

# CpGDB : A Comprehensive Database of Chloroplast Genomes

Bhupinder Pal Singh<sup>1,2</sup>, Ajay Kumar<sup>3</sup>, Harpreet Kaur<sup>4</sup>, Harpreet Singh<sup>5</sup>, Avinash Kaur Nagpal<sup>4\*</sup>

<sup>1</sup>I.K.Gujral Punjab Technical University, Kapurthala, 144 603, India; <sup>2</sup>Centre for I.T. Solutions, Guru Nanak Dev University, Amritsar, 143005, India; <sup>3</sup>Electronics and Communication Engineering Department, Beant College of Engineering and Technology, Gurdaspur, 143521, India; <sup>4</sup>Department of Botanical and Environmental Sciences, Guru Nanak Dev University, Amritsar, 143005, India; <sup>5</sup>Department of Bioinformatics, Hans Raj Mahila Maha Vidyalaya, Jalandhar, 144008, India. Bhupinder Pal Singh-E-mail: bhupinderpalsingh@hotmail.com; Ajay Kumar - E-mail: ajaykm\_20@yahoo.co.in; Harpreet Kaur - E-mail: preet.botany@gmail.com; Harpreet Singh - E-mail: harpreetsingh05@gmail.com; Avinash Kaur Nagpal - E-mail: avnagpal@yahoo.co.in; \*Corresponding author

Received January 30, 2020; Accepted February 20, 2020; Published February 29, 2020

DOI: 10.6026/97320630016171

## Declaration on official E-mail:

The corresponding author declares that official e-mail from their institution is not available for all authors

## Declaration on Publication Ethics:

The authors state that they adhere with COPE guidelines on publishing ethics as described elsewhere at <https://publicationethics.org/>. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

## Abstract:

Chloroplast Genome Database (CpGDB) is user friendly, web-based, freely available and dynamic relational database which provides a platform for researchers to search and download complete chloroplast genome sequences, individual gene sequences and feature records of plant species belonging to same or different families of spermatophytes. Presently, the database consists of genome sequences, individual gene sequences and feature records of chloroplast genomes of 3823 plant species belonging to 1527 genera from 256 families, which will be updated regularly with the availability of new sequences at NCBI. Extensive data mining of feature records from GenBank files, uniform nomenclature for majority of genes, enriched intron/exon feature records makes CpGDB a valuable resource for studies in chloroplast genomics while complementing existing chloroplast databases.

**Availability:** <http://www.gndu.ac.in/CpGDB>

**Keywords:** Chloroplast Genomes, Gene Sequences, Database.

## Background:

Chloroplasts are photosynthetic intracellular organelles of plants and are considered as Earth's main solar energy converters. Besides photosynthesis, chloroplasts are also involved in synthesis of amino

acids, vitamins, lipids, pigments, precursors of plant hormones etc. They contain their own distinct genome, which is known to be derived from a cyanobacterial ancestor. Among three genomes of a plant cell, chloroplast genome is the most gene dense constituting

more than 100 genes in genome size of 120 to 210 kb [1]. The chloroplast genome generally has a highly conserved organization in term of size, structure and gene content [2]. The presence of a pair of inverted repeats (IRs) is one of the outstanding features of chloroplast genome, which separates two single copy DNA regions, a large single-copy (LSC) and a small single-copy (SSC) region, on a single circular DNA molecule [3].

Recent advances in sequencing techniques have led to a significant increase in the availability of large number of chloroplast genome sequences. Currently, about 4000 eukaryotic chloroplast genomes have been sequenced completely with the best representation from flowering plants [4]. During the past two decades, few attempts have been made to develop specialized databases for chloroplast genomes such as GOBASE (Organelle Genome Database) [5], ChloroplastDB (Chloroplast Genome Database) [6], Database of PCR Primers of Chloroplast Genomes [7] and PCIR (Database of Plant Chloroplast Inverted Repeats) [8]. However, these databases are either non-functional or very much focused to cover only specific aspect of genomic information. The Organelle Genome Database available at NCBI does give information on chloroplast genomes of different plant species but it is more generalized database where annotations for some gene names and features are not uniform. Moreover, these databases does not allow family wise search of chloroplast genomes thereby necessitating the development of a specialized database of chloroplast genomes with enhanced user friendliness.

