Chemoinformatics and Advanced Machine Learning Perspectives:

Complex Computational Methods and Collaborative Techniques

Huma Lodhi
Imperial College, UK

Yoshihiro Yamanishi
Mines ParisTech—Institut Curie—Inserm U900, France

Medical Information Science REFERENCE

MEDICAL INFORMATION SCIENCE REFERENCE

Hershey · New York

Detailed Table of Contents

ecules. At the same time the physico-chemical properties of each single atom are considered as well as the neighborhood in the molecular graph. Later on our similarity measure is extended to deal with reduced graph representations, in which certain structural elements, like rings, donors or acceptors, are condensed in one single node of the graph. Comparisons of the optimal assignment kernel with other graph kernels as well as with classical descriptor based models show a significant improvement in prediction accuracy.

Chapter 3

We review an approach, proposed recently by Mahé, Ralaivola, Stoven, and Vert (2006), for ligand-based virtual screening with support vector machines using a kernel based on the 3D structure of the molecules. The kernel detects putative 3-point pharmacophores, and generalizes previous approaches based on 3-point pharmacophore fingerprints. It overcomes the categorization issue associated with the discretization step usually required for the construction of fingerprints, and leads to promising results on several benchmark datasets.

Chapter 4

This chapter analyses the use of similarity fusion in similarity searching of chemical databases. The ranked retrieval of molecules from a database can be modelled using both analytical and simulation approaches describing the similarities between an active reference structure and both the active and the non-active molecules in the database. The simulation model described here has the advantage that it can handle unmatched molecules, i.e., those that occur in one of the ranked similarity lists that are to be fused but that do not occur in the other. Our analyses provide insights into why the results of similarity fusion are often inconsistent when used for virtual screening.

Chapter 5

The objective of this chapter is to investigate the chemical information encompassed by autocorrelation descriptors and elucidate their role in QSAR and drug design. After a short introduction to molecular descriptors from a historical point of view, the chapter will focus on reviewing the different types of autocorrelation descriptors proposed in the literature so far. Then, some methodological topics related to multivariate data analysis will be overviewed paying particular attention to analysis of similarity/diversity of chemical spaces and feature selection for multiple linear regressions. The last part of the chapter will deal with application of autocorrelation descriptors to study similarity relationships of a set

of flavonoids and establish QSARs for predicting affinity constants, Ki, to the GABAA benzodiazepine receptor site, BzR.

Section 2
Graph-Based Approaches in Chemoinformatics

In standard QSAR (Quantitative Structure Activity Relationship) approaches, chemical compounds are represented as a set of physicochemical property descriptors, which are then used as numerical features for classification or regression. However, standard descriptors such as structural keys and fingerprints are not comprehensive enough in many cases. Since chemical compounds are naturally represented as attributed graphs, graph mining techniques allow us to create subgraph patterns (i.e., structural motifs) that can be used as additional descriptors. In this chapter, we present theoretically motivated QSAR algorithms that can automatically identify informative subgraph patterns. A graph mining subroutine is embedded in the mother algorithm and it is called repeatedly to collect patterns progressively. We present three variations that build on support vector machines (SVM), partial least squares regression (PLS) and least angle regression (LARS). In comparison to graph kernels, our methods are more interpretable, thereby allows chemists to identify salient subgraph features to improve the druglikeliness of lead compounds.

To annotate the biological function of a protein molecule, it is essential to have information on its 3D structure. Many successful methods for function prediction are based on determining structurally conserved regions because the functional residues are proved to be more conservative than others in protein evolution. Since the 3D conformation of a protein can be represented by a contact map graph, graph matching algorithms are often employed to identify the conserved residues in weakly homologous protein pairs. However, the general graph matching algorithm is computationally expensive because graph similarity searching is essentially a NP-hard problem. Parallel implementations of the graph matching are often exploited to speed up the process. In this chapter, we review theoretical and computational approaches of graph theory and the recently developed graph matching algorithms for protein function prediction.

Section 3
Statistical and Bayesian Approaches for Virtual Screenin

In quantitative structure-activity/property relationships (QSAR and QSPR), multivariate statistical methods are commonly used for analysis. Partial least squares (PLS) is of particular interest because it can analyze data with strongly collinear, noisy and numerous X variables, and also simultaneously model several response variables Y. Furthermore, PLS can provide us several prediction regions and diagnostic plots as statistical measures. PLS has evolved or changed for copying with sever demands from complex data X and Y structure. In this review article, we picked up four advanced PLS techniques and outlined their algorithms with representative examples. Especially, we made efforts to describe how to disclose the embedded inner relations in data and how to use their information for molecular design.

In many areas of research and industrial situations, including many data analytic problems in chemistry, a strong nonlinear relation between different sets of data may exist. While linear models may be a good simple approximation to these problems, when nonlinearity is severe they often perform unacceptably. The nonlinear partial least squares (PLS) method was developed in the area of chemical data analysis. A specific feature of PLS is that relations between sets of observed variables are modeled by means of latent variables usually not directly observed and measured. Since its introduction, two methodologically different concepts of fitting existing nonlinear relationships initiated development of a series of different nonlinear PLS models. General principles of the two concepts and representative models are reviewed in this chapter. The aim of the chapter is two-fold i) to clearly summarize achieved results and thus ii) to motivate development of new computationally efficient nonlinear PLS models with better performance and good interpretability.

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, we will introduce and discuss selected methodologies that are based on Bayesian principles.

