Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Application Note

Junker: An Intergenic Explorer for Bacterial Genomes

Jayavel Sridhar[1,2#], Radhakrishnan Sabarinathan[3#], Shanmugam Siva Balan[3], Ziauddin Ahamed Rafi[1], Paramasamy Gunasekaran[2], and Kanagaraj Sekar[3*]

[1]*Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai Kamaraj University, Madurai 625021, Tamilnadu, India;*
[2]*UGC-Networking Resource Centre in Biological Sciences, School of Biological Sciences, Madurai Kamaraj University, Madurai 625021, Tamilnadu, India;*
[3]*Bioinformatics Centre, Indian Institute of Science, Bangalore 560012, Karnataka, India.*

Genomics Proteomics Bioinformatics 2011 Oct; 9(4-5): 179-182 DOI: 10.1016/S1672-0229(11)60021-1

Received: Feb 01, 2011; Accepted: Jun 21, 2011

Abstract

In the past few decades, scientists from all over the world have taken a keen interest in novel functional units such as small regulatory RNAs, small open reading frames, pseudogenes, transposons, integrase binding attB/attP sites, repeat elements within the bacterial intergenic regions (IGRs) and in the analysis of those "junk" regions for genomic complexity. Here we have developed a web server, named Junker, to facilitate the in-depth analysis of IGRs for examining their length distribution, four-quadrant plots, GC percentage and repeat details. Upon selection of a particular bacterial genome, the physical genome map is displayed as a multiple loci with options to view any loci of interest in detail. In addition, an IGR statistics module has been created and implemented in the web server to analyze the length distribution of the IGRs and to understand the disordered grouping of IGRs across the genome by generating the four-quadrant plots. The proposed web server is freely available at the URL http://pranag.physics.iisc.ernet.in/junker/.

Key words: bacterial genome, intergenic region, web server, statistics module

Introduction

The genomic era has witnessed the sequencing of over 1,400 prokaryotic genomes and this enables scientists to analyze the genome to get a clear insight into its functional aspects. Prokaryotic intergenic regions (IGRs) are a natural home to a variety of functional elements, thus the annotation of IGRs is essential for the complete understanding of bacterial physiology. In the past years, bacterial IGRs were routinely analyzed to identify structural non-coding RNAs (tRNA, rRNA and sRNA), which have multiple roles in the survival of the cell (*1, 2*). It was identified that IGRs carry important functional units like transposons (*3*), integrase binding sites (*4*), transcription factor binding sites, small open reading frames (ORFs), pseudogenes and inverted repeats (*5*). Recently, the traces of potential coding genes were also determined in IGRs (*6*). Thus, a few qualitative and quantitative studies were performed to identify the dynamics of bacterial IGRs. One such study on the *Escherichia coli* K12-MG1655 genome (*7*) compared the cumulative length distribution of IGRs between two repli-

#Equally contribution.
*Corresponding author.
E-mail: sekar@physics.iisc.ernet.in

cores (left and right) to identify the impact of IGR on the distribution of sRNA-encoding genes. They found that most of the sRNA genes were located in the left core, though the proportions of IGRs were equal on both segments. They also pointed out that a high number of sRNAs were residing within the IGRs of length between 300 to 900 nucleotides. On the other hand, the sum of the total non-coding DNA or IGR content was found to be associated with the increased biological complexities of the organisms (*8*). Although a few computational methods were developed to retrieve the genes and their intergenic contexts (*9, 10*), no specific tool is available for the identification of the distribution pattern and statistical analysis of IGRs at a genome level. Thus, we have developed a web-based tool, named Junker, to identify the length distribution pattern of IGRs in a complete genome. The proposed server can also be used to calculate the cumulative intergenic content of the four equal segments of the genome (quadrants) or left and right replicores (*2, 7*). The flanking distance between the neighboring genes provides a measurement of local gene density (LGD) (*11*), which indicates that the quadrant specific intergenic content has inverse relationship with the LGD and is positively correlated with variable segments of the genome (*12, 13*). The proposed web server is freely available at the URL http://pranag.physics.iisc.ernet.in/junker/.

