Frequency analysis of sequence data at low coverage assembly

Computational sequence alignment can be utilized to compare closely related genomes at the nucleotide level. However, on the evolutionary timescale, the genomes of different species can diverge significantly due to frequent nucleotide insertions, deletions, and rearrangements. This inherent genetic variability makes straightforward sequence alignment nearly impossible, so genomes are instead typically compared via the absence/presence of a library of common genes. However, in the absence of easily identifiable common genes, there are few available computational methods capable of comparing different whole genome sequences. In this invention, the analysis of the frequencies of nucleotides or amino acids in a genome provides an alternative and phylogenetically deeper way of understanding and identifying relations between and within organisms.

Computational method identifies deep phylogenetic relationships

This invention establishes the method to extract frequencies of words from the genomes of organisms. It also creates the measures to evaluate the likelihood that a sequence is derived from another one, where alternative “distances” are designed and implemented. The same strategy can be applied to identify specific genes within a genome and produce genome profiles that indicate how the distribution of a section of the genome is related to the rest. These genomic profiles reflect very deep phylogenetic relationships that cannot be elucidated by traditional sequence alignment and phylogenetic techniques.

Lead Inventor:

Raul Rabdan, Ph.D.

Applications:

• To identify and classify organisms that have a very distant relationship to other previously sequenced organisms
• New Viruses: to analyze the family, subfamily, genus, and species of emerging viruses and other new pathogens
• Low Coverage Assembly: to identify and cluster fragments of genomes to reflect that these fragments originate from the same organisms. This allows identifying and “assembling” organisms with coverage less than 1
• The grouping of divergent genes, proteins, and related organisms based on the frequency analysis of their subsequences. A particular application is the identification of reads from sequence data that are coming from a gene
• Drug discovery: to aid in the discovery and development of drugs and antiviral agents

Advantages:

• The genome profiles generated by this method reveal phylogenetic relationships that cannot be elucidated by standard sequence alignment techniques
• The technology is especially suitable for identifying emerging viruses

Patent Information:
Patent Pending
Tech Ventures Reference: IR 2546

Inventors
Raul Rabadan