Statistics of gene clusters and gene families

David Sankoff
Département de mathématiques et de statistique
Université d'Ottawa
585 King Edward Ave.
Ottawa, Ontario
CANADA K1N 6N5

Abstract

Comparing chromosomal gene order in two or more related species is an important approach to studying the forces that guide genome organization and evolution. Linked clusters of similar genes found in related genomes are often used to support arguments of evolutionary relatedness or functional selection. However, as the gene order and the gene complement of sister genomes diverge progressively due to large scale rearrangements, horizontal gene transfer, gene duplication and gene loss, it becomes increasingly difficult to determine whether observed similarities in local genomic structure are indeed remnants of common ancestral gene order, or are merely coincidences.

A rigorous comparative genomics requires principled methods for distinguishing chance commonalities, within or between genomes, from genuine historical or functional relationships. In this lecture, we construct tests for significant groupings against null hypotheses of random gene order, taking incomplete clusters, multiple genomes and gene families into account. We consider both the significance of individual clusters of pre-specified genes, and the overall degree of clustering in whole genomes.