BOOK REVIEW

David Sankoff and Joseph B. Kruskal, Editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison Wesley Publishing Company, 1983, 382 pp., $31.95 hardback

---

Linear, sequential data are an important part of the information we receive, store, and process. This page of letters and words is an example of these data. Much biological data also fall into this category. Examples include 1) population sizes recorded at regular intervals, 2) recordings of vocalizations of animals, such as birds, and 3) nucleic acid and protein sequence data.

The readers of this journal are of course interested in the analysis of biological data. Sequential data require special methods of analyses. Statistical analyses come to mind, where the researcher might want to estimate the order of a Markov chain, to test for stationarity, or to find periodicities. Curve fitting is also important and is closely linked to prediction of future data points. Finally, the techniques of pattern recognition, developed because of the availability of computers, can be applied to sequences. One such technique is the subject of the book being reviewed here.

In the early 1970s, motivated by evolution of DNA sequences, Stan Ulam popularized the following problem: Given finite sequences x and y, find the minimum number of substitutions, insertions, and deletions required to change x into y. Needleman and Wunsch, Sankoff, and Sellers all gave solutions of the problem using the methods of dynamic programming. At the same time, Wagner and Fischer, motivated by the problem of string comparisons in computer science, gave essentially the same solution. These solutions utilized dynamic programming and the methods have been modified for a wide variety of problems. The book *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* looks at many aspects of sequence comparison by dynamic programming.

Dynamic programming methods have often been applied to macromolecular sequences. In this book various authors have fully discussed the comparison of two molecular sequences. The formulation above needs modification because of different weights required for different substitutions and for long insertions and deletions. In addition, finding highly similar segments of two otherwise unrelated sequences is of much interest and requires new algorithms. In fact in one article, Erickson and Sellers locate a cyclically permuted pattern in two protein sequences. Comparison of several sequences is another important topic and is the subject of a chapter by Sankoff and Cedergren.

Another problem involving macromolecular sequences arises from the fact that single-stranded RNA molecules fold back on themselves to form helical

52 Vanderbilt Ave., New York, NY 10017 0025-5564/85/$03.30

regions. Such configurations are called secondary structure. The prediction of RNA secondary structure was put on a rigorous basis a few years ago by a novel application of dynamic programming. This topic is discussed and advanced here by Sankoff, Kruskal, Mainville, and Cedergren.

Speech recognition is a popular topic these days. Sound waves are sampled at regular intervals and converted into discrete sequences. Compression and expansion of speech (time warping) can be handled in much the same way as insertion and deletion of letters. Time warping allows segments of the recorded sequence to match sequences in a dictionary of syllables. Then the sequence of syllables is compared with a dictionary of phrases. Chapters by Kruskal and Liberman and by Hunt, Lennig, and Mermelstein treat these topics, while a highly interesting chapter by Bradley and Bradley discusses bird song comparisons.

Computer science topics are not omitted. One chapter, by Wagner, explores the computational complexity of including transpositions, while another chapter treats formal language error correction. Chapters by Hirschberg and by Masek and Paterson also study computational complexity. Noetzel and Selkow study the tree edit problem with insertion, deletion, and substitution of nodes.

The topic of statistical significance of sequence distances is not fully developed because of the difficulty of the problems. Chvatal and Sankoff began the subject in 1975 with a paper on the length of the longest common subsequence of two random sequences of length n. It can be shown that the expected length will approach some constant times n, but the constant has not been obtained theoretically. Papers by Chvatal and Sankoff, Deken, and Sankoff and Mainville all study the longest common subsequence problem.

This is a well-written and well-edited book on an important topic. In molecular sequence analysis, there are so many ad hoc and incorrect methods proposed that it is a pleasure to read this book in contrast. Anyone involved in the theory or practice of sequence comparison should own a copy. It is not a textbook; however it might be used for a seminar. The book does limit itself to method of solution. Many problems, such as determining consensus sequences in many sequences, will never yield to dynamic programming. This book's topic is the theory and practice of dynamic programming sequence comparisons, and it gives an excellent treatment of that subject.

MICHAEL S. WATERMAN
Departments of Mathematics and of Molecular Biology
University of Southern California
Los Angeles, California 90089-1113