
Preface

In the literature, several terms are used synonymously to name the topic of this book: chem-, chemi-, or chemo-informatics. A widely recognized definition of this discipline is the one by Frank Brown from 1998 (*1*) who defined chemoinformatics as the combination of “all the information resources that a scientist needs to optimize the properties of a ligand to become a drug.” In Brown’s definition, two aspects play a fundamentally important role: decision support by computational means and drug discovery, which distinguishes it from the term “chemical informatics” that was introduced at least ten years earlier and described as the application of information technology to chemistry (not with a specific focus on drug discovery). In addition, there is of course “chemometrics,” which is generally understood as the application of statistical methods to chemical data and the derivation of relevant statistical models and descriptors (*2*). The pharmaceutical focus of many developments and efforts in this area—and the current popularity of gene-to-drug or similar paradigms—is further reflected by the recent introduction of such terms as “discovery informatics” (*3*), which takes into account that gaining knowledge from chemical data alone is not sufficient to be ultimately successful in drug discovery. Such insights are well in accord with other views that the boundaries between bio- and chemoinformatics are fluid and that these disciplines should be closely combined or merged to significantly impact biotechnology or pharmaceutical research (*4*). Clearly, from an algorithmic or methodological point of view, bio- and chemoinformatics are much more similar to each other than many of their applications would suggest, at least on a first glance. It is fair to assume that the application of information science and technology to chemical or biological problems will further develop and mature, as well as continue to define, and redefine, itself.

If we wish to focus on chemoinformatics in a more narrow sense, what should we really consider? First, methods that support decision making in the context of pharmaceutical research (*2*) (such as compound design and selection) or methods that help interfacing computational and experimental programs (*4*) [such as virtual and biological screening (*5*)] are without doubt essential components. Second, equally important to developing methods and research tools is building and maintaining computational infrastructures to collect, organize, manage, and analyze chemical data. Third, I would propose

that it has also become increasingly difficult to distinguish between chemoinformatics and chemometrics, since statistical methods, models, and descriptors play a crucial role in, for example, similarity and diversity analysis or virtual screening. Fourth, approaches to explore (and exploit) structure–activity or structure–property relationships can hardly be excluded from chemoinformatics research, much of which aims at helping to identify or make better molecules. This means that approaches that are long disciplines in their own right such as QSAR or structure-based design can—and perhaps should—also be considered to contribute and belong to chemoinformatics. Lastly, evaluation of drug-likeness and prediction of downstream ADME characteristics of compounds have become highly relevant topics for chemoinformatics and drug discovery research and are approached using rather different concepts and algorithms.

Being confronted with the task of putting *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery* together, I decided to focus on authors and their individual contributions, rather than trying to address everything possible that could be covered under the chemoinformatics umbrella, as discussed above. It was my sincere hope that this approach would do justice to this still evolving and rather diverse field. Therefore, a variety of researchers (including well-recognized pioneers, senior scientists, and junior-level investigators) from diverse professional environments (academia, large pharmaceutical industry, and biotech companies) were asked to contribute. Chemoinformatics-relevant subject areas were initially outlined to provide some guidance, but authors were given as much freedom as possible in choosing their topics and designing their chapters. The result we are looking at is the rather diverse array of chapters I had initially hoped for. Certainly, many chapters go well beyond the introduction of single methods and protocols that is a major theme of the *Methods in Molecular Biology* series, at least as far as experimental science is concerned. Our contributions range from the description of specific methods or applications to the discussion of fundamentally important concepts and extensive review articles. On the other hand, some of the topics I initially envisioned to cover are missing, for example, neural network simulations or chemical genetics, to name just two. By contrast, some contributions present and discuss similar methods, for example, compound selection or library design, in rather different ways, which I find particularly interesting and stimulating.

