UNIVERSITY OF SYDNEY

SCHOOL OF MATHEMATICS AND STATISTICS

Statistics Seminar

Friday, 14 March 2008, 2.00pm

Carslaw 375

Evaluation of Similarity Between Two Sequences

Susan Wilson
(Australian National University)

Abstract

The question "How should we measure and evaluate the similarity between two sequences?" is the focus of this presentation. A commonly encountered situation, particularly in biology, is to have a query sequence and want to find which, say DNA or protein, sequences in a large database have "significant" similarity to this query sequence. The widely accepted solution to the question is based on the notion of alignment, and the BLAST program is the most frequently used method. There are very many situations though for which the inherent assumption underlying the use of local alignment methods is violated. Hence there has been much interest in development of alignment-free sequence comparison algorithms. Arguably the best of these is the number of k-words shared between two sequences. The statistic, called D_2, is simple and extremely fast to compute. Its distribution and asymptotic properties are being studied, and recent results and unsolved problems will be overviewed.

Enquiries about the Statistics Seminar should be directed to
Jean Yang (jeany@maths.usyd.edu.au)