Journal of Advanced Bioinformatics Applications and Research
ISSN 0976-2604.Online ISSN 2278–6007
Vol 5, Issue3, 2014, pp206-209
http://www.bipublication.com

Retrieval and Statistical Analysis of Genbank Data (RASA-GD)

Anop Singh Ranawat, Mohan Kumar Yadav,
Anoop Kumar Vaishya and Sumit Govil
School of Life Sciences, main campus,
Jaipur National University, Jaipur, Rajasthan, India

[Received 15/11/2014, Accepted-30/11/2014]

ABSTRACT

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. It comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI [1].The information retrieved in a Genbank file is complex and requires considerable time for analyzing the data. In this work we wish to convert the data present in Genbank file into a more readable, graphical format by developing a computational tool RASA-GD, using Perl code, Perl module and php. The tool accept file in Genbank format and will extract information such as coding region, exon, intron, gene, mRNA, tRNA and other data contained in gene bank file and also calculate the length of these elements and will generate a graphical output of the length and statistical values such as mean, median, mode, standard deviation, variance, sample range, minimum value, maximum value, unique gene number. This tool will be very useful for various kinds of analysis to both molecular biologists and computational biologists and will help in hypothesis generation. We have used our tool for visualizing variation of foxp2 gene from various species such as in human, monkey, gorilla, house mouse and gallusgallus. foxp2 [3] gene is responsible for speech behavior. In order to explore whether there is an evolutionary significance of the length of various genetic elements of this gene, The data generated has provided useful insight into the evolutionary history of this gene.

Keywords: GI number, Genbank, foxp2 gene, RASA-GD, DDBJ

INTRODUCTION

Bioinformatics is an interdisciplinary scientific field where biological problems can be solve easily and efficiently by using computational approach [2].It plays a very important role in the field of medical research, now it is the time of genomic era so there is a need of computers with biology merge approach which can be used to deal with the genomic data, to study and analyze the variation in gene. Nowadays, experimental data's are stored in databases to study and analyze the result computationally. Various computational tools have been designed to analyze the data and predict the result by applying mathematical and statistical approach to perform the experiment Insilco, using computational tools and database. The handling of databases and selection of desirable data from the databases is a tough task for the researchers and students from a non bioinformatics background. To make the retrieval work easier we have designed a tool

RASA-GD, which helps to retrieve the desire data from the complex file obtained from the database of Genbank.

METHODS AND RESULT

We have designed this tool by using perlcode, Perl module and PHP. The tool used statistical and graphical module to calculate statistical value and visualize the desire output graphically. For study we have taken six samples files from Genbank database (www.genbank.org) of foxp2 gene in different organisms such as human GI: 568815591, chimpanzee GI: 319999821, monkey GI: 109156890, gorilla GI: 401623081, house mouse GI: 372099104, gallusgallusGI: 358485511. We have putted all Genbank sample files in single Genbank file and browse in tool.

