Preface

Drug discovery is a complicated process, requiring skills and knowledge from many fields: biology, chemistry, mathematics, statistics, computer science, and physics. Almost without exception, even the simplest drug interactions within the body involve extremely complicated processes. Professionals in numerous sub-fields (cell biologists, medicinal chemists, synthetic chemists, statisticians, etc.) make up teams that, in sequence and in parallel, form a network of research and process that leads — hopefully — to new drugs and to a greater understanding of human health and disease, as well as, more generally, the biological processes that constitute life. The advent of computer technology in the latter half of the 20th century allowed for the collection, management, and manipulation of the enormity of data and information flowing from these endeavors. The fields of cheminformatics and bioinformatics are the outgrowth of the combination of computing resources and scientific data. Modern drug discovery increasingly links these two fields, and draws on both chemistry and biology for a greater understanding of the drug–disease interaction.

Cluster analysis is an exploratory data analysis tool that naturally spans many fields, and bioinformatics and cheminformatics as they relate to drug discovery are no exception. Exploratory data analysis is often used at the outset of a scientific enquiry; experiments are designed, performed, and data are collected. The experiments that collect observations may be entirely exploratory, in that they do not fit under the more rigorous statistical confines of *design of experiments* methodology, whereby hypotheses are being tested directly. Experiments can simply be designed to collect observations that will be *mined* for possible structure and information. Namely, the scientific process is often not quite as straightforward nor as simple as a recanting of the scientific method might suggest; hypotheses do not necessarily appear whole cloth with a few observations. There are a great many fits and starts in even understanding what questions might be asked of a set of observations. Intuitions about observations may or may not be forthcoming, especially in the face of enormous complexity and vast numbers of observations. Finding structure in a set of observations may help to provide intuitions based on known facts that lead to hypotheses that may bear fruit from subsequent testing — where the scientific method earns its keep. Cluster analysis is a set of computational methods that quantify structure or classes within a data set: are there groups in the data? This effort is in turn frought with the possible misunderstanding

that any structure found through these methods must be something other than a random artifact. The observations may not mean much of anything!

This book is an exploration of the use of cluster analysis in the allied fields of bioinformatics and cheminformatics as they relate to drug discovery, and it is an attempt to help the practitioners in the field to navigate the use and misuse of these methods—that they can in the end understand the relative merits of clustering methods with the data they have at hand, and that they can evaluate and inspect the results in the hope that they can thereby develop useful hypotheses to test.

How to Use This Book

The first three chapters set the stage for all that follows. Chapters 4, 5, and 6 can be used to provide a student with sufficient detail to have a basic understanding of cluster analysis for bioinformatics and drug discovery. Such a selection of the text could be used as an introduction to cluster analysis over a 2- or 3-week period as a portion of a semester-long, introductory class. Students or practitioners interested in additional and more advanced methods, and subtle details related to specific types of data, would do well to cover the remaining chapters, 7 through 11. This would represent a more advanced, semester-long course or seminar in cluster analysis.

For supplementary materials, a great deal can be obtained online, including descriptions of methods, data, and free and open source software. However, for more extensive treatment of cluster analysis in general there are numerous texts of varying quality published over the past 40 years, the most outstanding of which, though somewhat outdated from a computational point of view, is *Algorithms for Clustering Data* by Jain and Dubes. It is, however, out of print and difficult to find. A more accessible and obtainable text is *Finding Groups in Data* by Kaufman and Rousseuw. Algorithms in this latter text can be found in R and S-Plus. A passing familiarity with R or S-Plus is extremely handy in exploring cluster algorithms directly. Students and practitioners would do themselves a significant service by learning the rudiments of one of these nearly identical languages. Readers having some exposure to interpreted languages such as Python, or those found in software packages such as Mathematica, Maple, or MATLAB, should have little trouble picking up the syntax and constructs of R or S. Note: all of the figures in the text are in grayscale, but many of the more complex figures have color representations that can be found on the DVD included with this book.