In high throughput screening a large number of molecules are tested against a single target protein to determine binding affinity of each molecule to the target. The objective of such tests within the pharmaceutical industry is to identify potential drug-like lead molecules. Current technology allows for thousands of molecules to be tested inexpensively. The analysis of linking such biological data with molecular properties is thus becoming a major goal in both academic and pharmaceutical research. This chapter details how screening data can be augmented with high-dimensional descriptor data and how machine learning techniques can be utilised to build predictive models. The pyruvate kinase protein is used as a model target throughout the chapter. Binding affinity data from a public repository provide binding information on a large set of screened molecules. We consider three machine learning paradigms: Bayesian model averaging, Neural Networks, and Support Vector Machines. We apply algorithms from the three paradigms to three subsets of the data and comment on the relative merits of each. We also used the learnt models to classify the molecules in a large in-house molecular database that holds commercially available chemical structures from a large number of suppliers. We discuss the degree of agreement in compounds selected and ranked for three algorithms. Details of the technical challenges in such large scale classification and the ability of each paradigm to cope with these are put forward. The application of machine learning techniques to binding data augmented by high-dimensional can provide a powerful tool in compound testing. The emphasis of this work is on making very few assumptions or technical choices with regard to the machine learning techniques. This is to facilitate application of such techniques by non-experts.

Machine learning techniques have been widely used in drug discovery and development, particularly in the areas of cheminformatics, bioinformatics and other types of pharmaceutical research. It has been demonstrated they are suitable for large high dimensional data, and the models built with these methods can be used for robust external predictions. However, various problems and challenges still exist, and

new approaches are in great need. In this Chapter, we will review the current development of machine learning techniques, and especially focus on several machine learning techniques we developed as well as their application to model building, lead discovery via virtual screening, integration with molecular docking, and prediction of off-target properties. We will suggest some potential different avenues to unify different disciplines, such as cheminformatics, bioinformatics and systems biology, for the purpose of developing integrated in silico drug discovery and development approaches.

The problem of modeling and predicting complex structure-property relationships, such as the absorption, distribution, metabolism, and excretion of putative drug molecules is a fundamental one in contemporary drug discovery. An accurate model can not only be used to predict the behavior of a molecule and understand how structural variations may influence molecular property, but also to identify regions of molecular space that hold promise in context of a specific investigation. However, a variety of factors contribute to the difficulty of constructing robust structure activity models for such complex properties. These include conceptual issues related to how well the true bio-chemical property is accounted for by formulation of the specific learning strategy, algorithmic issues associated with determining the proper molecular descriptors, access to small quantities of data, possibly on tens of molecules only, due to the high cost and complexity of the experimental process, and the complex nature of bio-chemical phenomena underlying the data. This chapter attempts to address this problem from the rudiments: we first identify and discuss the salient computational issues that span (and complicate) structure-property modeling formulations and present a brief review of the state-of-the-art. We then consider a specific problem: that of modeling intestinal drug absorption, where many of the aforementioned factors play a role. In addressing them, our solution uses a novel characterization of molecular space based on the notion of surface-based molecular similarity. This is followed by identifying a statistically relevant set of molecular descriptors, which along with an appropriate machine learning technique, is used to build the structure-property model. We propose simultaneous use of both ratio and ordinal error-measures for model construction and validation. The applicability of the approach is demonstrated in a real world case study.

Predicting mutagenicity is a complex and challenging problem in chemoinformatics. Ames test is a biological method to assess mutagenicity of molecules. The dynamic growth in the repositories of molecules establishes a need to develop and apply effective and efficient computational techniques to solving chemoinformatics problems such as identification and classification of mutagens. Machine learning methods provide effective solutions to chemoinformatics problems. In this chapter we review learning techniques that have been developed and applied to the problem of identification and classification of mutagens.

The purpose of the olfactory system is to encode and classify odorants. Hence, its circuits have likely evolved to cope with this task in an efficient, quasi-optimal manner. In this chapter we present a three-step approach that emulate neurocomputational principles of the olfactory system to encode, transform and classify chemical data. In the first step, the original chemical stimulus space is encoded by virtual receptors. In the second step, the signals from these receptors are decorrelated by correlation-dependent lateral inhibition. The third step mimics olfactory scent perception by a machine learning classifier. We observed that the accuracy of scent prediction is significantly improved by decorrelation in the second stage. Moreover, we found that although the data transformation we propose is suited for dimensionality reduction, it is more robust against overdetermined data than principal component scores. We successfully used our method to predict bioactivity of drug-like compounds, demonstrating that it can provide an effective means to connect chemical space with biological activity.

Section 5
Machine Learning Approaches for Chemical Genomics

In silico prediction of compound-protein interactions from heterogeneous biological data is critical in the process of drug development. In this chapter we review several supervised machine learning methods to predict unknown compound-protein interactions from chemical structure and genomic sequence information simultaneously. We review several kernel-based algorithms from two different viewpoints: binary classification and dimension reduction. In the results, we demonstrate the usefulness of the methods on the prediction of drug-target interactions and ligand-protein interactions from chemical structure data and genomic sequence data.

Chemical genomics is one of the cutting-edge research areas in the post-genomic era, which requires a sophisticated integration of heterogeneous information, i.e., genomic and chemical information. Enzymes play key roles for dynamic behavior of living organisms, linking information in the chemical

space and genomic space. In this chapter, we report our recent efforts in this area, including the development of a similarity measure between two chemical compounds, a prediction system of a plausible enzyme for a given substrate and product pair, and two different approaches to predict the fate of a given compound in a metabolic pathway. General problems and possible future directions are also discussed, in hope to attract more activities from many researchers in this research area.