BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Young BITS-RSG Symposium
Title	GBRAP: A Comprehensive Tool and a Database to Measure Genomic Characteristics
All Authors	Sachithra Kalhari Yaddehige(1), Cristian Taccioli(1)
Affiliation	
(1) Department of Animal Medicine, Production and Health, University of Padova, Italy	

Motivation

In the field of genomics, understanding the evolutionary relationships across different life kingdoms requires an extensive examination of genomic data. The vast array of genomes available through the National Center for Biotechnology Information (NCBI) offers a unique opportunity to explore these relationships. By downloading thousands of genomes from all domains of life, our research aims to use detailed genomic information to uncover evolutionary patterns. Despite the availability of a huge number of genomes, the effective visualization of the information these genomes contain is limited due to their size. This necessitates the development of advanced tools and databases capable of handling and interpreting complex genomic data.

Background

In the pursuit of unravelling evolutionary dynamics of different species through genomic scrutiny, we developed GBRAP (Genome-Based Retrieval and Analysis Parser), a pioneering software proficient in parsing .gbff files (Genome Bank Flat Files). With GBRAP, we have successfully downloaded, analysed, and summarized the genomic content of all organisms available in the Reference Sequence Database (RefSeq), ranging from the smallest viral genomes to the largest plant and animal genomes. Here, we present the GBRAP database, an online repository housing data curated by GBRAP. This repository, with easily accessible data, yields invaluable insights into genomic diversity, complexity, and base composition parity. Offering a quantitative, multifaceted perspective on genomic content, GBRAP facilitates a more profound understanding of the genome architecture of organisms.

Methods

Programmed in Python 3 and operated through command line, GBRAP boasts the ability to fetch gbff files from RefSeq which may contain one or multiple sequences, and subsequently produce a multitude of genomic statistics. Its sole reliance on standard Python libraries eliminates the need for any additional installations. The script discerningly filters out alternate loci and scaffolds from .gbff files, ensuring data extraction solely from complete chromosomal and mitochondrial sequences, thus preserving data integrity. Furthermore, we have meticulously curated several filtering mechanisms within the script to prevent data redundancy. For instance, the 'cds_selector' function of GBRAP selectively identifies a singular CDS isoform per gene. The function operates by first removing CDS sequences that do not start with a start codon, then eliminating isoforms that are not composed entirely of triplets of bases. If multiple isoforms still persist, the function selects the longest among them. The GBRAP database, constructed using MySQL database management system, features a user-friendly website developed with HTML and PHP.

Results

Output files from GBRAP's current version encompass over 200 columns of data, categorized into various genomic features, and distributed across rows dedicated to individual chromosomes. These data encompass a range of genomic information metrics, not limited to base counts and frequencies but also GC content, Codon usage, Shannon entropy, Chargaff score and more. To enhance scrutiny, these quantitative metrics are also segregated by genomic features such as non-coding RNAs (ncRNAs), CDS, Introns, and transposable elements etc. as well as by the whole chromosome.

The online database currently provides two methods for data retrieval. Users can either search by typing the organism’s name, or navigate pre-categorized zoological classes (Mammals, Birds, Plants, Bacteria etc.) to access organism listings, which can subsequently lead to organism-specific data tables, facilitating targeted inquiries into desired genomic features. For instance, to ascertain the GC content in CDS, or introns of a chicken, users can navigate via Birds -> Gallus Gallus, and locate the GC_cds, and GC_cds_intron columns in the data table. While presently all data columns for an organism are displayed, forthcoming updates of the database will empower users to selectively filter data columns or rows as per their requirements.

Conclusion

The creation and implementation of the GBRAP tool and Database represent a significant breakthrough in the field of genomics. Prior to this, no existing tool had the capability to extract such an extensive range of genomic information from the entirety of the NCBI genome repository. Moreover, the GBRAP database is unprecedented in its comprehensive inclusion of genomic data from a diverse array of organisms, ranging from viruses and bacteria to primates. This inclusivity allows for an unparalleled overview of genomic diversity and complexity across all domains of life, providing a unique resource for researchers aiming to understand evolutionary trends and relationships. Looking ahead, we aim to leverage GBRAP-generated data to discern evolutionary patterns among various species, thus charting new frontiers in genomic exploration.

Info	
-	
filename	-
Figure	
-	
Availability	http://www.bioinformatics.maps.unipd.it/gbrap/
Dissemination Material	
Social	
-	
Summary	
-	
Corresponding Author	
Name, Surname	Sachithra Kalhari, Yaddehige
Email	sachithrakalhari.yaddehige@phd.unipd.it
Submitted on	23.04.2024