Book Reviews

Book Review

Introduction to Computational Biology

Author: Michael S. Waterman;
published by Chapman & Hall, 1995
ISBN: 0-412-99391-0

Reviewed by *Randolph Chung*

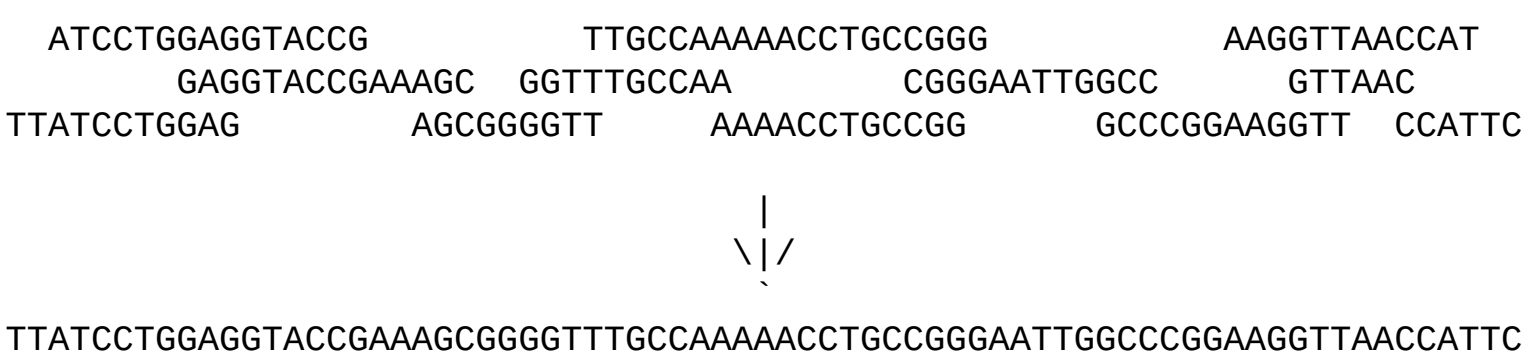
Biology and computer science used to be two distinct fields. However, with the recent rapid developments in molecular biology -- especially in the realms of DNA sequencing, restriction mapping and studies of protein structures -- biologists are turning to computer scientists for efficient methods to handle and analyze the vast amounts of data becoming available. A new and exciting field known as **Computational Biology** has subsequently evolved.

Waterman's book, Introduction to Computational Biology, offers a treatment of the mathematical structures of many problems biologists are trying to solve. After delineating these mathematical characterizations, Waterman presents several algorithms in use today for analyzing biological data.

To give an example of problems treated in the book, consider the immense efforts that are being put into sequencing the genome of bacteria, viruses, yeasts, human beings and other animals (e.g. in the Human Genome Project [1]). One of the things in which biologists are interested is how similar the genome of one organism is to another. Among other things, the degree of similarity gives us an idea of the relationship between organisms and their ancestors. It may also allow us to deduce the function of one gene based on its similarity to another gene whose function is known.

Many commericial applications can do these types of homology searches and comparisons. Among them, the BLAST system [2] is probably the most well known. Waterman presents a detailed mathematical treatment of the problem followed by a dynamic programming algorithm for attacking it. The mathematical treatment allows one to organize the problem so that an efficient solution is possible.

Another application that is discussed in the text is shotgun sequencing [3]. Shotgun sequencing is a method for sequencing DNA whereby several copies of the genome of an organism is first broken down into many small pieces (see figure). Each piece is sequenced individually. The problem then becomes: how does one piece together the sequenced fragments in order to uncover the entire genomic sequence?



Shotgun Sequencing

This can be phrased as the *shortest substring problem* (SSP): given a collection C of strings over an alphabet, what is the shortest string S such that each string in C is a substring of S?

SSP is known to be a NP-complete problem, which means that it is believed that no efficient (polynomial time) algorithms exists for solving SSP. After proving this fact, Waterman presents an approximation algorithm which always finds a sequence that is at most twice the length of the optimal string in polynomial time. These approximation algorithms are often very useful in real-life situations where the sheer volume of data to be analyzed does not permit the use of exponential-time algorithms.

Despite its title as an introductory text, Introduction to Computational Biology assumes a fair amount of background in statistics, calculus and an understanding of algorithms and their running time complexity. Although the author opens with a chapter in introductory molecular biology, additional background in this area is useful if one wants to get the most out of this text.

I recommend this book to anyone interested in learning how computer science can be applied to solve important problems in the field of molecular biology.

References:

1 Schuler, et al. *Science* **274**: 540-546

2 http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html

3 For an example, see Fleischmann, R. D. et al. *Science* **269**: 496-512