Early stages in modern drug discovery often involve screening small molecules for their effects on a selected protein target or a model of a biological pathway. In the past 15 years, innovative technologies that enable rapid synthesis and high-throughput screening of large libraries of compounds have been adopted in almost all major pharmaceutical and biotech companies. As a result, there has been a huge increase in the number of compounds available on a routine basis to quickly screen for novel drug candidates against new targets/pathways. In contrast, such technologies have rarely become available to the academic research community, thus limiting its ability to conduct large-scale chemical genetics or chemical genomics research. However, the landscape of publicly available experimental data collection methods for chemoinformatics has changed dramatically in very recent years. In 2005, the National Health Institute (NIH) launched a Molecular Libraries Initiative (MLI) that included the formation of the national Molecular Library Screening Centers Network (MLSCN). MLSCN aims to offer to the research community the results of testing about a million compounds against hundreds of biological targets.

Due to the broad application of high-throughput synthetic and analytical chemical technologies, scientists who generate large volumes of data are no longer equipped with adequate tools and approaches to manage and analyze this data. At the same time, the revolutionary development of information and communication technologies during the last few decades has changed dramatically our capabilities of collecting and accessing all sorts of molecular data and has afforded the creation of huge heterogeneous data depositories. For instance, the PubChem database has been developed by the NIH as the central repository for chemical structure–activity data. PubChem currently contains over 18 million chemical compound records, more than 700 bioassay results, and links from chemicals to bioassay description, literature, references and assay data for each entry. This data requires the development and application of sophisticated mathematical and statistical tools for the discovery of new patterns and structures in large chemical datasets and to achieve a deeper...
understanding of the relationship between chemical structure and physical and biological properties. From the computational knowledge mining prospective, the availability of large collections of chemical structures affords the opportunity for virtual screening of these collections in silico to identify and prioritize promising candidate compounds for experimental validation.

Virtual screening has been typically considered an area of computer-aided drug discovery where three-dimensional protein structures are used to discover small molecules that fit into the active site (docking) and have high predicted binding affinity (scoring). Traditional docking protocols and scoring functions rely on explicitly defined three-dimensional coordinates and standard definitions of atom types of both receptors and ligands. Albeit reasonably accurate in many cases, conventional structure-based virtual screening approaches are relatively computationally inefficient, which has precluded them from screening really large compound collections. Significant progress has been achieved over many years of research in developing many structure-based virtual screening approaches. However, several recent publications comparing many available scoring and docking approaches suggest that their accuracy still needs to be improved considerably to afford their automated and successful application to solve practical problems in drug design.\textsuperscript{3,4} Yet the availability of millions of compounds in chemical databases and billions of compounds in synthetically feasible virtual chemical libraries for virtual screening calls for the development of approaches that are both fast and accurate in their ability to identify a small number of viable computational hits that deserve subsequent experimental investigation.

Here we discuss the use of chemoinformatics as a powerful virtual screening methodology that presents both an alternative as well as complement to traditional structure-based docking and scoring approaches. The first published definition of cheminformatics\textsuperscript{5} defined it as:

\textit{the use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.}

This definition introduced by an industrial pharmaceutical scientist was obviously biased towards pharmaceutical applications. However, many years of research in multiple areas of chemistry, computational chemistry, chemometrics, molecular modeling, computer science and statistics, both before and after that publication, provide clear evidence that modern chemoinformatics appeals to almost any area of chemical research, including organic, physical, analytical chemistry and, more recently, systems biology.\textsuperscript{5} In this sense, following an early definition by G. Paris\textsuperscript{6} we describe chemoinformatics broadly

\textsuperscript{5}Both cheminformatics and chemoinformatics are used in the literature interchangeably and both spellings will be found in this book, depending on the personal preferences of the authors.
as a scientific discipline encompassing the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information.

We note that chemoinformatics is distinct from other computational molecular modeling approaches in that it uses unique representations of chemical structures in the form of multiple chemical descriptors, has its own metrics for defining similarity and diversity of chemical compound libraries, and applies a wide array of statistical, data mining and machine learning techniques to very large collections of chemical compounds in order to establish robust relationships between a chemical structure and its physical or biological properties. Chemoinformatics addresses a broad range of problems in chemistry and biology; however, the most commonly known applications of chemoinformatics approaches have arguably been in the area of drug discovery, where chemoinformatics tools have played a central role in the analysis and interpretation of structure–property data collected by the means of modern high-throughput screening.

