Intergenics: A tool for extraction of intergenic regions

Kavitha Kurup*, Sujitha Mary, Ziauddin Ahamed Rafi

Centre of excellence in Bioinformatics, School of Biotechnology, Madurai Kamaraj University, Madurai 625 021, Tamil Nadu, India; kavitha kurup - Email- kavitha@mkustrbioinfo.com; Phone: +91 452 2459141; Fax: +91 452 2459105 ; *Corresponding author

Abstract :
For the past one decade, there has been considerable explosion of interest in searching novel regulatory elements in the intergenic region between the protein coding regions. The microbial genomes are the most exploited in terms of intergenic (noncoding) regions due to its less complexity. We think, the increasing pace of genome sequencing calls for a tool which will be useful for the extraction of intergenic regions. IntergenicS (Intergenic Sequence) is a tool which can extract the intergenic regions of microbial genomes at NCBI. All the unannotated regions between annotated protein coding genes and noncoding RNA genes can be extracted. It also deals with the calculation of GC base composition of the intergenic regions. This will be a useful tool for the analysis of noncoding regions of both bacterial and archael genomes.

Background :
As the intergenic regions are home to a wide diversity of functional elements, its analysis will be adding more annotations in genomes. In the past few years, the intergenic regions in microbial genomes are largely analyzed for noncoding RNAs such as rRNA, tRNA and especially small RNAs (functionally active regulatory sRNAs) which play an important role in cellular functions, including RNA processing, mRNA stability, translation, protein stability and secretion [1]. Exploring the locations of genes encoding the known sRNAs, Argaman *et al* found that they are primarily located in the empty intergenic regions with no other annotated genes on either strand [2]. Based on computational analysis in the intergenic regions, several sRNAs are identified in microbes and its number is increasing day-by-day. Genes encoding peptides are often missed in genomic annotations due to their small size. Search for unannotated small ORFs in the intergenic regions has revealed several peptides which are important regulators of growth, development and physiology of organisms [3]. Besides this, several promoters, terminators, riboswitches etc… have been identified and annotated from the intergenic regions.Microbial genomes are highly diversified in terms of chromosomal size, copy number, topology, and GC content. For example, *Mycoplasma mobile* is having 24.9% GC content [4] while for *Micrococcus luteus* it is 74% [5]. The base composition of a sequence tells us the nature of that sequence. The unusual GC content in the intergenic regions may reveal horizontally transferred genomic islands, sRNAs, or other functionally important elements. GC content was used to detect sRNA genes in AT-rich archael genomes such as *Pyrococcus furious* and *Methanococcus jannasehii* [6]. GC content is an important parameter of a particular bacterial species that can be accurately obtained from the genome sequence.We have developed an online tool for extraction of intergenic regions and calculation of its GC content from microbial genomes at NCBI. The first step in finding any of the functional elements in the intergenic region is the extraction of intergenic sequence. So we hope this tool may reduce time of script writing and make the analysis of microbial intergenic regions much easier.

Implementation :
The web server IntergenicS (see **Figure 1**) is implemented on Apache server and the web interface is designed using HTML and PHP. The algorithms to extract intergenic regions of bacterial genomes *in silico* are complemented with the language of python.

Main functions include: 1. Intergenic region extraction ; 2. Unannotated intergenic region extraction (excludes annotated noncoding RNAgenes in the IRs) ; 3. %GC calculation. Once the genome is selected, the intergenic region or unannotated region of the selected genome is displayed based on user's choice. Initially user has to select the microbe of their interest then they have to choose either intergenic region or unannotated region for displaying corresponding coordinates and sequence. The output can be further filtered by specifying the size of the region. GC content calculation which helps in the further analysis of intergenic regions is also incorporated into this tool. Results are displayed in tabular format. User can either view the results or save the results into their in-house systems and do further analysis. The work flow of the server is given in **Figure 2.**

As these intergenic regions will be having annotated noncoding RNA genes also, one who is interested in the analysis of completely unannotated regions, this tool will help to omit all the annotated protein coding genes and noncoding RNA genes and will extract the rest of the sequence. Due to the small size of small noncoding RNAs is (below 500bases), many researchers avoid the intergenic regions of size less than 500 bases for more accurate prediction. For purposes like this, this tool will allow the user to specify intergenic regions of particular size range. The calculation of GC base composition will give an idea of the nature of the sequence in the selected intergenic region.

