABL Pharmacophore Model

feature type	radius	Cartesian coordinates		
		X	Y	Z
lipophilic aromatic	1.5	0.494	−6.948	5.402
lipophilic aromatic	1.5	2.543	−2.759	−0.716
lipophilic aromatic	1.5	−0.060	2.315	0.538
lipophilic aromatic	1.5	−3.027	4.833	−0.210
HBA-F	1.5	−0.535	−4.037	3.324
(projected point)	2.0	−0.027	−1.893	6.132
HBA-F	1.5	−4.318	4.220	−1.277
(projected point)	2.0	−6.600	3.136	−2.507

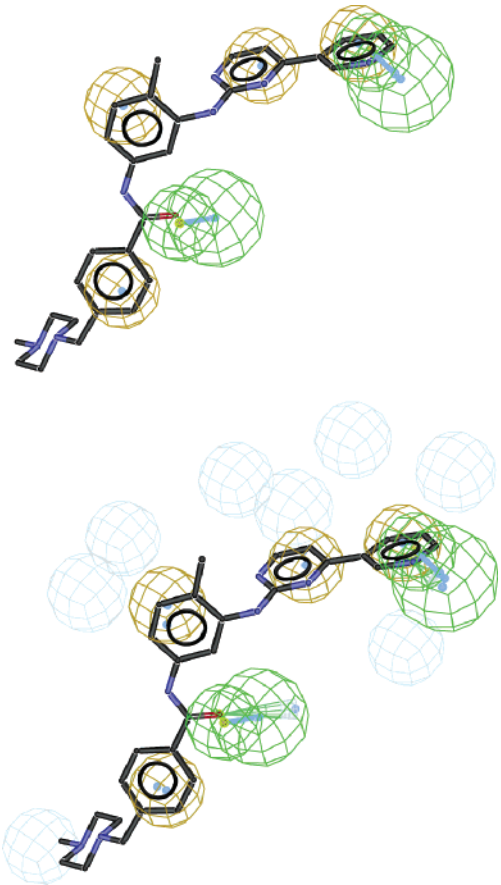


Figure 6. Merged pharmacophore from 1fpu, 1iep, and 1opj with STI-571 consisting of four lipophilic aromatic points (brown) and two vectorized HBA-F (green); with and without excluded volume spheres.

Table 6. Screening Results for the BCR-ABL Pharmacophore

database	hits	drug-like hits	STI-571 or PRC hits
PDB singleConf	7	7	7
PDB multiConf	2	2	2
Maybridge	19	7	0

Maybridge database, 19 compounds were identified out of which seven correspond to the simple drug-likeness criteria described above.

BENCHMARKS

The most computing-intense part of LigandScout is the extraction of the ligand including interpretation and identification of the relevant amino acids, which has to be performed only once for the whole PDB and can subsequently be updated incrementally. For this part, a distributed computing architecture was designed that is capable of splitting computation on a per PDB record basis. These calculations were carried out on an 11-node Linux cluster each equipped with a single Pentium 4 processor (2.8 GHz) and 1 GB RAM sharing a single standard IDE hard disk via NFS. Processing the whole PDB (15.2 GB of uncompressed data), saving and indexing ligands including the interacting amino acids in LigandScout's internal file format was performed in less than 6 h. Pharmacophore generation can be carried out interactively within a graphical user interface; a single pharmacophore creation typically takes a few seconds on standard PC hardware (same configuration as described above) depending on the size of the ligand and the complexity of the interaction patterns.

SUMMARY

This paper presents data mining for pharmacophores in the Protein Data Bank (PDB), which is the largest available public repository of biological relevant proteins complexed with small organic molecules. The major focus of this work has been put on the ligands with the aim of extracting relevant information on the respective binding mode. Due to poor data quality of the ligands in some complexes resulting from historic growth, existing algorithms were adopted and new strategies were developed in order to interpret ligand topology adequately. A step-by-step interpretation is performed on the PDB ligand entries: planar ring detection, assignment of functional group patterns, hybridization state determination, and last Kekulé pattern assignment.

The interpretation procedure has formed the basis for the next step, the fully automated creation of pharmacophore models, which is a state-of-the-art approach to generalizing biological interactions.[2,3] A rule-set was presented that automatically detects and classifies protein−ligand interactions into hydrogen bond interactions, charge transfers, and lipophilic regions. The entire set of interactions forms a pharmacophore model, which can be used for rapid virtual screening. The application examples show that the resulting pharmacophores can be sufficiently selective and represent effective filters. By omitting features due to detailed knowledge of the target or merging several selective pharmacophore models the user can transparently enhance and adapt models, which were generated in a fully automated way.

ACKNOWLEDGMENT

We thank Theodora Steindl, Eva Krovat, Oliver Funk, Christian Laggner, and Konstantin Poptodorov for helpful discussions.

REFERENCES AND NOTES

(1) Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960−1964.

(2) Langer, T.; Hoffmann, R.; Bachmair, F.; Begle, S. Chemical function based pharmacophore models as suitable filters for virtual screening. *J. Mol. Struct. (THEOCHEM)* **2000**, *503*, 59−72.

