# Preface

Bioinformatics is a new and interdisciplinary field which is concerned with the development of methods for recording, organising, and analysing biological data. With the advent of human genome-sequencing, microarrays and high-performance computing, developing software tools to generate useful biological knowledge has been a major activity.

Clustering techniques have been increasingly used in the analysis of high-throughput biological datasets. Although many generic clustering methods have been used successfully to analyse biological datasets, many specific properties of those datasets require customised methods which are specifically designed to meet such properties. Therefore, both aspects – the design and the application of clustering methods in this field – have been under active investigations and considerations by many research groups around the world. Indeed, there have been many activities in assimilating the work in this field, especially by designing state-of-the-art methods which uniquely address various issues that are particularly relevant in biological data.

This book attempts to outline the complete pathway from the basics of molecular biology to the generation of biological knowledge. It is supplied with an introductory part to molecular biology at a level which can be understood by researchers coming from a numeric background, such as computer scientists and information engineers. The introductory part helps those readers to get introduced to the basic biological knowledge needed to appreciate the specific applications of the methods in this book. The book also explains the structure and properties of many types of high-throughput datasets commonly found in biological studies, including public repositories/databases, pre-processing like normalisation and the identification of differentially expressed genes. A major part of the book will cover various clustering methods and clustering-validation techniques as well as their applications to biological datasets, representing an integrative analysis. It should be remarked that not all clustering methods have been utilised in bioinformatics yet. Some of the most recent state-of-the-art clustering methods, which can deal with specific problems that appear in biological datasets, are paid much attention, especially in how they, and their possible successors, could be used to enhance the pace of biological discoveries in the future. Although proposed in the context of bioinformatics, some

specialised sophisticated methods can also be used by other researchers to apply them to other analogous problems. Therefore, the general community of researchers in the field of machine learning are also targeted by most of the contents of this book.

There are books mainly focusing on various aspects related to microarrays, such as biological experimental design, image processing, identification of differentially expressed genes, supervised classification, etc., while covering clustering analysis at a less thorough level. In any case, these tend to focus on clustering of microarray datasets rather than considering a wider range of biological datasets. Yet, other books provide more thorough coverage of clustering than the aforementioned books, but they do not provide sufficient background in the field of molecular biology for researchers coming from numeric backgrounds to appreciate and understand the origins of the available datasets. These kinds of book tend to be mainly data-clustering books and naturally belong to the computational side of bioinformatics rather than to the interface between the computational and the biological sides. This book does sit at this interface and goes far beyond those currently available, in that it presents some biological preliminaries as well as some state-of-the-art clustering methods that are specifically designed to suit specific issues which appear in biological datasets.

As the field is still developing, such a book cannot be definitive or complete. This book is designed to target researchers in bioinformatics ranging from entry level researchers (e.g. senior bachelor students and master students) to the most senior researchers (e.g. heads of research groups). It is hoped that graduate students should be able to learn enough basics before studying journal papers; researchers in related fields should be able to get a broad perspective on what has been achieved; and current researchers in this field should be able to use it as a reference.

Further to the material provided in the book, a companion website hosting a selected collection of software and links to publicly available datasets can be accessed by using the following URL: https://code.google.com/p/integrative-cluster-bioinformatics/.

A work of this magnitude will unfortunately contain errors and omissions. We would like to take this opportunity to apologise unreservedly for all such indiscretions in advance. We would welcome comments and corrections; please send them by email to a.k.nandi@ieee.org or by any other means.

Basel Abu-Jamous, Rui Fa and Asoke K. Nandi
London, UK
Feb 2015