Pure Appl. Chem., Vol. 76, No. 5, pp. 991–996, 2004.

Virtual combinatorial chemistry and in silico screening: Efficient tools for lead structure discovery?*

Thierry Langer[1,‡] and Gerhard Wolber[2]

[1]Institute of Pharmacy, University of Innsbruck Innrain 52, A-6020 Innsbruck, Austria; [2]Inte:Ligand GmbH Clemens-Maria Hofbauer-G. 6, A-2344 Maria Enzersdorf, Austria

Abstract: In this article, an overview of the most common ligand-based in silico screening techniques is given together with an example on the recent successful application of combined use of pharmacophore modeling, database mining, and biological assays. Additionally, a new approach for structure-based high-throughput pharmacophore model generation is presented. The LigandScout program contains an automated method for creating pharmacophore models from experimentally determined structure data, e.g., publicly available from the Brookhaven Protein Databank (PDB). In a first step, known algorithms were implemented and improved to extract small-molecule ligands from the PDB including assignment of hybridization states and bond orders. Second, from the interactions of the interpreted ligands with relevant surrounding amino acids, pharmacophore models reflecting functional interactions like H-bonds or ionic transfer interactions were created. These models can be used for screening molecular databases for similar modes of actions on the one hand, or for screening one single compound for potential side-effects (reversed screening) on the other hand. The implementation was done using the ilib framework, which also formed the basis of the software tool Comb^{i}Gen, a fragment-based virtual combinatorial library generation program enabling the user to obtain in silico compound collections with high drug-likeness.

INTRODUCTION

The increasing economic pressure on the pharmaceutical industry to develop new drugs in a faster and more efficient way than in the past has led to the development of a large number of new methods aimed at a more efficient and rapid lead structure discovery process. Recent advances in combinatorial chemistry have made it possible for chemists to synthesize large libraries of compounds, and today, high-throughput screening (HTS) allows considerable reduction of the time amount needed for the discovery of new molecules possessing biological activity for a certain target. However, the number of compounds that can be synthesized is still a small percentage of the total number of compounds that are possible in principle. The experimental efforts to carry out the biological screening of billions of compounds are still considerably high, and, therefore, computer-aided drug design approaches have emerged as a promising tool for helping medicinal chemists to decide what to synthesize [1]. In this context, one of the major goals of computer-aided ligand discovery strategies is the identification of small subsystems from large groups of chemical compounds. Combinatorial libraries can contain several 1000–100 000 compounds as already demonstrated [2], and, furthermore, libraries with a size

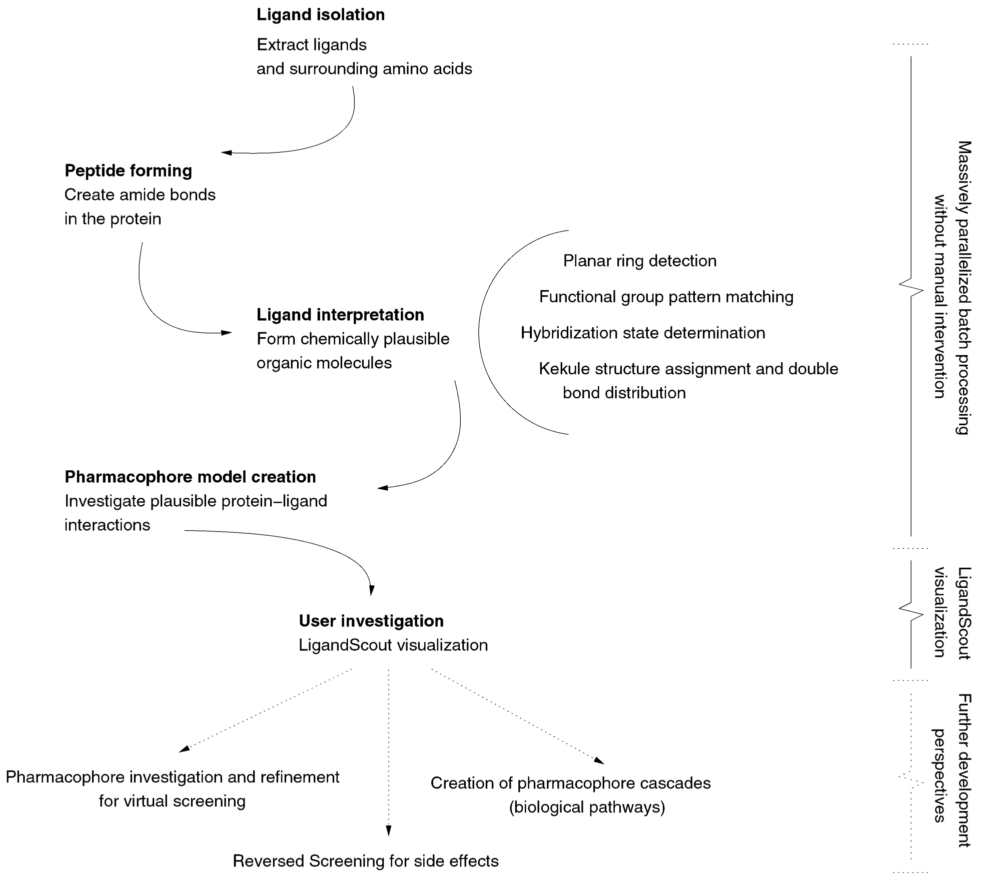
[*] Lecture presented at the Polish–Austrian–German–Hungarian–Italian Joint Meeting on Medicinal Chemistry, Kraków, Poland, 15–18 October 2003. Other presentations are published in this issue, pp. 907–1032.

