

# Subsequence Counting and Statistics

Jonathan B. Myers  
Virginia Tech University  
Robert W. Robinson \*  
University of Georgia

## Abstract

By a  $k$ -sequence is meant a sequence of length  $k$  over a fixed but arbitrary finite alphabet. An  $i$ -subsequence of a  $k$ -sequence is an  $i$ -sequence obtained by deleting any  $k-i$  elements. Equivalently, the  $k$ -sequence is a  $k$ -supersequence of the  $i$ -sequence in that case. Dynamic programming recurrences are derived for the number of different  $i$ -subsequences of any given sequence, and also for the number of different common  $k$ -supersequences of any two given sequences.

Let  $X(i,k)$  denote the random variable which gives the number of different  $i$ -subsequences of a  $k$ -sequence which is chosen according to the uniform measure, i.e., considering all  $k$ -sequences over the fixed alphabet to be equiprobable. The above recurrences appropriately summed lead to efficient algorithms for exactly calculating the mean and variance of  $X(i,k)$ . Numerical results are presented which indicate that as  $k$  grows, the standard deviation tends to be comparable in magnitude to the mean for all values of  $i$  which are not close to 0 or  $k$ .