The holistic integration of virtual screening in drug discovery

Yusuf Tanrikulu[1], Björn Krüger[1] and Ewgenij Proschak[2]

[1] Merz Pharmaceuticals GmbH, Chemical R&D - Drug Design, Eckenheimer Landstrasse 100, D-60318 Frankfurt, Germany
[2] University of Frankfurt, Institute of Pharmaceutical Chemistry, Max-von-Laue-Strasse 9, D-60438 Frankfurt, Germany

During the past decade, virtual screening (VS) has come of age. In this review, we document the evolution and maturation of VS from a rather exotic, stand-alone method toward a versatile hit and lead identification technology. VS campaigns have become fully integrated into drug discovery campaigns, evenly matched and complementary to high-throughput screening (HTS) methods. Here, we propose a novel classification of VS applications to help to monitor the advances in VS and to support future improvement of computational hit and lead identification methods. Several relevant VS studies from recent publications, in both academic and industrial settings, were selected to demonstrate the progress in this area. Furthermore, we identify challenges that lie ahead for the development of integrated VS campaigns.

Introduction

The identification of novel lead structures is a central task at the beginning of a drug discovery campaign. There are many ways to identify hits, which can then be used as starting points for hit-to-lead optimization. The systematic experimental testing of large compound libraries (i.e. HTS) has been established since the 1980s. The costs of HTS experiments are tremendous and, thus, VS, an *in silico* analog of HTS, was developed ten years later. Comparison of the appearance of literature related to VS and HTS highlights this development (Fig. 1).

Notably, the most cited HTS-related publication, that by Lipinski *et al.* [1], discusses the application of both HTS and VS to estimate the solubility and permeability of chemical compounds. Although VS was initially seen as a cost-saving substitute for HTS, both techniques are of a more complementary nature and recent developments in the area of lead identification approaches make use of the advantages of both. In this article, we do not intended to review the exhaustive applications of VS; instead, we present and analyze the evolution of VS over the past two decades. VS is currently maturing as a hit identification strategy, as occurred with HTS a decade before. This process becomes more evident as we observe the development of VS from a more isolated procedure toward a fully integrated technique for hit and lead identification [2]. Experimental data are no longer only collected after a VS campaign but are instead incorporated into the process.

Ten years ago, a trend toward the integration of VS and HTS had been documented [3], for which a classification has been proposed recently [4]. Here, we emphasize current progress in VS from selected recent publications and give an overview of the emerged integral strategies in drug discovery. We suggest a categorization of the global VS technique according to its level of integration into: classic VS, parallel VS, iterative VS and integrated VS (Fig. 2). We provide a definition of each category and focus on the benefits and bottlenecks of each.

Classical applications of virtual screening

VS is often compared to a funnel, where a large number of molecular compounds, often referred to as a VS library, is reduced by a computational algorithm to a smaller number that will then be tested experimentally (Fig. 2a). The screening library often contains 10^5–10^7 molecules, whereas the desirable output of these protocols is in the range of 10^0 to 10^3, depending on the study. The role of VS algorithms is to enrich active compounds in the highly reduced output. The protocol often comprises several 'filtering layers', which hold back inactive or undesired molecules or prioritize compounds according to their predicted activity (so-called

Corresponding author:. Proschak, E. (proschak@pharmchem.uni-frankfurt.de)

