ISSN: 2153-0602
Journal of Data Mining in Genomics & Proteomics

OPEN ACCESS Freely available online

Commentary

Genome Annotation and its Stepwise Approach

Cui Lachlan*

Department of Microbiology, Nanjing Agricultural University, Nanjing, China

DESCRIPTION

Genome annotation is the process of identifying functional elements along the sequence of the genome, thereby giving meaning to the genome. DNA sequencing is necessary because it produces sequences of unknown function. Over the last three decades, genomic annotations have been made up of thousands of single nucleotides, from computational annotations of long protein-coding genes on individual genomes (one per species) and experimental annotations of short regulatory elements on a small number of genomes. It has evolved into a collective annotation. This increased resolution and inclusiveness of genomic annotations (genotype to phenotype) leads to accurate insights into the biology of species, populations, and individuals. The main goal of annotation is to describe these sequences and ultimately determine how universal these sequences are to the promoter of a particular gene. The first step is to write such an array in the reference species and use this information for further comparative analysis. Gene annotations are the analysis and interpretation needed to capture the raw DNA sequences generated by the genome sequencing project, extract biologically important information, and place such derived details in context. An easy method of gene annotation is to search for homologous genes in a specific database using a homology-based search tool such as BLAST. The information obtained is used to annotate genes and genomes. Structural annotations consist of the identification of genomic elements. It consists of three main steps: identifying parts of the genome that do not encode proteins, identifying elements on the genome, a process called gene prediction, and attaching biological information to those elements.

Genome annotation consists of explaining the function of the predicted gene product by a silico approach. It has specific functions including (1) signalling sensors such as TATA boxes, start and stop codons, or polyA signalling and (2) content sensors (such as G + C content, codon frequency, or dicodon abundance). This can be achieved using the bioinformatics software provided detection and (3) Similarity detection e.g., proteins from closely related organisms, mRNAs from the same organism, or between reference genomes. However, methods for predicting genes and genomic structures (tRNA, rRNA, promoter regions, etc.) are linked to the assembly strategy and sequencing platform used.

Genome annotations can be divided into three basic categories. The first is a nucleotide-level annotation that seeks to pinpoint the physical location of DNA sequences and where components such as genes, RNA, and repeat elements are located. Sequencing and/or assembly errors at this stage can result in false pseudogenes by Indel. The second is protein-level annotations that seek to determine and identify possible functions of genes that a particular organism may or may not have. The third is process-level annotations aimed at identifying pathways and processes in which different genes interact in order to construct efficient functional annotations. At the last two levels, sequencing and/or assembly errors reduce similarity and can compromise the inference of true genetic function.

Stepwise approach to gnome annotation

Advances in sequencing technology are accelerating the generation of genomic data. Access to these genomic resources allows scientists to compare genomes in a variety of biomedical, ecological, and/or geographical situations. The first step in these comparisons is genomic annotation, where biological information is extracted from raw nucleotide sequence data. Many gene prediction tools are used to predict the location of genes. The predicted gene is used in a sequence similarity search against the database to assign cellular function to the gene product. Annotations are enhanced for the biological context by integrating information at the level of regulatory and metabolic networks and protein-protein interactions.

CONCLUSION

Prior to genome annotation, there is a process of genome assembly using the reference genome-based method or the *de novo* approach. Gene discovery tools; Prodigal, GeneMark, Metagene Annotator, etc. are used to identify Open Reading Frames (ORFs) of genomic sequences. These ORFs are BLAST searched against databases such as GENBANK and UniProt to identify estimation function and protein evidence. The ORF is mapped to the pathway using the KEGG database. Protein domains are identified by an InterProScan search. In this search, each protein domain is assigned GO terms, and these features will be used later to perform enhanced analysis. The ORF is searched against the stored domain database containing the COG to identify the corresponding ortholog.

Correspondence to: Cui Lachlan, Department of Microbiology, Nanjing Agricultural University, Nanjing, China, E-mail: CuiLachlan@yahoo.com

Received: 10-Jan-2022, Manuscript No. JDMGP-22-15245;**Editor assigned:** 12-Jan-2022, PreQC No.JDMGP-22-15245 (PQ);**Reviewed:** 24-Jan-2022, QC No. JDMGP-22-15245;**Revised:** 27-Jan-2022, Manuscript No. JDMGP-22-15245 (R); **Published:**31-Jan-2022, DOI: 10.4172/2153-0602.22.13.245.

Citation: Lachlan C (2022) Genome Annotation and its Stepwise Approach. J Data Mining Genomics Proteomics. 13:245.