Owing to the broad nature of chemoinformatics, several monographs have appeared recently that discuss various aspects of chemoinformatics research. The present book presents a unique focus on chemoinformatics approaches that are used for virtual screening of available collections of chemical compounds to identify novel biologically active molecules. The approaches discussed by the contributors rely on chemoinformatics concepts such as representation of molecules using multiple descriptors of chemical structures, advanced chemical similarity calculations in multidimensional descriptor spaces, the use of advanced machine learning and data mining approaches for building quantitative and predictive structure–activity models, the use of chemoinformatics methodologies for the analysis of drug-likeness and property prediction, and the emerging trend of combining chemoinformatics and bioinformatics concepts in structure-based drug discovery.

The chapters are organized in a logical flow that a typical chemoinformatics project would follow, i.e., from structure representation and comparison to data analysis and model building to applications of structure–property relationship models for hit identification and chemical library design. Chapter 1, by I. Baskin and A. Varnek, discusses the fundamental chemoinformatics concept of chemical structure representation by the means of molecular descriptors, focusing on fragment descriptors and their use in Quantitative Structure–Activity Relationship (QSAR) studies and database mining. This introductory chapter is followed by chapters by D. Horvath (Chapter 2) and by T. Langer and colleagues (Chapter 3) that discuss recent advances in pharmacophore identification and their use in virtual screening. Naturally, the pharmacophore is the major concept in medicinal chemistry and computational drug discovery, and many research papers and monographs have been published on this subject over the years. Still, these two chapters that have different focuses on pharmacophores derived from (two-dimensional) chemical graphs (Chapter 2) and, the more common, three-dimensional pharmacophores (Chapter 3) offer unique perspective on pharmacophore identification as a tool for knowledge discovery and mining in molecular databases.
Whereas pharmacophore identification can be viewed as an example of chemical data mining approaches focusing on specific descriptors of chemical structures, much information about structure–activity relationships can be obtained by using another major concept of chemoinformatics, *i.e.*, that of chemical (or molecular) similarity. Chapter 4 by L. Peltason and J. Bajorath summarizes recent advanced studies into this fundamental chemoinformatics problem and discusses the use of molecular similarity calculations in virtual screening. The next two chapters focus on recent methodologies that establish and explore quantitative structure–activity relationships (QSAR). E. Radchenko, V. Palyulin and N. Zefirov (Chapter 5) cover the use of topological molecular fields in drug design and virtual screening whereas D. Filimonov and V. Poroikov (Chapter 6) present an analysis of promising probabilistic approaches in QSAR modeling.

Chemoinformatics approaches are finding growing and important application in developing a better understanding of the chemical features that distinguish drugs and drug-like molecules from other organic molecules. In fact, this area so far has almost exclusively relied on ligand based approaches. Chapter 7 by G. Schneider and colleagues addresses the issue of drug-likeness and discusses ligand-based methodologies that can be used in designing viable drug candidates. Chapter 8, by I. Tetko and T. Oprea, presents an overview of chemoinformatics methods that are used in early stages of drug discovery to identify and prioritize compounds with optimal ADMET (Adsorption, Distribution, Metabolism, Excretion and Toxicity) properties.

Chemical library design has always been an important component of chemoinformatics studies and it could be viewed in fact a special case of virtual screening. Chapter 9 by W. Zheng and S.R. Johnson provides an expert overview of computational approaches that are employed in the design of targeted and diverse chemical libraries, including the use of property (*i.e.*, ADMET) filters.

The final chapter by A. Tropsha looks into chemoinformatics methodologies that rely on compound representation in multidimensional chemical descriptor space and chemical similarity searches that could be employed in structure-based drug discovery. These approaches could enrich traditional structure-based virtual screening and docking methodologies. The chapter may serve to illustrate the importance of building natural bridges between structural bioinformatics and structural chemoinformatics approaches in addressing the common problem of virtual screening that is the major theme of this book.

In conclusion, we believe that the focus on extending the experiences accumulated in chemoinformatics research towards virtual screening makes the theme of this monograph highly attractive for all computational and experimental researchers in the area of drug discovery or, more broadly, chemical biology. We stated at the beginning that virtual screening is one important area of modern chemoinformatics research that deserves special attention, which motivated us to develop this monograph. We believe, however, that due to its generic data-analytical focus we will see growing application of chemoinformatics approaches in multiple areas of chemical and biological research such as
synthesis planning, nanotechnology, proteomics, physical and analytical chemistry and, of course, chemical genomics.

Alexandre Varnek and Alexander Tropsha

Acknowledgements

The editors are grateful to Dr Dennis Fourches for designing the front cover illustration.

References

Chemoinformatics
An Approach to Virtual Screening
Varnek, A.; Tropsha, A. (Eds.)
2009, Hardcover
A product of Royal Society of Chemistry