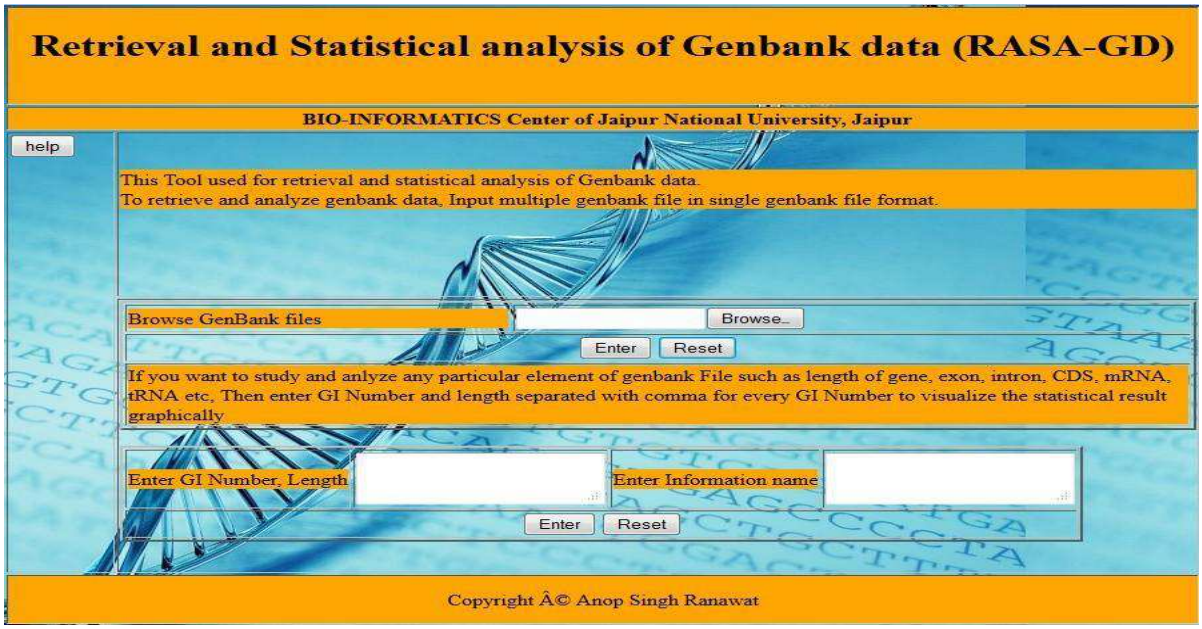
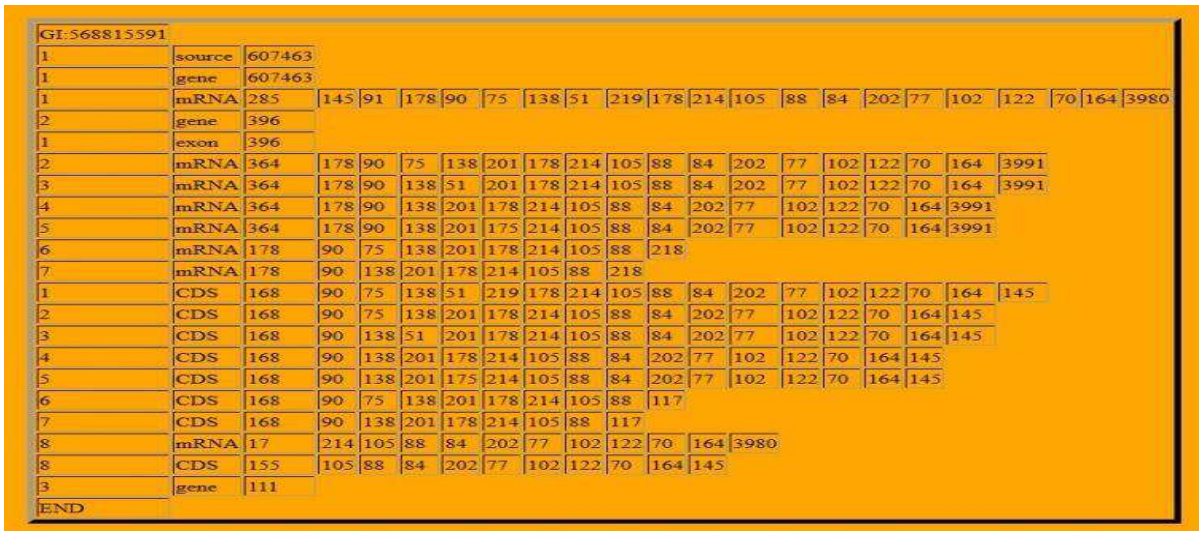
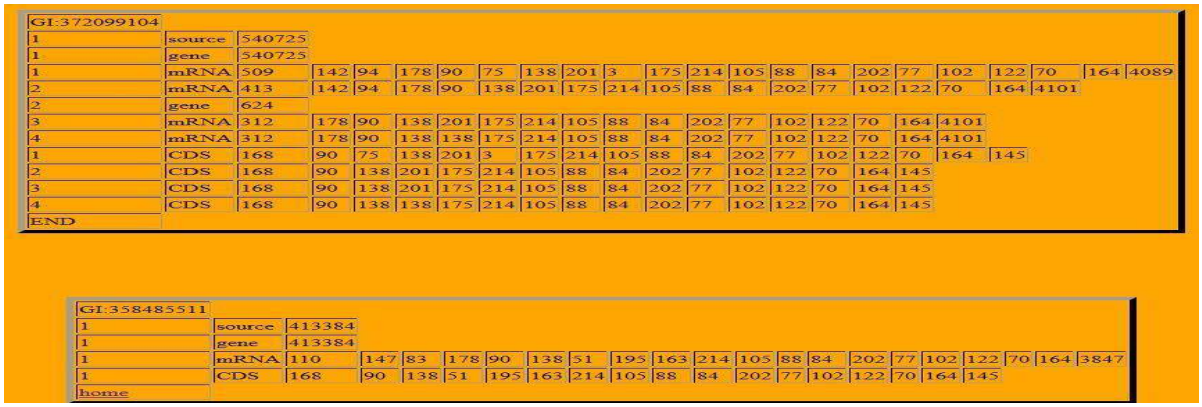


