Preface

Volume 183 of *Methods in Enzymology* dealing with the computer analysis of protein and nucleic acid sequences has proved very popular with molecular biologists and biochemists. Computers and computer programs evolve rapidly, however, and can become outmoded very quickly. As a result, there was pressure to issue an updated volume that covers much the same general subject areas.

Like the earlier volume, this one is divided into several sections, the first of which deals with databases and some aspects related to their holdings. Also, there have been some relocations of major databases. GenBank is now centered at the National Center for Biotechnology Information (NCBI) at the National Library of Medicine in Bethesda, Maryland, and the EMBL Database has relocated to the European Bioinformatics Institute (EBI) at a site just outside Cambridge, England. More than ever, of course, geographic location is becoming moot, thanks to the World Wide Web (WWW) and extended hyperlink access.

There is some new vocabulary in this volume that did not appear in Volume 183. The use of neural nets, for example, is discussed in several places, including chapters dealing with the classification of sequences, on the one hand, and with predicting secondary structure, on the other. The kinds of databases are also changing. For instance, it has been found that the fragmentary data known as Expressed Sequence Tags (EST) are extremely useful.

Searching newly determined sequences remains the first order of business. More often than not, a simple search of a new sequence provides both functional and structural information. New pattern searching programs have greatly extended the power of this approach so that very distant relatives of well-characterized families can be identified.

The multiple alignment of protein sequences continues to have a prominent role in protein characterization. Whether the sequences are of the "same" protein from different organisms or are paralogs that have resulted from gene duplications, the alignment problems are the same. Interestingly, the most popular algorithms have not changed much, but the amino acid substitution tables that support them have. This is chiefly the result of there being so much comparative data in the current databases that empirical measures of relationships can be obtained by simply tallying the occurrences of the amino acids in blocks of obviously aligned sequences. As discussed in Chapter [6] by Henikoff and Henikoff, these BLOSUM tables have been remarkably effective.

Among their many uses, multiple alignments are used to construct profiles for more sensitive searching than is possible by single-searching. They are also used in the consensus mode for better predictions of secondary structure and for three-dimensional searches. And, of course, they are used in the construction of phylogenetic trees.

Recent advances have led to some changes in emphasis in some of the sections. Most of the chapters focus on protein sequences, even though the vast majority of those are determined by DNA sequencing. Accordingly, a section on RNA folding that appeared in the earlier volume has been dropped, and instead a number of chapters that relate to the secondary structure and three-dimensional aspects of proteins have been added.

Indeed, three-dimensional searching is following the course of sequence searching a decade ago. As a new protein structure is characterized, the first matter of general interest is to determine whether the fold resembles that of any that were reported previously. The remarkable thing is that not only are most new structures falling into well-defined families, but often there is no hint in advance on the basis of either structure or function. The problems associated with structure searching are similar to those experienced by sequence searchers in the past: a burgeoning data bank (PDB is the Protein Data Bank), choices of search programs, and, finally, the problem of judgment on how significant a resemblance may be. Many of these problems are addressed in Section V of this volume.

As with Volume 183, authors were encouraged to make their programs or databases available to readers. Many chapters make reference to a WWW home page or an Internet email address from which additional information can be extracted.

Finally, I thank all the authors who wrote such interesting and informative chapters under a very strict and compressed timetable. Academic Press, and especially our editor, Shirley Light, outdid themselves in getting the manuscripts through the publication process in record time. As in the case of the previous volume dealing with this topic, I must also acknowledge that the task could not have been accomplished without the help of my assistant, Karen Anderson. Her relentless but always gentle prodding of authors to produce manuscripts and her remarkable organizational skills that kept the courier traffic flowing in the right direction were indispensable.

RUSSELL F. DOOLITTLE