Web Server

Implementation and utilities

The proposed web server integrates and reports information about the IGRs present in the bacterial genomes. To create a local intergenic database, all the available bacterial genomes have been downloaded from the NCBI portal (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). Next, the corresponding IGRs from the bacterial genomes were filtered out by excluding the protein and RNA encoding regions. In view of the above, two options, one is protein annotations and the other is protein and RNA annotations, are implemented in the server. Thus, the web interface enables users to search for the distinct IGRs based on their length and location. By default, Junker searches for IGRs with minimum length of 20 nucleotides, but an option is also provided to increase the minimum length. In addition, the nucleotide sequences extracted from IGRs are subjected to various analyses. For example, the experimentally determined pseudogenes are mapped using the annotated gene information from GenBank file. The presence of functional protein coding regions or ORFs in the IGRs is predicted using the gene prediction tools GeneMark2.5 (*14*), Glimmer3 (*15*) and Prodigal2.50 (*16*). In addition, the identical, tandem and inverted repeats in the IGR sequences are identified using FAIR (*17*) and "etandem" and "einverted" programs from EMBOSS suite (*18*).

All the IGR extractions, file handling modules and search engine were designed and implemented using Perl/CGI scripts. The histograms and circular maps presented were created using GD graph module (v1.43) (http://search.cpan.org/~mverb/GDGraph-1.43/). The web server runs under Solaris (v10.0) operating system on a 64 bit Quad-core Intel Xenon 5430 processor of 2.67 GHz with 4 GB of random access memory. The web server is implemented with user-friendly options to give explicit results. Presently, the local genome database of Junker contains 1,023 bacterial genomes.

Features

Users can select their genome of interest from the list provided in the index page of the server. Additional options are provided for the users to change the minimum length of IGRs and their location.

Selection of IGR of interest

The web server enables users to select a particular region from the whole genome by using physical genome map viewer. In general, the selected region covers the interval of 200,000 base pairs and is used to list the IGRs present in the selected region. The detailed report of IGRs extracted from the selected map position contains their start and end position, adjacent flanking gene IDs with their length information, different types of repeat elements and known pseudogenes present in the IGR sequence (Figure S1). In addition, options are provided for the users to download (in FASTA format) or display the interested IGR sequence.

IGR statistics module

The IGR statistics module has two major utilities to calculate the length distribution and the four-quadrant plots. The length distribution of the IGRs in different length intervals is represented in an interactive histogram, which also enables users to get the IGR sequences in FASTA format. Similarly, the cumulative lengths of the IGRs within the four quadrants of the genome are displayed using a pie chart known as four-quadrant plots. There are four scale points used in the pie chart to represent the complete genome in four quadrants.

Application

The genome of *Sodalis glossinidius* str. Morsitans (NC_007712) is reported to have the least coding capacity among the prokaryotes (*19*). Analysis of the *S. glossinidius* genome using the method indicated by Taft *et al* (*8*) shows that the genome has an ncDNA/tgDNA ratio of only 50.91%. This fact was confirmed in our study by comparing the *S. glossinidius* genome with other Gammaproteobacteria genomes (Figure S2). We analyzed the *S. glossinidius* genome sequence using Junker with the default options and found a total of 1,837 IGRs. Moreover, the length and positional distribution of these IGRs were analyzed using IGR statistics module (Figure S3). Figure S3A indicates that the genome contains many IGRs in different lengths with the maximum of 16 Kb. In addition, the four-quadrant plot of the genome indicates that the disordered grouping of IGRs accumulated mostly in the fourth quadrant compared to the others (Figure S3B). Furthermore, similar analysis with other genomes has shown that *Orientia tsutsugamushi* Boryong (NC_009488) (*20*) and *Thermocrinis albus* DSM14484 (NC_013894) have the highest (51.21%) and the lowest IGR ratio (3.98%), respectively.