Fig.1: This is home page of RASA-GD tool



GI:568815591																					
1	source	607463																			
1	gene	607463																			
1	mRNA	285	145	91	178	90	75	138	51	219	178	214	105	88	84	202	77	102	122	70	164	3980
2	gene	396																			
1	exon	396																			
2	mRNA	364	178	90	75	138	201	178	214	105	88	84	202	77	102	122	70	164	3991		
3	mRNA	364	178	90	138	51	201	178	214	105	88	84	202	77	102	122	70	164	3991		
4	mRNA	364	178	90	138	201	178	214	105	88	84	202	77	102	122	70	164	3991			
5	mRNA	364	178	90	138	201	175	214	105	88	84	202	77	102	122	70	164	3991			
6	mRNA	178	90	75	138	201	178	214	105	88	218										
7	mRNA	178	90	138	201	178	214	105	88	218											
1	CDS	168	90	75	138	51	219	178	214	105	88	84	202	77	102	122	70	164	145		
2	CDS	168	90	75	138	201	178	214	105	88	84	202	77	102	122	70	164	145			
3	CDS	168	90	138	51	201	178	214	105	88	84	202	77	102	122	70	164	145			
4	CDS	168	90	138	201	178	214	105	88	84	202	77	102	122	70	164	145				
5	CDS	168	90	138	201	175	214	105	88	84	202	77	102	122	70	164	145				
6	CDS	168	90	75	138	201	178	214	105	88	117										
7	CDS	168	90	138	201	178	214	105	88	117											
8	mRNA	17	214	105	88	84	202	77	102	122	70	164	3980								
8	CDS	155	105	88	84	202	77	102	122	70	164	145									
3	gene	111																			
END																					

Fig.2: First output file generated from RASA-GD



GI:372099104																						
1	source	540725																				
1	gene	540725																				
1	mRNA	509	142	94	178	90	75	138	201	3	175	214	105	88	84	202	77	102	122	70	164	4089
2	mRNA	413	142	94	178	90	138	201	175	214	105	88	84	202	77	102	122	70	164	4101		
2	gene	624																				
3	mRNA	312	178	90	138	201	175	214	105	88	84	202	77	102	122	70	164	4101				
4	mRNA	312	178	90	138	138	175	214	105	88	84	202	77	102	122	70	164	4101				
1	CDS	168	90	75	138	201	3	175	214	105	88	84	202	77	102	122	70	164	145			
2	CDS	168	90	138	201	175	214	105	88	84	202	77	102	122	70	164	145					
3	CDS	168	90	138	201	175	214	105	88	84	202	77	102	122	70	164	145					
4	CDS	168	90	138	138	175	214	105	88	84	202	77	102	122	70	164	145					
END																						

GI:358485511																						
1	source	413384																				
1	gene	413384																				
1	mRNA	110	147	83	178	90	138	51	195	163	214	105	88	84	202	77	102	122	70	164	3847	
1	CDS	168	90	138	51	195	163	214	105	88	84	202	77	102	122	70	164	145				
home																						

Fig.3: This is a output result showing lengths of Genbank Elements according to different GI.

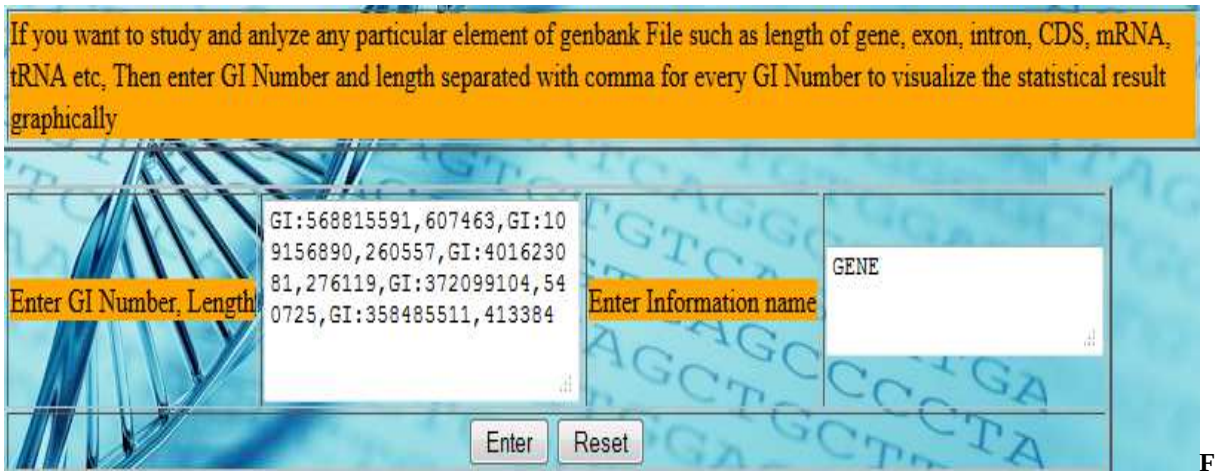


Fig .4: This shows the second input values.

