



Statistics Seminar Series

Session 2, 2009



Conrad Burden

Australian National University

Alignment-free sequence comparisons using k-word matches

A common problem faced by biologists is finding a close match in a database to a given DNA or protein sequence. This is used, for example, to identify homologous genes or proteins in a particular species, or to find genes related by a common ancestor in two different species. The most popular, currently available sequence matching algorithms attempt to align long sequences. This may not always be appropriate when related sequences have been rearranged or spliced, or when identifying short regulatory motifs. We are developing an alignment free method, called k-word matches. The idea is to use as a comparison statistic the number of exact or partial short word matches of a given pre-specified length. We have found accurate representations of the statistical properties of word match counts under suitable null hypotheses, and are developing fast computer algorithms for biological applications.

About the speaker: Dr. Conrad Burden is a Fellow at the Centre for Bioinformation Science Mathematical Sciences Institute and John Curtin School of Medical Research, ANU. He is interested in bioinformatics with a particular focus on the statistical analysis of oligonucleotide microarrays, on analysing genetic regulatory networks, on analysing protein structure, on alignment-free DNA sequence comparison using k-word matches and on phylogenetic reconciliation of gene and sequence trees.

Time: 4pm, Friday, 4th September

Location: Room RC4082

Seminar co-ordinator: Spiridon Penev

e-mail: S.Penev@unsw.edu.au