The calculated percentages of known CDS (percentage of genome coding for proteins) and IGR (percentage of IGR in the genome) ratios for 1,023 bacterial genomes are available in the form of a table in the web server (http://pranag.physics.iisc.ernet.in/cgi-bin/junker/table.pl).

Conclusion

Junker is a web-based tool designed to efficiently access and analyze the IGRs in bacterial genomes. The selected query genome is represented in the form of a physical genome map, which facilitates the users to select a genome region of interest. In addition, the IGR sequences are checked for the presence of known pseudogenes, probable coding regions and other repetitive elements. Moreover, the length distribution of IGRs over the whole genome is shown as histograms and their disordered grouping is plotted onto a four-quadrant pie chart. It is believed that Junker will be helpful for the in-depth analysis of IGRs.

Acknowledgements

This paper is dedicated to the memory of late Prof. Ziauddin Ahamed Rafi who was the inspiration behind this study. We thank the support from Bioinformatics Centre and the Interactive Graphics Facility, Indian Institute of Science. JS thanks the Department of Biotechnology, Government of India for funding the projects. PG and JS thank the University Grants Commission for funding the Networking Resource Centre in Biological Sciences, Madurai Kamaraj University, Madurai.

Authors' contributions

JS conceived and coordinated the construction of the web server. JS and RS drafted the manuscript. RS and SSB developed the web interface and the scripts for prediction. ZAR and KS improved the web server and revised the manuscript. PG conceived the idea of the study and helped the revision of the manuscript. All authors read and approved the final manuscript.

References

1 Wassarman, K.M., *et al*. 2001. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 15: 1637-1651.
2 Hershberg, R., *et al*. 2003. A survey of small-RNA encoding genes in *Escherichia coli*. *Nucleic Acids Res.* 31: 1813-1820.
3 Siguier, P., *et al*. 2006. Insertion sequences in prokaryotic

genomes. *Curr. Opin. Microbiol.* 9: 526-531.
4 Doublet, B., *et al*. 2008. Secondary chromosomal attachment site and tandem integration of the mobilizable *Salmonella* genomic island 1. *PLoS One* 3: e2060.
5 Sharples, G.J. and Lloyd, R.G. 1990. A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes. *Nucleic Acids Res.* 18: 6503-6508.
6 Fu, L.M. and Shinnick, T.M. 2007. Genome-wide analysis of intergenic regions of *Mycobacterium tuberculosis* H37Rv using Affymetrix GeneChips. *EURASIP J. Bioinform. Syst. Biol*. 2007: 23054.
7 Blattner, F.R. *et al*. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453-1462.
8 Taft, R.J., *et al*. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29: 288-299.
9 Ray, W.C. and Daniels, C.J. 2003. PACRAT: a database and analysis system for archaeal and bacterial intergenic sequence features. *Nucleic Acids Res.* 31: 109-113.
10 Oberto, J. 2008. BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence. *Bioinformatics* 24: 424-425.
11 Haas, B.J., *et al*. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461: 393-398.
12 Chiapello, H., *et al*. 2008. MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. *BMC Bioinformatics* 9: 498.
13 Banos, R.C., *et al*. 2009. Differential regulation of horizontally acquired and core genome genes by the bacterial modulator H-NS. *PLoS Genet.* 5: e1000513.
14 Borodovsky, M. and McIninch, J. 1993. GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.* 17: 123-133.
15 Delcher, A.L., *et al*. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27: 4636-4641.
16 Hyatt, D., *et al*. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.
17 Banerjee, N., *et al*. 2008. An algorithm to find all identical internal sequence repeats. *Curr. Sci.* 95: 188-195.
18 Rice, P., *et al*. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16: 276-277.
19 Toh, H., *et al.* 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res.* 16: 149-56.
20 Cho, N.H., *et al.* 2007. The *Orientia tsutsugamushi* genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. *Proc. Natl. Acad. Sci. USA* 104: 7981-7986.

Supplementary Material

Figures S1-S3

DOI: 10.1016/S1672-0229(11)60021-1