(3) Walters, W.; Stahl, M.; Murcko, M. Virtual Screening − an overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(4) SYBYL 6.6/FlexX, Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri, 63144, U.S.A.

(5) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(6) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. Multiple automatic base selection: Protein−ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 369−384.
(7) Oshiro, C.; Kuntz, I. Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 113−130.
(8) Weiner, S.; Kollman, P.; Case, D.; Singh, U.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.
(9) Morris, G. M.; Goodsell, D. S.; Halliday, R.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(10) Jones, G.; Willett, P.; Glen, R. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(11) Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(12) Schrödinger, Portland, OR97201.
(13) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldan, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289−307.
(14) Böhm, H.-J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593−606.
(15) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(16) Cambridge Crystallographic Data Centre, Relibase: A program for searching protein−ligand databases, http://relibase.ccdc.cam.ac.uk/, 2003.
(17) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 774−778.
(18) Sayle, R.; Milner-White, E. RASMOL: Biomolecular graphics for all. *TIBS* **1995**, *20*, 374−376.
(19) Hooft, R.; Vriend, G.; Sander, C.; Abola, E. Errors in protein structures. *Nature* **1996**, *381*, 272−272.
(20) PDB File Format Contents Guide, Version 2.2, http://www.rcsb.org/pdb/docs/format/pdbguide2.2/, 1996.
(21) Sayle, R. PDB: Cruft to Content: Perception of Molecular Connectivity from 3D Coordinates, http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html, 2001.
(22) Gasteiger, J.; Jochum, C. An Algorithm for the Perception of Synthetically Important Rings. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43−48.
(23) Wipke, W.; Dyott, T. Use of Ring Assemblies in Ring Perception Algorithm. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140−147.
(24) Hendrickson, J.; Grier, D.; Toczko, A. Condensed structure identification and ring perception. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 195−203.
(25) Balducci, R.; Pearlman, R. Efficient exact solution of the ring perception problem. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 822−831.
(26) Fan, B.; Panaye, A.; Doucet, J.; Barbu, A. Ring perception. A new algorithm for directly finding the smallest set of smallest rings from a connection table. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 657−662.
(27) Downs, G.; Gillet, V.; Holliday, J.; Lynch, M. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172−187.
(28) Figueras, J. Ring Perception using Breadth-First Search. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986−991.
(29) Wolber, G.; Langer, T. CombGen: A novel software package for the rapid generation of virtual combinatorial libraries. In *Rational Approaches to drug design*; Höltje, H.-D., Sippl, W., Eds.; Prous Science: 2000; pp 390−399.
(30) Green, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.
(31) Krovat, E.; Langer, T. Non-Peptide Angiotensin II Receptor Antagonists: Chemical Feature Based Pharmacophore Identification. *J. Med. Chem.* **2003**, *46*, 716−726.
(32) Catalyst, Version 4.7, Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752, U.S.A., 2001.
(33) Pauling, L. *The Nature of the Chemical Bond*, 2nd ed; Cornell University Press: New York, 1948.
(34) Khan, A. A liquid water model: Density variation from supercooled to superheated states, prediction of H-bonds and temperature limits. *J. Phys. Chem.* **2000**, *104*, 11268−11274.
(35) Lipinski, A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(36) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868−873.
(37) Kabsch, W. A solution for the best rotation to relate to sets of vectors. *Acta Crystallogr.* **1976**, *A32*, 922−923.
(38) Wimmer, E. Genome-linked proteins of viruses. *Cell* **1982**, *28*, 199−201.
(39) Viropharma, 405 Eagleview Boulevard, Exton, PA 19341.
(40) Hadfield, A. T.; Diana, G. T.; Rossmann, M. G. Analysis of three structurally related antiviral compounds in complex with human rhinovirus 16. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14730.
(41) Carroll, M.; Ohno-Jones, S.; Tamura, S.; Buchdunger, E.; Zimmermann, J.; Lydon, N. B.; Gilliland, D. G.; Druker, B. J. CGP 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing BCR-ABL, TEL-ABL, and TEL-PDGFR fusion proteins. *Blood* **1997**, *90*, 4947.
(42) Zimmermann, J.; Buchdunger, E.; Mett, H.; Meyer, T.; Lydon, N. B. Potent and selective inhibitors of the ABL-kinase: phenylaminopyrimidine (PAP) derivatives. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 187.
(43) Buchdunger, E.; Cioffi, C. L.; Law, N.; Stover, D.; Ohno-Jones, S.; Druker, B. J.; Lydon, N. B. Abl protein-tyrosine kinase inhibitor STI571 inhibits in vitro signal transduction mediated by c-kit and platelet-derived growth factor receptors. *J. Pharmacol. Exp. Ther.* **2000**, *295*(1), 139−145.
(44) Schindler, T.; Bornmann, W.; Pellicena, P. W.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Structural mechanism for STI-571 inhibition of Abelson Tyrosine Kinase. *Science* **2000**, *289*, 1938−1942.

CI049885E