Chapter 10
Virtual Screening Methods Based on Bayesian Statistics

Martin Vogt
Rheinische Friedrich-Wilhelms-Universität, Germany

Jürgen Bajorath
Rheinische Friedrich-Wilhelms-Universität, Germany

ABSTRACT
Computational screening of in silico-formatted compound libraries, often termed virtual screening (VS), has become a standard approach in early-phase drug discovery. In analogy to experimental high-throughput screening (HTS), VS is mostly applied for hit identification, although other applications such as database filtering are also pursued. Contemporary VS approaches utilize target structure and/or ligand information as a starting point. A characteristic feature of current ligand-based VS approaches is that many of these methods differ substantially in the complexity of the underlying algorithms and also of the molecular representations that are utilized. In recent years, probabilistic VS methods have become increasingly popular in the field and are currently among the most widely used ligand-based approaches. In this contribution, the authors will introduce and discuss selected methodologies that are based on Bayesian principles.

INTRODUCTION
During the 1990ies, virtual screening (VS) approaches became increasingly popular in pharmaceutical research (Walters, Stahl, & Murcko, 1998). The early 1990s were also the time when rational drug design (Blundell, Jhoti, & Abell, 2002, Kuntz, 1992), or better structure-based drug design (because ligand-based design methods are no less “rational”), became a new paradigm in drug discovery, with all the promises coming along with it. The ensuing substantial increase in the number of X-ray structures of target proteins that became available from then on also opened the door for extensive applications of computational structure-based screening methods that are conventionally termed protein-ligand “docking” (Brooijmans & Kuntz, 2003, Kuntz, 1992) and that were originally developed during the 1980s (Kuntz, Blaney, Oatley, Langridge, & Ferrin,
Virtual Screening Methods Based on Bayesian Statistics

Thus, a first wave of VS that drew significant attention outside the expert community was largely structure-oriented, although ligand-based approaches to chemical database mining had also been extensively applied since the 1980s. These early ligand-based approaches (that play an important role to this date) included compound clustering algorithms (R. D. Brown & Martin, 1996, Willett, 1987), quantitative structure-activity relationship (QSAR) models (Hopfinger, 1980, Hopfinger et al., 1997), 2D molecular fingerprint representations (Willett, 2005) for similarity searching, and pharmacophore models for 3D database searching (Martin, 1992, Willett, 2005).

Ligand-based virtual screening approaches were further boosted during the 1990s by two important developments. Computational methods focusing on the assessment of whole-molecule similarity as an indicator of similar biological activity were increasingly introduced and widely recognized (Johnson & Maggiora, 1990). Furthermore, during the mid 1990s, the structure-based drug design paradigm was succeeded by another one focusing on high-throughput technologies and the pharmaceutical “number’s game”, i.e. more and more compounds were to be made and tested in a highly efficient manner in order to identify sufficient numbers of new candidates. Combinatorial chemistry (Czarnik, 1997) and HTS (Fox, Farr-Jones, & Yund, 1999) took center stage, compound libraries began to grow at unprecedented rates, and novel computational approaches were required for synthesis planning, library design and analysis, hence introducing the new field of molecular diversity analysis (Martin, 2001). Together with molecular similarity methods (Johnson & Maggiora, 1990), methodologies for diversity analysis including cell-based partitioning techniques (Pearlman & Smith, 1998) have substantially influenced the field of ligand-based VS by complementing more traditional QSAR and pharmacophore approaches. The term “chemoinformatics” first appeared in the literature in 1998 (F. K. Brown, 1998), and the majority of ligand-based VS methods are at present covered under the chemoinformatics umbrella (Bajorath, 2001). By the end of the 1990s, both structure- and ligand-based VS approaches were widely applied (Bajorath, 2001, 2002) and the popularity of ligand-based methods has grown ever since (Stahura & Bajorath, 2005).

Today, the leading publication venues for chemoinformatics and VS approaches including the Journal of Chemical Information and Modeling and the Journal of Medicinal Chemistry report novel methods and practical applications in this area on a regular basis and the field is steadily branching out and becoming increasingly heterogeneous from a methodological point of view. During the past few years the clear trend could be observed that machine learning methods originally developed in computer or information science were, and continue to be, adopted for ligand-based VS (Stahura & Bajorath, 2005). In addition to neural networks (Livingstone, Mann-lack, & Tetko, 1997) and self-organizing maps (Kohonen, 1989), currently especially popular approaches include kernel-based methodologies (Schölkopf & Smola, 2002) such as support vector machines (Jorissen & Gilson, 2005, Warmuth et al., 2003) and binary kernel discrimination (Harper, Bradshaw, Gittins, Green, & Leach, 2001, D. J. Wilton et al., 2006) and in addition probabilistic methods, in particular, Bayesian modeling (Duda, Hart, & Stork, 2000).

In this contribution, we will discuss different methodologies of varying complexity that are based on Bayesian statistics and that are relevant for ligand-based VS. These approaches include relevance weighting for substructural analysis, binary kernel discrimination, binary QSAR, and Bayesian classification. The theory underlying these methodologies and exemplary application examples will be presented. Special emphasis will be put on Bayesian VS methods that incorporate information-theoretic components in order to prioritize descriptors for the analysis of compounds with specific biological activity, dis-
Related Content

Symbolic Equation for the Instantaneous Amount of Substance in Linear Compartmental Systems: Software Furnishing the Coefficients Involved in it
[www.igi-global.com/chapter/symbolic-equation-instantaneous-amount-substance/56463?camid=4v1](www.igi-global.com/chapter/symbolic-equation-instantaneous-amount-substance/56463?camid=4v1)

Experimental-Theoretical Method for Defining Physical-Mechanical Properties of Polymer Materials with Regard to Change of Their Physical-Chemical Properties
[www.igi-global.com/chapter/experimental-theoretical-method-defining-physical/77078?camid=4v1](www.igi-global.com/chapter/experimental-theoretical-method-defining-physical/77078?camid=4v1)

3D Ligand-Based Virtual Screening with Support Vector Machines
[www.igi-global.com/chapter/ligand-based-virtual-screening-support/45464?camid=4v1](www.igi-global.com/chapter/ligand-based-virtual-screening-support/45464?camid=4v1)

[www.igi-global.com/chapter/virtual-screening-emergent-key-methodology/56458?camid=4v1](www.igi-global.com/chapter/virtual-screening-emergent-key-methodology/56458?camid=4v1)