Chloroplast Genome Database (CpGDB) is an attempt to organize and integrate enormous amount of data related to chloroplast genomes of spermatophytes. It has two levels of information: NCBI based sequence data and curated annotations. Its main focus is to make available complete chloroplast genome sequences and their different features e.g. gene, CDS, rRNA, tRNA, intron, exon etc. for different types of analysis. This is a dynamic relational database available at <http://www.gndu.ac.in/CpGDB> and supports different queries specifically with respect to comparative genome analysis across different species/genera/families. It allows search by family, genus and gene names along with provision to download complete chloroplast genome sequences or individual gene sequences for selected plant species belonging to same or different families. Additionally, it provides unified annotations with respect to gene names. This database would be a valuable platform for researchers in the field of chloroplast genomics.

## Methodology:

### Data mining and organization

The CpGDB was designed using MS-Visual Studio 2015 as front end and SQL Server 2012 as back end (Figure 1). A web interface was developed to import complete chloroplast genome sequences belonging to spermatophytes directly into CpGDB from NCBI RefSeq nucleotide database [7] using Entrez Programming Utilities (E-Utilities). The Entrez text query “(Chloroplast [Title] OR Plastid[Title]) AND complete[Title] AND genome[Title] AND refseq[filter] AND plants[filter] AND Spermatophyta[Organism]” was passed as web request to NCBI nucleotide database to download sequences both in FASTA and GenBank format and store in local repository for providing faster and efficient sequence retrieval system.

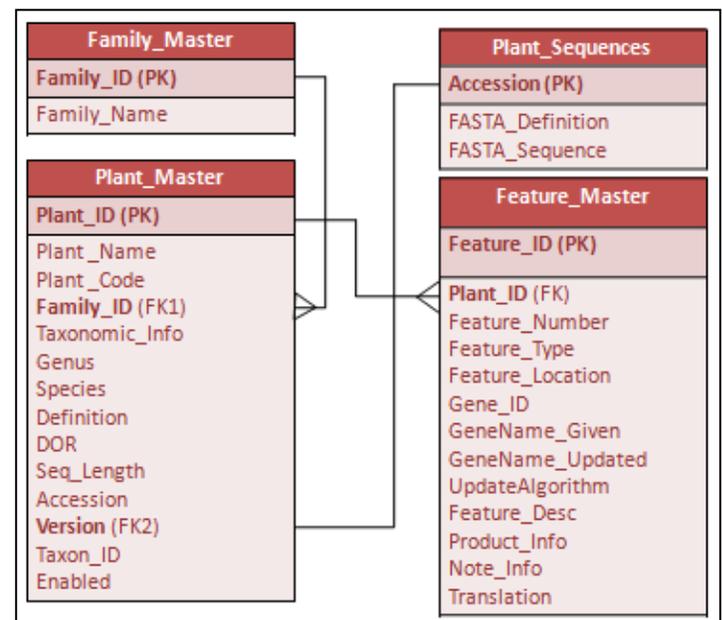
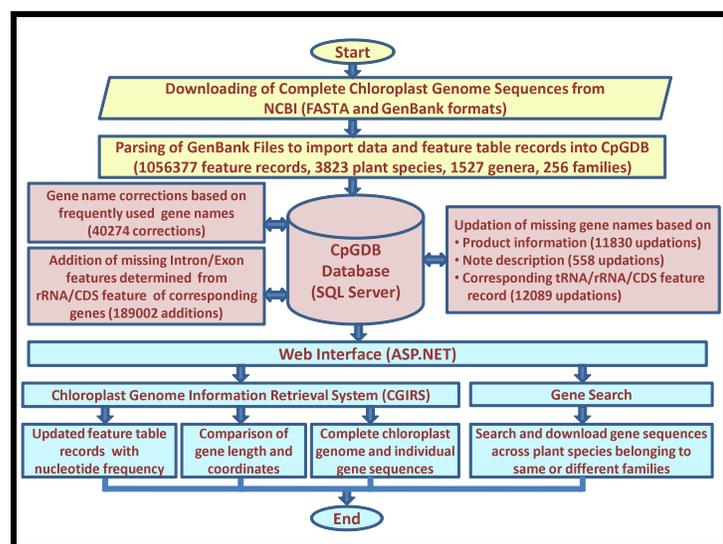


