

Preface

Clustering can be defined as the partitioning of a data set into subsets (called *clusters*), so that each subset consists of elements that are similar with respect to some similarity criterion. The measure of similarity can be the distance between the data points (used in *distance-based clustering*) or some descriptive concept (as in *conceptual clustering*), and can be chosen differently depending on the type of the data set of interest and the purpose of clustering. The typical objectives include the data classification and reduction, detection of natural modules based on their properties, and the determination of outliers. Clustering algorithms have been successfully used to analyze the data sets arising in many important applications of diverse origin, including biology. In fact, applications of clustering in biology can be traced back to Aristotle's *History of Animals*, in which he classified plants and animals according to complexity of their structure and function.

Many biological systems can be conveniently modeled using graphs or networks, with vertices representing the data points and edges connecting pairs of vertices corresponding to data points that are related in a certain way. Network clustering and cluster detection algorithms represent an important tool in structural analysis of such networks. For example, in gene networks, the vertices correspond to genes and the edges represent functional relations between these genes that are identified using the comparative genomics methods. Solving clustering problems in gene networks allows to identify groups of genes with similar expression patterns. This information is crucial for understanding the nature of genetic diseases. Other examples of biological networks include the protein interaction networks, metabolic networks, and signaling networks.

Network clustering problems present a number of formidable research challenges, many of which are still to be addressed. On the one hand, developing a proper mathematical model describing the clusters that are interesting from biological perspective may be very tricky. On the other hand, most known optimization problems on graphs used as the basis for network clustering appear to be NP-hard, making it extremely difficult to solve large-scale instances of such problems to optimality. Based on the objective that clustering aims to achieve for

a particular application, one has to choose an appropriate graph-theoretic definition of a cluster, formulate the corresponding optimization problem, and develop a network clustering algorithm for the developed model. From the practical perspective, the effectiveness of any clustering algorithm has to be confirmed through empirical evidence, and this process is complicated by possible errors in the data used to construct the network.

This volume presents a collection of papers, several of which have been presented at DIMACS Workshop on Clustering Problems in Biological Networks that took place at Rutgers University on May 9 - 11, 2006. It consists of two parts, with the first part containing surveys of selected topics and the second part presenting original research contributions. While clustering in biological networks represents the central theme of this volume, some of the chapters deal with other related problems in computational biology that may not necessarily fall within the vaguely defined network clustering domain.

This book will be a valuable source of material to faculty, students, and researchers in mathematical programming, data analysis and data mining, as well as people working in computer science, engineering and applied mathematics. In addition, the book can be used as a supplement to any course in data mining or computational/systems biology.

We would like to thank the authors of the chapters, the anonymous referees and the staff of World Scientific for their cooperation, without which the publication of this volume would not have been possible. We also acknowledge the support of the National Science Foundation in organizing the DIMACS Workshop mentioned above.

Sergiy Butenko, W. Art Chaovalitwongse, and Panos M. Pardalos
September 2008