[‡] Corresponding author

of 10^9 or more molecules can be assembled. Up to this, the existing virtual chemistry space may contain perhaps 10^{60} possible molecules. The filtering of large databases or libraries of candidate compounds through the use of computational approaches based on discrimination functions that permit the selection of series of compounds to be tested for biological activity has been termed “virtual screening” (VS). There are problems encountered with the generation and screening of very large virtual libraries, however, within the next years all the necessary components for processing virtual libraries with as many as 10^{15} compounds will be in place. Now, by using a state-of-the-art virtual screening strategy, one should already be able to reduce the number of candidates to be examined experimentally by at least 9 orders of magnitude, thus ending up with some 1000 of compounds to be assayed for their biological activity. Moreover, the main goal of medicinal chemists is to develop compounds that do not fail in later research phases. Therefore, the prediction of side-effects and ADMET properties is crucial. Side-effects, however, are often linked to low target specificity. We propose a new approach, parallel screening, for assessing bioaffinity profiles of virtual compounds for multiple targets in silico. For this task, we have developed a novel software tool for high-throughput, structure-based pharmacophore model generation: The LigandScout program is capable of analyzing and collecting pharmacophoric patterns from experimentally determined, biological macromolecules and their co-crystallized ligands in a fully automated high-throughput batch process.

LIGANDSCOUT PROGRAM

The LigandScout program [3] performs data mining in the Brookhaven Protein Data Bank (PDB), which is the largest available public repository of biological relevant proteins complexed with small organic molecules [4]. The major focus of this application has been put on the ligands with the aim of extracting relevant information on the binding mode for each particular ligand. Due to historic growth, the data quality for the ligands in most complexes available is poor. In order to obtain plausible results, existing algorithms were adopted and new algorithms were developed. The algorithms developed perform a step-by-step interpretation of the ligand molecules: Planar ring detection, assignment of functional group patterns, hybridization state determination, and, lastly, Kekulé pattern assignment. Due to errors in some PDB files and distorted or wrongly typed structures, this assignment still produces errors in some cases. However, due to the quantitative geometry model introduced, it is possible to automatically detect complexes showing a high error probability. In order to speed up the interpretation process, a distributed computing environment was designed and implemented that permits rapid processing of the whole Protein Data Bank: From 20 622 PDB entries, 42 471 ligand conformations out of 3.387 different ligands were extracted in less than 8 h on 5 standard PCs running Linux. 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. A rule-set was presented that automatically detects and classifies protein–ligand interactions into hydrogen bond interactions, charge transfers, and lipophilic regions. The set of all interactions is collected in order to form a pharmacophore model, which can be used for rapid virtual screening. Finally, a high-quality graphical user interface was created for the purpose of manually investigating and viewing the ligand complex simultaneously with the pharmacophore. The overall workflow is depicted in Scheme 1.


Ligand isolation
Extract ligands
and surrounding amino acids
Peptide forming
Create amide bonds
in the protein
Ligand interpretation
Form chemically plausible
organic molecules
Planar ring detection
Functional group pattern matching
Hybridization state determination
Kekule structure assignment and double bond distribution
Pharmacophore model creation
Investigate plausible protein–ligand interactions
User investigation
LigandScout visualization
Pharmacophore investigation and refinement for virtual screening
Reversed Screening for side effects
Creation of pharmacophore cascades (biological pathways)
Massively parallelized batch processing without manual intervention
LigandScout visualization
Further development perspectives

Scheme 1 LigandScout overall workflow overview.