Conclusion :
IntergenicS is expected to be a useful tool for those biologists who are interested in regulatory elements in the noncoding regions of microbial genomes. Currently this user friendly tool is designed only for the purpose of sequence extraction and to get GC content. In future, we would like to extend this tool so that the user will be able to retrieve all the functional elements in the intergenic regions from any organism of their interest. Project home page: http://bicmku.in:8081/intergenics
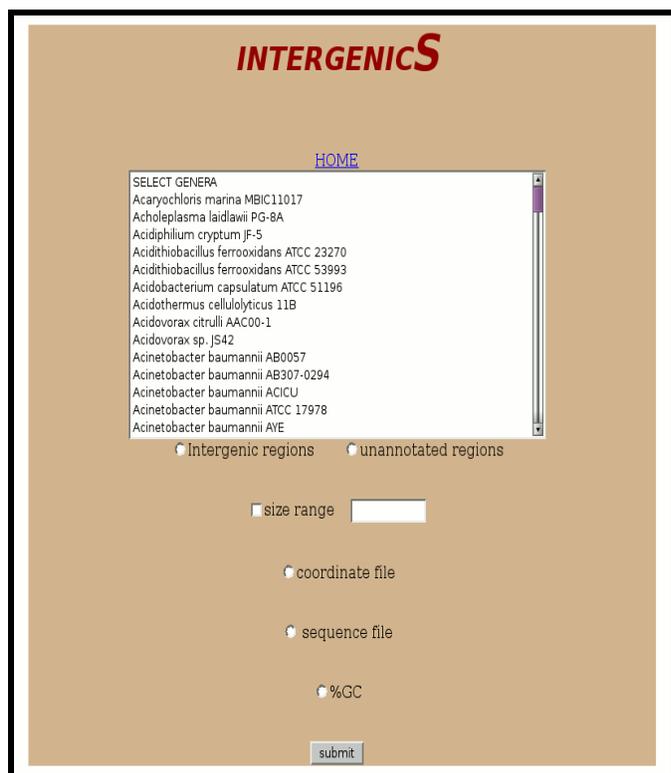
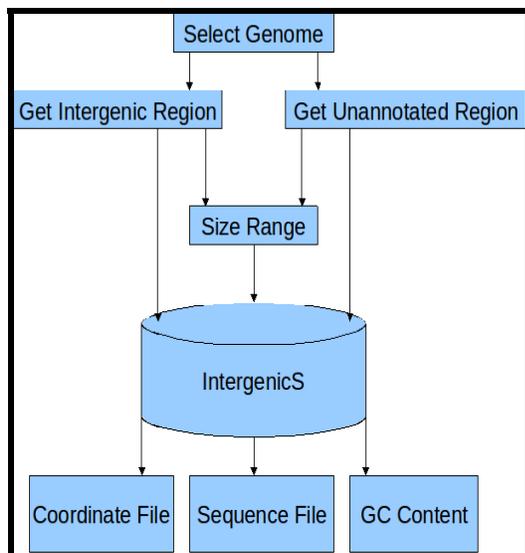**Figure**1: The search page of IntergenicS

Figure 2: Diagram showing the work flow of the server

References:
[1]   Wassarman *et al. Trends Microbiol* **7**: 37 (1999) [PMID: 10068996]
[2]   Argaman *et al. Current Biology* **11**: 941 (2001) [PMID: 11448770]
[3]   A Kevin *et al. Plant Physiol* **142**(3): 831 (2006) [PMID: 16998087]
[4]   JD Jaffe *et al. Genome Res* **14**:1447 (2004) [PMID: 15289470]
[5]   S Casjens *Ann. Rev.Genet.* **32**: 339 (1998) [PMID: 9928484]
[6]   RJ Klein *Proc Natl Acad. Sci. USA* **99**:7542 (2002) [PMID: 12032319]