Figure 1: Entity-Relationship diagram of CpGDB.

The GenBank files were parsed using parsing script written in ASP.NET and database tables Family\_Master, Plant\_Master and Plant\_Sequences were updated for each accession. In order to avoid duplicate occurrences of chloroplast genome sequences due to availability of more than one accession for some of the plant species, the most recently modified versions were considered. Also accessions containing hybrid plant species or with missing species name were not considered. Hence, out of total 3881 accessions available upto 30/11/2019, 3823 were considered for further

analysis (**Supplementary Table S1 - see excel file**). These accessions belong to 1527 different genera from 256 plant families. Each plant species was assigned a unique four letter code to make data retrieval, analysis and comparison simple and more convenient. An algorithm was developed to extract and export feature records of plant species from GenBank files. The data related to gene name, geneID, location, product and note information etc. was extracted and stored separately for each plant species. The execution of this algorithm resulted in compilation of 10,56,377 feature records belonging to 3823 plant species. Any missing information was indicated with '-' value in the record.



**Figure 2:** Complete workflow for creation of CpGDB.

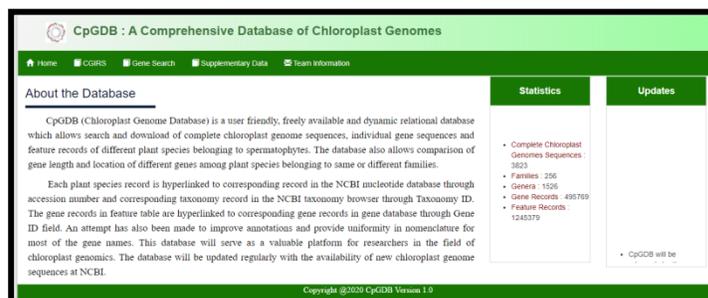
### Data cleaning:

It is important to mention here that gene names in most of the downloaded GenBank files were not uniform. For example 16S ribosomal RNA was mentioned as *rrn16*, *16rrn*, *16s*, *16S* ribosomal RNA, *16S rRNA*, *16S-rRNA* etc. among various chloroplast genomes. To bring uniformity in gene names, an algorithm was developed to determine frequency of different gene names in the database. Out of 781 gene names, 461 gene names were replaced with most frequently used gene names. For example gene names for 16S ribosomal RNA were updated with most frequently used gene name '*rrn16*'. In total, 40,274 corrections were made across 979 plant species and a record of changes was also maintained (**Supplementary Table S2 - see excel file**).

A total of 11,830 feature records across 1160 plant species were updated based on their product information given in the feature description due to nonavailability of gene names in the feature records (**Supplementary Table S3 - see excel file**). Similarly, 558 records spanning 115 plant species with missing gene name as well as product information were updated using note description of the corresponding record (**Supplementary Table S4 - see excel file**). Finally, 12,089 records belonging to 1171 plant species with missing gene name, product and note description were updated from the corresponding tRNA/rRNA/CDS feature records (**Supplementary Table S5 - see excel file**). To enrich information about insufficiently annotated split genes, 1,89,002 missing exon/intron features belonging to 3802 plant species were determined using tRNA/CDS features of the corresponding genes and added in the database (**Supplementary Table S6 - see excel file**). All these manual curations resulted in 12,45,379 feature records available in the database. **Figure 2** shows the complete workflow along with data for creation of CpGDB.