IN SILICO SCREENING

Experimental screening for lead structure determination suffers from limitation with respect to the possible number of compounds that can be submitted to a high-throughput bioassay even if there has been much improvement within the last decade. However, this is also true for in silico VS, and therefore it is highly important to reduce the search space a priori. Obviously, the concept of maximal diverse libraries is no longer valid: Sensible and focused, synthesizable, and lead-like libraries are needed. Moreover, if VS should fulfill the time demands given by the pharmaceutical industry, there is a need for high-speed algorithms using highly approximate and intelligent filtering methods. Approaches used for VS range from 2D similarity analysis to evolutionary methods. In the 2D approach, a variety of descriptors is calculated from a molecular graph, which can be done very rapidly. Going a step further, descriptors derived from 3D structures can also be used for VS. The pharmacophore concept has been widely used over the past decades for rational drug design approaches and can now be incorporated into a VS strategy. A pharmacophore (pharmacophore model, pharmacophoric pattern) can be considered as the ensemble of steric and electrostatic features of different compounds that are necessary to ensure optimal supramolecular interactions with a specific biological target structure and to trigger or to block its biological response [5]. Following this definition, a pharmacophore is not a real molecule or a real association of functional groups, but a pure abstract concept which, however, accounts for the common molecular interaction capacities of a group of molecules toward their target structure. For implementation of this concept into VS, the chemical function-based approach is the most generic one. The originality of this type of pharmacophore mostly resides in the fact that their definition is general and represents the different types of interactions between small organic molecules and proteins. The utility of such models as queries for 3D database search has been recently reviewed [6]. Such pharmacophores can be generated indifferently from ligand sets or from an active site structure. At the end of VS filtering, a reliable method for ranking the hits obtained according to their expected bioactivity is required. Successful application of VS can be considered as the identification of a set of ca. 10^3 compounds bioactive compounds from a large-scale library containing up to 10^{12} different molecules. However, such a strategy could not be envisaged without the existence of well-designed compound libraries (virtual libraries).

FEATURE-BASED PHARMACOPHORE MODELS

When the 3D structure of the protein target has not been characterized, or when a certain number of ligands (with or without associated binding affinity) are available, pharmacophore models can be generated and used as search queries to screen a library. This approach can even be undertaken if only information on hits from HTS is available. Feature-based pharmacophores have turned out to be the most effective type of pharmacophore models in which pharmacophoric points represent chemical features like hydrogen-bond acceptors/donors, hydrophobic points, acidic or basic features, etc. Several automated methods can be used to generate this type of pharmacophore [7]. The strength of this type of pharmacophore models is the general definition of the pharmacophoric points. The chemical function character allows searching for very diverse structural scaffolds since multiple structural elements can express the same chemical function. In the past decade, an increasing influence on rational drug design has been exerted by different software programs that rely on chemical feature-based pharmacophore models. Several successful applications within this subject have been performed using the Catalyst program [8], one of the leading software packages in chemical feature-based pharmacophore modeling. A review covering such feature-based methods has been published recently by Kurogi and Güner [9] and by one of us and Krovat [10]. They outline the theoretical background as well as describe several significant studies including 3D database search strategies.

APPLICATION EXAMPLE: THE SEARCH FOR ENTOTHELIN ET_A RECEPTOR ANTAGONISTS

The endothelins (ETs) are 21 amino acid peptides which are mainly released by vascular endothelial cells. There are three known subtypes, namely ET-1, ET-2, and ET-3, binding to at least two types of G-protein coupled receptors, ET_A and ET_B. ET-1 is the most potent endogenous vasoconstrictor known so far. Elevated levels of ET-1 and of cardiac ET_A receptors and down-regulation of ET_B receptors were verified in disease states like essential and pulmonary hypertension, congestive heart failure, and arteriosclerosis, which are associated with excessive vasoconstriction and smooth muscle cell proliferation. Therefore, antagonism of the ET_A receptors is expected to be an effective way of treating these diseases.

In order to gain more insight into structure–activity relationships and to obtain access to new potential lead candidates by using in silico screening techniques, we generated chemical feature-based pharmacophore models of ET_A-selective antagonists [11] by using Accelry's Catalyst software [8]. The models were finally used as queries for 3D database mining, and biological testing was performed on selected molecules in order to validate the results obtained: Four of the retrieved compounds originating from the commercial vendor Maybridge, Ltd. were selected for testing of their affinity to the ET_A receptor. Two of these proved to be active at the ET_A receptor (low µM range). Summarizing, it could be demonstrated that chemical feature-based pharmacophore models are effective tools for the in silico identification of new potential lead structures [11].

