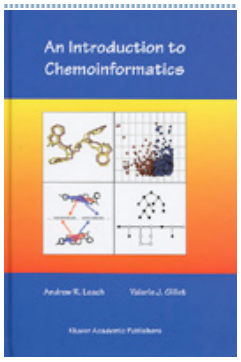
An Introduction to Chemoinformatics

Andrew R. Leach & Valerie J. Gillet
£ 52.00 $ 83.00 EUR 75.00
Kluwer Academic Publishers, pp 260
Hardback ISBN 1402013477



Chemoinformatics is the application of computational methods to chemical problems, with particular emphasis on the manipulation of structural information.

While this type of work has been done for a long time, its application has increased dramatically in recent years, in part to deal with the huge quantities of data generated by high-throughput screening and combinatorial chemistry. The concomitant increases in computing power have allowed for development of many new methods as well.

As the title suggests, this volume serves as introduction for those who are not at all familiar with the field, and was targeted for graduate students, senior undergraduates, and professional scientists. It assumes a basic knowledge of chemistry and some math. It is a brief skimming of many topics that each could fill a volume in their own right. There are numerous literature references throughout the chapters, and at the end of the book (where I almost missed it) is a list of suggestions for further reading on the topic of each chapter for those interested in pursing the topics further.

This volume has 9 chapters, each of which is described in more detail below.

The first chapter is entitled 'Representation and Manipulation of 2D Molecular Structures'. It discusses some of the methods used to store chemical structures in large databases, as well as methods for searching for specific substructures within compounds. The methods described in this chapter are concerned only with connection information between atoms, not three-dimensional data. They include graph theory, connection tables and linear notations (SMILES), and canonical structure representations. Mention is made of reaction databases, and Markush representations which are important in chemical patents. Two-dimensional databases are used to store and retrieve complete chemical structures and their associated properties, but to search for substructures, or fragments of structures that are identical. For example, all structures containing a benzene ring may be retrieved from a database. Various methods of implementing such substructure searches are described.

'Representation and Manipulation of 3D Molecular Structures' are covered in the next chapter. In addition to the types of atoms and connectivity between them, 3D descriptors also incorporate

related

links to ChemWeb.com:

An Introduction to Chemoinformatics

Bookstore

external site(s) related to this article:

Cambridge Crystallographic Data Centre

Protein Data Bank

services

EMAIL THIS STORY TO FRIEND

Type e-mail address then return:

SEND

information about the relative orientation of those atoms in space. Those orientations in turn affect the steric and electronic properties of the molecules. While 3D methods are particularly useful for identifying molecules with similar properties but very different underlying structure, they are much more complicated to work with than 2D methods. This complication arises mainly from the fact that molecules large enough to be of interest can adopt multiple low energy conformations.

Many methods have been devised to balance the competing needs of computational efficiency and accurately representing the multitude of low-energy conformations for a single molecule. The two best known databases of 3D structures are the Cambridge Structural Database and the Protein Databank, each containing single, high resolution crystal or NMR structures of individual molecules, and both are briefly described. Methods of searching 3D databases, including two-stage methods are briefly mentioned. Since crystal structures are not available for all compounds, databases of theoretical structures are also used. Various ways to create and search (including constrained systematic search, clique detection, maximum likelihood, genetic algorithms) databases of theoretical structures are described.

Sometimes molecular descriptors are used, instead of working with complete structures. Molecular descriptors are numerical values used to characterize specific properties, and are the subject of chapter 3. The emphasis here is on descriptors that represent properties of whole molecules, rather than substituents. Two-dimensional descriptors described in some detail include simple counts, physiochemical properties, molar refractivity, topological indices, kappa shape indices, electrotopological state indices, 2D fingerprints, atom-pairs and topological torsions, and BCUT descriptors. Three-dimensional descriptors include fragment screens and pharmacophore keys. A section is devoted to validating descriptors, and another section to methods to removing redundancy from multiple descriptors.

Chapter 4 discusses 'Computational Models', specifically Quantitative Structure-Activity Relationships (QSARs) or Quantitative Structure-Property Relationships (QSPRs). A brief historical review is followed by sections describing simple and multiple linear regression, least squared correlation coefficient (R^2), and the cross-validation and standard error of prediction methods of performing a reality check on calculated values. Guidelines for designing a QSAR experiment are given. The principal components regression method to reduce the number of variables used to fit the data is discussed, as is the partial least squares method and molecular field analysis.

Similarity searching is an alternative to substructure or pharmacophore searching. Instead of identifying a small piece of the molecule and searching for other molecules that contain that same piece, similarity searching takes an entire molecule and searches for other molecules that are similar to the query molecule. Each molecule in the database is scored as to how similar it is to the query molecule. How to define 'similar' is a complex question, and a number of both 2D and 3D similarity metric methods are discussed. The 2D methods include fingerprints, similarity coefficients, and maximum common subgraph. Most 3D methods require alignment of the molecules, and many methods to determine the molecular overlay that maximizes similarity are discussed.

Chapter 6 is entitled 'Selecting Diverse Sets Of Compounds'. Instead of selecting similar compounds, sometimes what is required is a set of compounds that maximally cover structure space. Diversity methods allow navigation through relevant chemical space, and the identification of potentially useful subsets. Some of the methods discussed are cluster analysis, including both hierarchical and non-hierarchical, dissimilarity-based methods, cell-based methods, and optimization methods. An overview at the end of the chapter discusses how the various methods work, and when each might be

appropriate.

One of the main driving forces behind developments in chemical information techniques is the need to analyse high-throughput screening data, and this is the subject of chapter 7. The typical types of data collected are described, as well as methods of data visualization and data mining (substructure analysis, discriminant analysis, neural networks, and decision trees). The summary section points out that these methods are currently being developed, and further research is required to determine which method works best under which circumstances.

Virtual Screening, or the scoring and ranking of predicted activity of compounds completely *in silico*, is the subject of the next chapter. Since virtual combinatorial libraries can run into billions of compounds, the speed of the scoring function is crucial. Simple rules and filters used to predict drug-likeness are discussed, as are structure-based screening methods such as protein–ligand docking. Practical matters are discussed — how to set up computational experiments so the results are both scientifically valid and timely. A substantial section talks about prediction of ADMET (absorption, distribution, metabolism, excretion and toxicity) properties, those pesky biological processes that keep active compounds from becoming marketable drugs.

'Combinatorial Chemistry and Library Design' is the final chapter. It discusses strategies that can be used in the design of the types of libraries that were searched in the previous chapter. Achieving a balance between diversity and focus is crucial. The chapter begins with a brief introduction to how combinatorial libraries are synthesized, then moves on to enumeration strategies and library design strategies.

Lisa M. Balbes, PhD, is a freelance consultant and technical writer. She specialises in scientific software and bio/chemoinformatics, as Balbes Consultants.

Opinions expressed here are those of the reviewer and not necessarily those of Elsevier.

Lisa M. Balbes
16 March 2004