### Database Interface and Utility:

The database can be accessed easily through Chloroplast Genome Information Retrieval System (CGIRS) link available on the top menu bar of home page of CpGDB (**Figure 3**).



**Figure 3:** Screen shot showing home page of CpGDB

CGIRS provides a tabular view of families along with list of genera and corresponding species whose chloroplast genome records are available in the database. These records can be filtered and sorted out as per number of genera and number of species. The web interface allows selection of one or more families and/or plant species using check boxes provided against each record. The web interface also provides updated feature records of individual plant species along with the pre-determined nucleotide frequency. The feature records can be filtered based on feature type (All features, Gene, CDS, tRNA, rRNA etc.). The plant species, taxonomy and gene records are hyperlinked to corresponding records in the

nucleotide core [9] taxonomy [10] and gene [11] databases, respectively at NCBI. Specific details are displayed on hovering mouse over the plant species name and feature type records. Feature table data as well as chloroplast genome sequences can be downloaded for further analysis. Additionally, user can select multiple plant species and compare lengths and coordinates of all genes across selected plant species. 'Gene Search' link gives an option to select a feature type corresponding to a given gene and download the related sequences or records for selected species.

#### Conclusion:

CpGDB provides curated information on complete chloroplast genome sequences available upto 30/11/2019. The database will be regularly updated with the availability of new sequences. The strength of the database lies in its user friendliness, ability to retrieve specific data and uniform nomenclature of majority of genes. CpGDB will serve as a valuable resource for chloroplast genomic studies.

#### Supplementary Data:

Supplementary files are available for downloading at journal website and CpGDB website.

#### References:

- [1] Wakasugi T *et al. Photosynth. Res.* 2001 **70**: 107 [PMID: 16228365].
- [2] Fu J *et al. PLoS One.* 2016 **11**: e0159566 [PMID: 27442423].
- [3] Sugiura M *et al. Annu. Rev. Genet.* 1998 **32**: 437. [PMID: 9928487].
- [4] <http://www.ncbi.nlm.nih.gov/genomes>
- [5] O'Brien EA *et al. Nucleic Acids Res.* 2009 **37**:D946. [PMID: 18953030].
- [6] Cui L *et al. Nucleic Acids Res.* 2006 **34**:D692. [PMID: 16381961].
- [7] Heinze B *Plant Methods* 2007 **3**:4 [PMID: 17326828].
- [8] Zhang R *et al. Database.* 2019 Volume 2019: 1-7 [PMID: 31696928].
- [9] <https://www.ncbi.nlm.nih.gov/nuccore>
- [10] <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>
- [11] <https://www.ncbi.nlm.nih.gov/gene>

Edited by P Kanguane

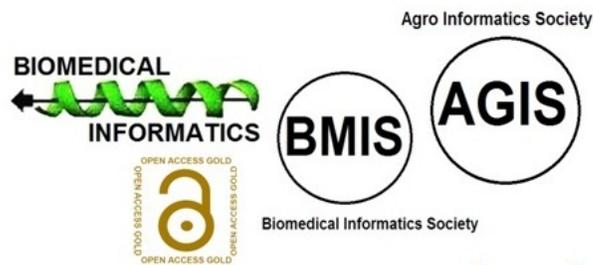
Citation: Singh *et al. Bioinformation* 16(2): 171-175 (2020)

**License statement:** This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License

Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article for FREE of cost without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

# BIOINFORMATION

*Discovery at the interface of physical and biological sciences*



*since 2005*

# BIOINFORMATION

*Discovery at the interface of physical and biological sciences*

*indexed in*

