

# GGRP: GenBank Genome Record Parser

Kavita Yadav<sup>1</sup>, Sachin Pundhir<sup>2</sup>, Tiratha Raj Singh<sup>1\*</sup> and Anil Kumar<sup>1</sup>

<sup>1</sup>Bioinformatics Sub-Centre, School of Biotechnology, Devi Ahilya University, Khandwa Road, Indore-452001, INDIA

<sup>2</sup>Bioinformatics Sub-Centre, NIPGR, Aruna Asaf Ali Marg, New Delhi – 110067, INDIA

\*tiratharaj@gmail.com

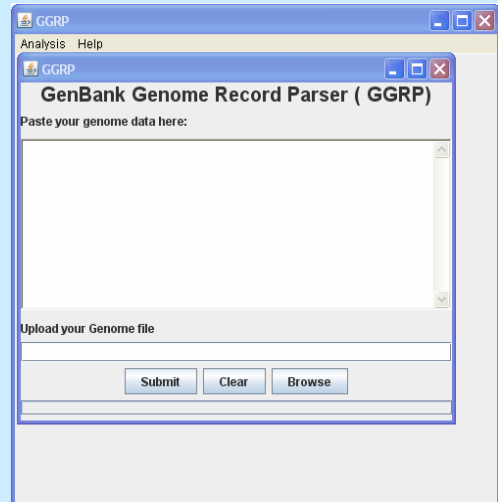
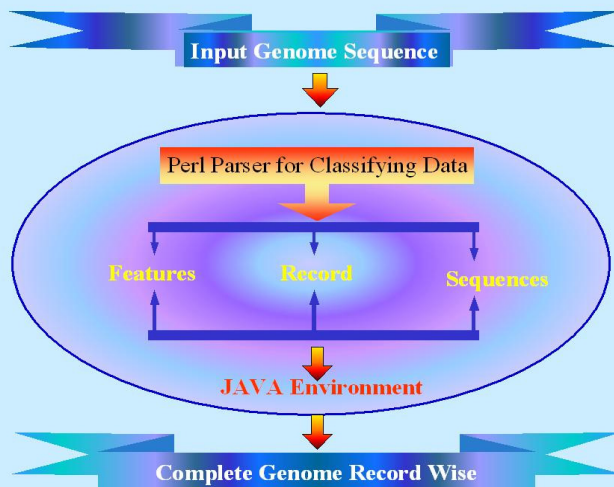
## Introduction

Browsing genomic information has been central to life science analyses since large scale genomes became available for myriad of species. With the advent of recent high throughput DNA sequencers, huge amount of genomic data has been produced, requiring the data analysis to be done as quickly as possible. NCBI completed the transition of its main genome annotation database from Locus link to Entrez Gene in spring 2005. However, there is need to parse a single genome according to user's requirements. Here We developed a parser i.e., GenBank Genome Record Parser (GGRP) for GenBank genome record to parse the complete genome record wise. In this Software input will be any complete genome record of GenBank and the software will parse the genome record wise. All the required information for a particular record will be displayed in one screen-shot. It will parse whole genome according to GeneID and will display respective information. GGRP has been developed in Java. Some additional Perl scripts have also been used to parse the genome. Parser have been generalized in way that it will work for any GenBank genome entry i.e. nuclear or mitochondrial.

## Objectives

- To develop a computational tool to parse the GenBank genome in its fragments.
- To visualize and navigate all the records in some user friendly sequence.
- To provide all the annotations for each record in one screen shot

## Methodology



## Results and Discussion

HFADR:		Date of Retrieval:	13-02-2008
Locus Tag:	NC_000907	Definition:	Haemophilus influenzae R4 KW20, complete genome.
Length:	1930138 bp	Accession:	NC_000907
Molecule Type:	DNA	Version:	NC_000907.1
Molecule Form:	circular	GC:	1327187%
division:	BCT		
FEATURE:			
Gene Type:	CDS	Gene ID:	950899
Gene Region:	? 1071	Coding Region:	? 1071
Gene:	gapDH	GI:	13271877
PID:	NP_438174.1	LID:	H0001
EC No:		Note:	1 of 3-phospho-D-glyceryl phosphate from D-glyceraldehyde 3-phosphate
Function:		Product:	glyceraldehyde 3 phosphate dehydrogenase
Protein:		Nucleotide:	

Miscellaneous:

59.101	iron use [Carbohydrate transport and metabolism], Region: GapA, COG0057
59.45	flavate dehydrogenase, NAD-binding domain, Region: Gap_4h_V, pfam00044
452.95	sphate dehydrogenase, C-terminal domain, Region: Gap_4h_C, pfam02800

GGRP has been implemented to facilitate the parsing of GenBank genome record entries in to separate fragments with all annotation to be displayed in one screen shot. User can navigate all the records of a single genome entry. GGRP allow sequence annotations through a three-tier architecture where input genome is first scanned and classified through Perl Script. Output is then entered into Java Environment where object-oriented class architecture has been implemented through Java Swing to manage classified data. Finally Java environment will generate a screen-shot for each record with inbuilt navigation facilities.

GGRP accommodate record-specific data formats and therefore allowed users to navigate all the sub-records and their respective annotations for a complete genome record.

All the major as well as minor classification units from header, features and miscellaneous sections has been included in the final output to provide the user fully classified and annotated record entries.

## Conclusions and Future Prospects

GGRP provides a user friendly parsing of complete GenBank genome record entry. It will help in the fragmented analysis of genome record and will be helpful in functional genomics and other related analysis. We are looking to enhance GGRP in near future by providing more flexible query interface and converting it into a web server as well.

## Acknowledgment

The authors acknowledge Amey Kekre & Dheeraj S. Panwar for technical support and the Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi for funding this research.

Availability: <http://www.davvbiotech.res.in/GGRP/>