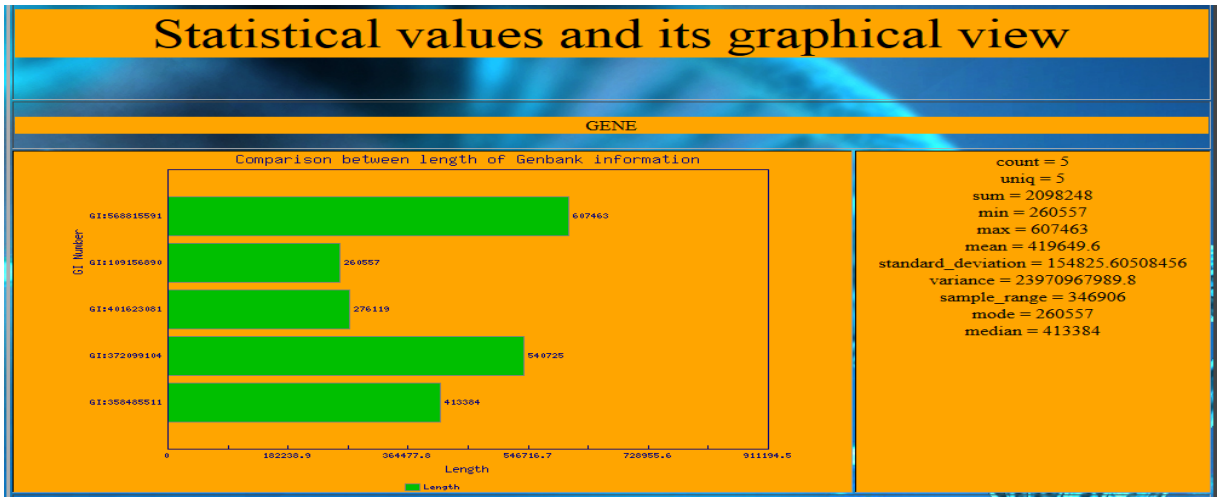


Fig .5: This is graphical and statistical value of different gene lengths are obtained from RASA-GD.

In first result we have obtained lengths of differenent elements contained in different GI number. Then we have selected the gene lengths of foxp2 in different organisms. The selected lengths are putted in second box shown in fig.2,as GI:568815591,607463, GI:109156890,260557, GI:401623081,276119, GI:372099104,540725, GI:358485511,413384 .

The gene length variation shown graphically in fig.4, can easily visualized. The statistical value gave following information such as unique sequence is 5 that are different from other sequences, count is 5 that is total number of sample genes, and sum is 2098298 that is total sum of all gene length, min. value 260557 that is minimum value, max. Is 607463 that is value of maximum length of the gene, mean is 419649.6 that gave the mean value of total gene length and also gave standard deviation that is 15425.6050, variance is 23970967989.6, sample range is 346906, mode is 260557 and median is 413384.

DISCUSSION AND ANALYSIS

The tool RASA-GD used in retrieval and graphical visualization of Genbank elements such as length of genes, exon, intron, mRNA,CDS, tRNA etc from Genbank file and also used to calculate the statistical values of desire length of elements contained in Genbank file. The different organisms of foxp2 genes found different lengths.

REFERENCES

1. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2012; 40:D48–D53. [PubMed]
2. Andrade M.A, Brown N.P, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C. Automated genome sequence analysis and annotation. 1990. *Bioinformatics* **15:** 391–412.
3. Fisher, S. E., Vargha-Khadem, F., Watkins, K. E., Monaco, A. P., Pembrey, M. E. Localisation of a gene implicated in a severe speech and language disorder. Nature Genet. 18: 168-170, 1998. Note: Erratum: Nature Genet. 18: 298 only, 1998. [PubMed: 9462748]
4. Руслан У. Закиров (02 August 2013) GDGraph-1.48 http://search.cpan.org/dist/GDGraph/Graph.pm [accessed 10/10/2014].
5. Rhet Turnbull (13 July 2002) Statistics-Descriptive-Discrete-0.07 http://search.cpan.org/~rhettbull/Statistics-Descriptive-Discrete-0.07/Discrete.pm#COPYRIGHT[accessed 10/10/2014].