VIRTUAL LIBRARY DESIGN AND GENERATION

Compounds to be screened in silico are not limited to those that already exist. A virtual library can be generated using a computational approach. The criteria for generating a valuable general virtual library are (i) large diversity; (ii) high degree of lead-likeness; (iii) low chance of affecting targets responsible for side-effects; (iv) favorable ADMET properties; and, last but not least, (v) synthetic accessibility. There are a lot of possibilities to generate a diverse virtual library. However, it is challenging to construct a virtual library that meets the criteria set forth above. If the objective is to generate a focused library, the same requirements have to be met. Diversity in this case, however, reflects scaffold diversity of compounds all matching the same pharmacophore space. Diversity estimation is still a field of large interest, and a considerable number of different new algorithms have been developed and used for assessing the structural diversity of libraries. Clearly, multidimensional optimization has to be performed for obtaining valuable virtual compound libraries. In our opinion, genetic algorithms combined with optimizing heuristics are the most appropriate and flexible approach to achieve such a goal. We will incorporate them into our library generation tools, the first one being a new software package *iLib diverse* building drug-like compound libraries starting from a well-balanced molecular fragment set [12]. We were able to generate structurally diverse compound databases containing more than 90 % of drug-like molecules as assessed by Sadowski and Kubinyi using their neural network described in ref. [13].

It is widely accepted that structural diversity alone does not represent the key to success for a compound library. Differences between drugs, natural products, and molecules from combinatorial chemistry within the chemical property space have been analyzed, and the results of this study suggest that by mimicking certain distribution properties of natural compounds (e.g., the prevalence of aromatic rings, the number of complex ring systems, the degree of saturation), combinatorial libraries might be designed that are substantially more diverse and have greater biological relevance, which is still an issue in this area [14]. Natural products present a large number of appealing scaffolds and, moreover, exhibit in most cases some kind of biological activity. Therefore, pharmacophoric properties and scaffold architecture of natural products and trade drugs, which share several topological pharmacophore patterns, have been analyzed [15], and applications have been elaborated based on self-organizing maps for the design of natural product-based combinatorial libraries. A chemically diverse virtual library can contain many non-drug-like compounds, and applications have been developed to recognize drug-like compounds from diverse compound libraries. The problem has partly been solved; however, it has been observed that many drug-like compounds, which should be potential candidates, do not appear as hits when they are screened against biological targets. Therefore, it is now common sense that further refinement of the filtering technologies should be made in order to retrieve lead-like compounds instead of drug-like compounds. Intrinsically, lead-likeness and drug-likeness are the descriptors of potency and selectivity, but also absorption, distribution, metabolism, toxicity, and scalability. Until now, these parameters were optimized sequentially, but nowadays it is believed that these parameters should be optimized simultaneously.

REFERENCES

1. H. Kubinyi. *J. Recept. Signal Transduct. Res.* **19**, 15 (1999).
2. D. Maclean, J. R. Schullek, M. M. Murphy, Z. J. Ni, E. M. Gordon, M. A. Gallop. *Proc. Natl. Acad. Sci. USA* **94**, 2805 (1997).
3. G. Wolber. Ph.D. thesis, University of Innsbruck (2003).
4. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. *Nucl. Acids Res.* **28**, 235 (2000).
5. C.-G. Wermuth and T. Langer. In *3D-QSAR in Drug Design: Theory, Methods and Applications*, H. Kubinyi (Ed.), pp. 117–136, Escom, Leiden (1993).
6. T. Langer and R. D. Hoffmann. *Curr. Pharm. Des.* **7**, 509 (2001).

7. R. D. Hoffmann, H. Li, T. Langer. In *Pharmacophore Perception, Development and Use in Drug Design*, O. F. Güner (Ed.), pp. 301–318, International University Line, La Jolla, CA (2000).
8. This software package is available from Accelrys Inc, San Diego, CA, USA.
9. Y. Kurogi and O. F. Güner. *Curr. Med. Chem.* **8**, 1035 (2001).
10. T. Langer and E. M. Krovat. *Curr. Opin. Drug Discov. Dev*. **6**, 370 (2003).
11. O. Funk, V. Kettmann, J. Drimal, T. Langer. *J. Med. Chem.* **47**, 2750 (2004).
12. G. Wolber and T. Langer. In *Rational Approaches to Drug Design*, H.-D. Höltje and W. Sippl (Eds.), pp. 390–399, Prous Science, Barcelona (2001).
13. J. Sadowski and H. Kubinyi. *J. Med. Chem.* **41**, 3325 (1998).
14. M. Feher and J. M. Schmid. *J. Chem. Inf. Comput. Sci*. **43**, 218 (2003).
15. M. L. Lee and G. Schneider. *J. Comb. Chem*. **3**, 284 (2001).