AUTOMATING THE SEARCH FOR LATERAL GENE TRANSFER

MARK A. RAGAN

Institute for Molecular Bioscience, University of Queensland and
Australian Research Council (ARC) Centre in Bioinformatics
St Lucia, Q 4072, Australia

Most genes have attained their observed distribution among genomes by transmission from parent to offspring through time. In prokaryotes (bacteria and archaea), however, some genes are where they are as the result of transfer from an unrelated lineage. To elucidate the biological origins and functional consequences of lateral gene transfer (LGT), we have constructed an automated computational pipeline to recognise protein families among prokaryotic genomes, generate high-quality multiple sequence alignments of orthologs, infer statistically sound phylogenetic trees, and find topologically incongruent subtrees (prima facie instances of LGT). This pipeline requires that we automate workflows, design and optimize algorithms, mobilise high-performance computing resources, and efficiently manage federated data. I will summarise results from the automated comparison of 422971 proteins in 22437 families across 144 sequenced prokaryotic genomes, including the nature and extent of LGT among these lineages, major donors and recipients, the biochemical pathways and physiological functions most affected, and implications for the role of LGT in evolution of biochemical pathways.