Integrative Biology-2013
OMICS Group Conferences
Accelerating Scientific Discovery
International Conference on
Integrative Biology Summit
August 5-7, 2013 Embassy Suites Las Vegas, USA

A comparative study of gene similarity measures

Anurag Nagar

Southern Methodist University, USA

Abstract

Genes are essential functional, regulatory and hereditary units present in all living organisms. They are identified by contiguous stretches of DNA present in chromosomes and carry the codes for a polypeptide or RNA chain that has a particular purpose. Genes also code for protein synthesis, which is one of the most important functions of the cells. Genes have been studied from various perspectives – such as identifying their DNA sequence and location on chromosomes, identifying their purpose and functions, studying evolution over time, and finding their similarity across multiple species. One of the most important repositories of gene related information is the Gene Ontology (GO) project. This project was started with the aim of providing a standard way of storing gene related information for multiple species and products. Gene Ontology contains three types of gene ontologies that describe biological terms and processes – Biological Processes (BP), Molecular Function (MF) and Cellular Components. These terms are organized in the form a directed acyclic graph with the terms getting more specific as we travel down the graph. Various genome projects such as the yeast genome project and the human genome project annotate species-specific genes with the terms of the GO. One of the important applications of GO is to compute similarities between genes – either from the same species or different ones. Various semantic similarity measures have been proposed so far. Some of them use the Information Content of the terms, such as that proposed by Resnik et al while others use the distances between edges of the terms, such as that proposed by Nagar et al. Researchers have tried to improve upon them continuously, for example the IntelliGO measure proposed by Benabderrahmane et al. The above similarity measures use different approaches so their results are different for the same pair of genes. With the development of next generation sequencing technologies, it has become possible to identify the exact location and sequence of the genes. It is thus possible to compute the sequence similarity between two genes and see how this correlates to their semantic similarity computed by using GO. Some amount of work in this direction has been done by P.W. Lord et al in their paper "*Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation*". However, this was in 2002 and they used only one GO-based similarity measure. Since then many new methods have been proposed and it is interesting to re-visit this problem and compare their similarities. In this talk, I will present a background of GO, review the existing semantic similarity measures and compare them with the sequence similarity measures obtained by using sequence comparison methods such as Clustal. The results show that there is still a need for better semantic similarity measures that utilize the structure of GO terms and the information content in them.