Chemoinformatics: Concepts, Methods, and Tools for Discovery begins with an elaborate theoretical discussion of the concept of molecular similarity by Maggiora & Shanmugasundaram that is one of the origins and cornerstones of chemoinformatics as we understand it today. Chapter 2 by Willett follows up on this theme and extends the discussion to molecular diversity, a related

—yet distinct—and equally fundamental concept. Following these methodological considerations, Bembenek & colleagues describe a computational infrastructure to enable pharmaceutical researchers to efficiently access basic chemoinformatics tools and help in decision-making. Chapters 4 and 5 by Parker & Schreyer and Lajiness & Shanmugasundaram describe efforts to interface chemoinformatics approaches with high-throughput screening and with screening and medicinal chemistry, respectively. As discussed above, the formation of such interfaces is one of the major challenges—and opportunities—for chemoinformatics in pharmaceutical research.

Esposito & colleagues provide an extensive discussion of QSAR approaches in Chapter 6. The authors review basic principles and methods and then focus on the latest developments in multidimensional QSAR analysis. In the following chapter, Gomar & colleagues describe the development of a lipophilicity descriptor that alleviates the molecular alignment problem in QSAR and discuss exemplary applications. In general, the majority of chemoinformatics applications critically depend on the use of descriptors of molecular structure and properties, and Chapter 8 by Labute presents a good example of descriptor design. The author describes the generation of a novel class of molecular surface property descriptors that can be readily calculated from 2D representations of molecular structures.

The next four chapters focus on partitioning algorithms and classification methods that have become very popular for the analysis of large compound databases, screening sets, and virtual screening for active molecules. Xue & colleagues describe cell-based partitioning based on principal component analysis and, to contrast with chemical space dimension reduction methods, Godden & Bajorath introduce a statistically based partitioning algorithm that directly operates in higher-dimensional, albeit simplified, chemical descriptor spaces. In the following back-to-back chapters, Lam & Welch first apply clustering and cell-based partitioning methods for the selection of active compounds from the HIV data set of the National Cancer Institute. Based on their computational scheme and results, Young & Hawkins apply recursive partitioning (another statistical approach) to the same data set, thus enabling direct comparisons.

Following these compound classification and selection methods, Chapters 13–15 describe different approaches to compound library design. Gillet discusses a genetic algorithm-based method to simultaneously optimize multiple objectives or properties when designing libraries. Schnur & colleagues describe various approaches to focus compound libraries on families of therapeutic targets, which represents a major trend in drug discovery, and Zheng introduces simulated annealing as a stochastic approach to library design.

In Chapter 16, Lavine & colleagues return to a compound classification problem by using a combination of principal component analysis and a genetic algo-

rithm that is here applied to an optimization problem different from the one discussed by Gillet. In the next chapters, Crippen introduces novel ways of describing molecular chirality and conformational parameters with relevance for the analysis of structure–activity relationships, and Pick provides a brief review of scoring functions for structure-based virtual screening. The book ends with an extensive and detailed description by Jalaie & colleagues of different types of methods, including structure-based approaches, to predict drug-like character of compounds and basic ADME properties based on modeling their putative interactions with cytochrome P450 isoforms, which are important drug metabolizing enzymes. This discussion complements other major themes represented herein including molecular similarity, structure-activity relationships, and compound classification and design.

First and foremost, I would like to thank our authors whose diverse contributions have made this project a (hopefully, interesting!) reality.

Jürgen Bajorath

References

1. Brown, F. K. (1998) Chemoinformatics: What is it and how does it impact drug discovery. *Ann. Rep. Med. Chem.* **33**, 375–384.
2. Goodman, J. M. (2003) Chemical informatics. *Chem. Inf. Lett.* **6** (2); http://www.ch.cam.ac.uk/MMRG/CIL/cil_v6n2.html#14
3. Claus, B. L. and Underwood, D. J. (2002) Discovery informatics: Its evolving role in drug discovery. *Drug Discov. Today* **7**, 957–966.
4. Bajorath, J. (2001) Rational drug discovery revisited: Interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* **6**, 989–995.
5. Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* **1**, 882–894.

Chemoinformatics

Concepts, Methods, and Tools for Drug Discovery

Bajorath, J. (Ed.)

2004, XIV, 524 p., Hardcover

ISBN: 978-1-58829-261-2

A product of Humana Press