 1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2013.01.007

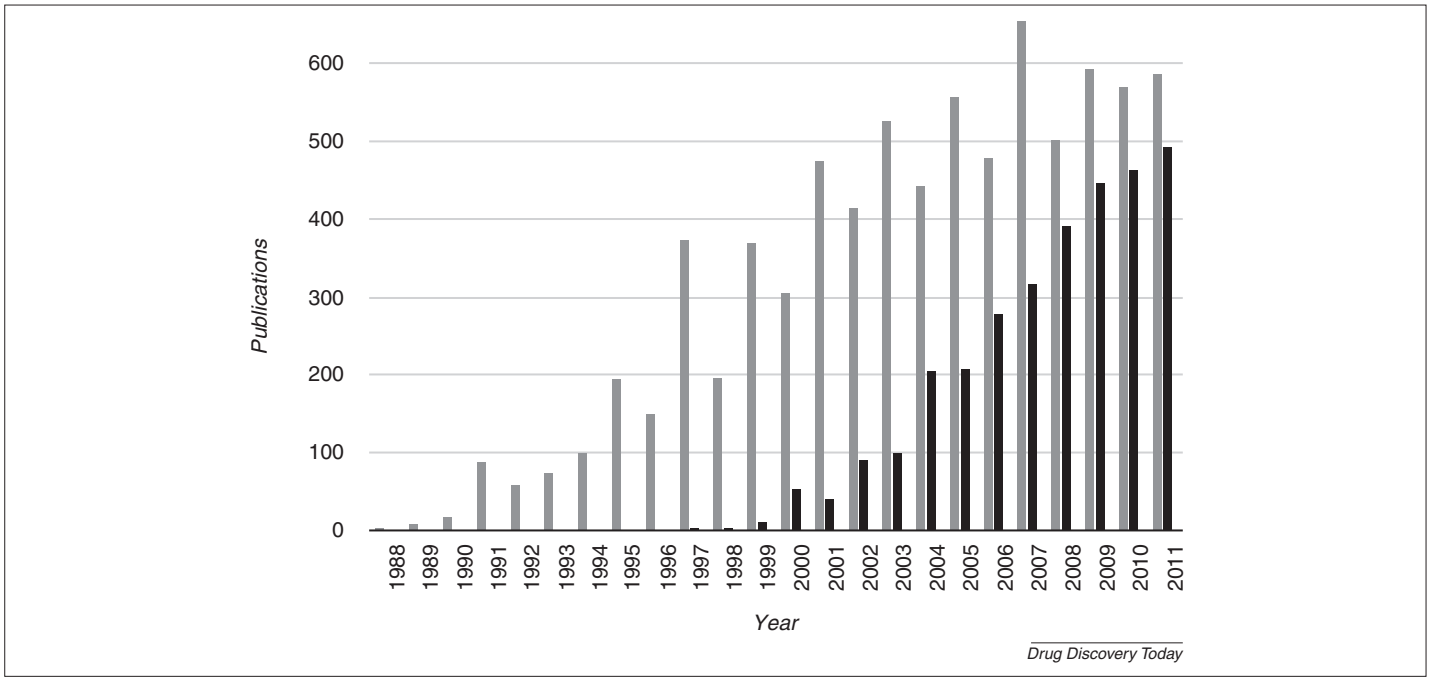

FIGURE 1

Chronological overview of the number of high-throughput screening (HTS; gray bars) and virtual screening (VS; black bars) publications according to ISI Web of Knowledge (Thomson Reuters, http://www.isiknowledge.com).

'ranking'). Often the layers are arranged according to the computational time required; however, the growth in computational power resulted in a tendency to apply computationally expensive methods even to large databases (e.g. high-throughput molecular docking). The final step usually comprises manual selection of compounds by experts, often referred to as 'cherry picking'.

Numerous classic VS studies have been extensively reviewed by Bajorath and coworkers [5,6]. The interested reader is referred to those articles, because here we emphasize the maturation of VS strategies.

Parallel applications of virtual screening

Another VS strategy is to apply multiple protocols in parallel and to combine the results (Fig. 2b). Often, these protocols cover various methods from different domains, including two- (2D) and three-dimensional (3D), ligand- and structure-based, similarity searching, machine learning and molecular modeling methods. The fundamental idea behind this parallelization is that each single method is complementary to the others in terms of the resulting virtual hit lists. Each single protocol is considered to be a classic approach, as described above. The fusion of multiple results helps to improve the overall performance by increasing the number of true positives and decreasing the number of false positives in the final selection [7]. Although the beneficial effect on the enrichment of true positive compounds has been studied thoroughly, the effect on true and/or false negatives remains largely unclear. The broad application of parallel VS emerged originally with the appearance of high-performance computational clusters in cheminformatics and computational chemistry working groups boosting the available processor time.

In general, parallel VS is a valid strategy to increase the enrichment rates. Thus, it is important to select the most suitable data fusion strategy for merging resulting virtual hit lists. Various fusion models (e.g. similarity or group fusion) have been described [7]. Furthermore, the application of an additional VS method as the last step of a fully parallelized approach has been observed. Here, we summarize selected studies exemplifying the use of parallel VS.

In 2005, coworkers from Sanofi-Aventis reported the discovery of blockers of the voltage-dependent potassium channel Kv1.5 by multiple VS approaches [8]. Given the lack of biological assays suitable for an HTS approach and the 3D protein structure, they used homology modeling to produce a receptor-based pharmacophore model. This was then used as a query in a VS of the compound library of the company, where 244 molecules had been selected for *in vitro* validation. In total, 19 were successfully confirmed as hits (a hit rate of 7.8%), and five compounds had an IC_{50} in the range of $<10\ \mu M$ up to 900 nM. Intermolecular pairwise distance measurements based on UNITY fingerprints (Tripos International, http://www.tripos.com) showed that if one of the five hits was used as query, none of the remaining hits would have been found, because of high structural dissimilarity. Repeating the same experiment based on Feature Trees [9] revealed only a single compound, because all the others had distances of less than a suggested similarity cutoff [10]. Interestingly, two additional VS approaches using 2D similarity searching and a ligand-based pharmacophore had been run previously. Both approaches also resulted in successful identifications of novel Kv1.5 blockers. However, the number of chemotypes identified was lower compared with the number of chemical classes identified via the receptor-based pharmacophore approach (five chemotypes). In addition, none of the identified hits was found by more than one of the VS approaches. This clearly shows the complementarity of VS techniques in terms of the identified hits. As a consequence, it was not necessary to apply more complex data fusion methods to