Editorial

Machine Learning for Virtual Screening (Part 1)

Computer-assisted drug design is used to increase the chances of finding valuable drug candidates, by applying a wide range of computational methods, such as machine learning, structure-activity relationships, quantitative structure-activity relationships, molecular mechanics, quantum mechanics, molecular dynamics, and drug-protein docking. Machine learning is an important field of artificial intelligence, and includes a diversity of methods and algorithms that extract rules and functions from large datasets. The most important algorithms are linear discriminant analysis, artificial neural networks, decision trees, lazy learning, *k*-nearest neighbors, Bayesian methods, Gaussian processes, support vector machines, and kernel algorithms. This special issue presents a representative selection of machine learning applications for the virtual screening of chemical libraries.

In the opening paper, Melville, Burke and Hirst review recent applications of machine learning techniques in ranking chemical libraries based on their biological activity against a particular protein target. Applications of ligand-based similarity searching and structure-based docking are critically evaluated, with an accent on the major algorithms, such as decision trees, naïve Bayesian classifiers, artificial neural networks, and support vector machines.

Chen *et al.* examine the technical aspects of ligand-based virtual screening, such as available software, molecular descriptors, and performance measures. The procedures reviewed include binary kernel discrimination, *k*-nearest neighbors, linear discriminant analysis, logistic regression, and probabilistic neural networks. The detailed comparison of various studies is especially valuable in providing an estimate of the level of success that may be expected in virtual screening.

The comparison of various machine learning techniques is further explored by Plewczynski, Spieser and Koch in a large-scale evaluation of the screening success. Based on the biological targets explored in the literature, it was found that there is no machine learning approach that consistently provides the best results. Thorough careful tuning of parameters, most chemical libraries may be modeled with existing algorithms. The study found that a promising class of methods is represented by fusion (or ensemble) classifiers, which combine predictions from several models and are thus able to outperform single classifiers.

Burton *et al.* present an in-depth overview of recent advances in screening ligands of cytochromes P450. The most effective methods, which may reach 90% accuracy, are support vector machines, decision trees, artificial neural networks, *k*-nearest neighbors, and partial least squares.

Schneider *et al.* investigate the *de novo* design of novel ligand structures for various biological targets based on the software Flux. Extensive simulations show that this evolutionary *de novo* algorithm may reconstruct 27% of all compounds from a set of known ACE inhibitors and 17% of known aldose reductase inhibitors. A lower success rate was obtained for angiotensin-II receptor antagonists, but the algorithms may be improved by considering retrosynthetic routes to ring systems. Overall, the experiments demonstrate that Flux is a valuable tool in discovering novel lead structures.

The multiple-target screening method evaluates the docking scores of a chemical compound against a panel of biological targets. Fukunishi presents several methods to improve the multiple-target screening by a machine-learning score modification that computes a new screening score as a combination of docking scores that results in a maximum database enrichment. It is suggested that a combination of structure-based screening and ligand-based similarity evaluation provides higher database enrichment.

Machine learning algorithms evaluate the molecular similarity with various structural descriptors computed from the chemical structure. However, the molecular graph and the three-dimensional molecular structure may be used directly to compute the chemical similarity, as reviewed by Mahé and Vert for support vector machines and kernel methods. Novel molecular kernels may be thus obtained by translating directly the chemical structure into numerical scores of chemical similarity. Several applications of molecular kernels in structure-activity relationships are presented, demonstrating the modeling potential of these novel similarity functions.

Williams and Schreyer present an original algorithm, mutual information based activity labeling and scoring (MIBALS), for screening molecules based on mutual information analysis of 2D fingerprints. MIBALS was extensively tested in screening ligands for 40 different biological targets, and the results were promising compared with those obtained with traditional similarity search methods. MIBALS may be applied to identify important pharmacophore fragments, and to highlight beneficial and detrimental groups in a congeneric series of chemicals.

Lazy learning consists of a group of memory-based local learning methods, such as *k*-nearest neighbors, that delay all computations until a request is made to predict the biological activity of a chemical compound. Kulkarni, Jayaraman, and Kulkarni present a comprehensive overview of regression lazy learning, with detailed theoretical algorithms, practical applications, and critical assessment of its advantages and limitations. Lazy learning is a simple and robust method, which may provide predictive structure-activity models.

This special issue of *Combinatorial Chemistry and High Throughput Screening* will appear in two parts because of redactional, technical reasons. For further details relevant to this topic, see the second part (CCHTS Vol. *12*, No. *5*).

Ovidiu Ivanciuc
(***Guest Editor***)
Department of Biochemistry and Molecular Biology
University of Texas Medical Branch
301 University Boulevard
Galveston
TX 77555-0857
USA
E-mail: ivanciuc@gmail.com