SynRio: R and Shiny based application platform for cyanobacterial genome analysis

Karthick Lakshmanan, Arul Prakasham Peter, Shylajanaciyar Mohandass, Sangeetha Varadharaj, Uma Lakshmanan & Prabaharan Dharmar*

National Facility for Marine Cyanobacteria ,Sub-Distributed Bioinformatics Center, Department of Marine Biotechnology, School of Marine Sciences, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India; Prabaharan Dharmar – Email: dharmarpraba@gmail.com; *Corresponding author

Abstract:
SynRio is a Shiny and R based web analysis portal for viewing *Synechocystis* PCC 6803 genome, a cyanobacterial genome with data analysis capabilities. The web based user interface is created using R programming language powered by Shiny package. This web interface helps in creating interactive genome visualization based on user provided data selection along with selective data download options.

Availability:
SinRio is available to download freely from Github - https://github.com/NFMC/SynRio or from http://www.nfmc.res.in /synrio/. In addition an online version of the platform is also hosted at nfmc.res.in/synrio, using shiny server (open source edition) installation.

Background:
In recent years, cyanobacteria have acquired more attention in the field of drug discovery as they source natural products **[1,2].** The cyanobacterial model organism-an autotrophic prokaryote-*Synechocystis* sp. PCC 6803 has been widely used in various genomic and transcriptomic studies for many years, to generate in abundance, a variety of genome level data sets which are now getting connected to better understand this organism. There are many databases available which gather genome wide information about this bacterium (Example: Cyanobase **[3]**, KEGG **[4]** etc.) and here is our attempt to represent this model cyano genome data with simple plotting functionality utilizing R programming language. As R is more suitable for such tasks of combining user defined data modification with filtering and effective graphical plotting, we chose this programming language to build our cyano genome

interface. In addition, for a cyanobacterial researcher without programming experience using R for data analysis pose a greater challenge, since R has a steep learning curve. With a simple web-based GUI, we are lessening this hurdle and highlighting the biological content to the researchers for better data analysis. The Shiny package from R studio is a latest introduction in R programming for developing web interfaces. To our advantage, Shiny and Shiny server packages provided us a great opportunity to design a minimalistic but dynamic web based platform. Here we introduce "SynRio": a simple frame-work which leverages dynamically the power of R with web server functionality of Shiny for interactive exploration of model cyanobacteria. We hosted a working version of SynRio at www.nfmc.res.in/synrio, using a remote web server with shiny server setup. Shiny server installed in ubuntu server is backing the R web interface. The session based user interactions are

BIOINFORMATION

included in the framework design and are handled smoothly by Shiny package.
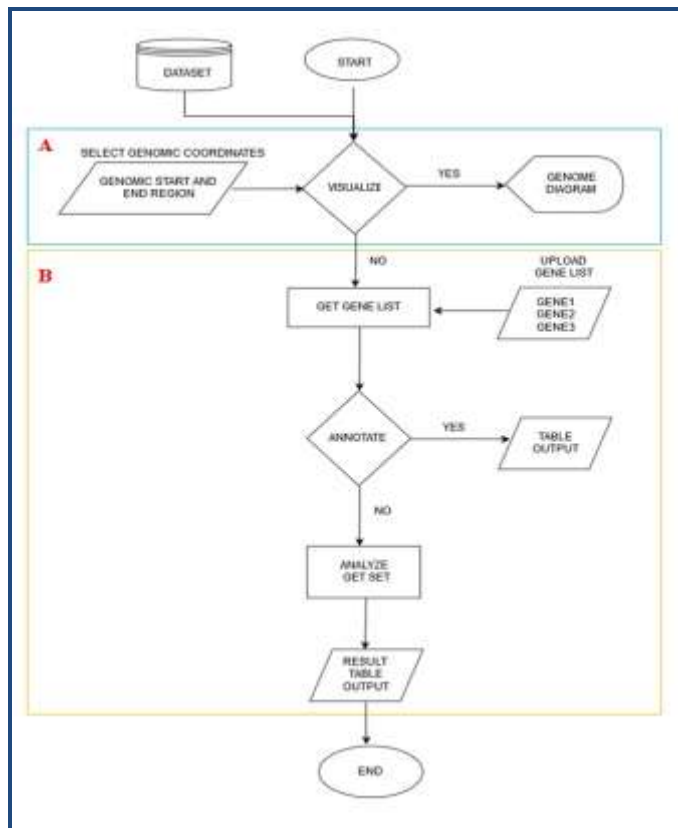
Figure 1: SynRio-data visualization and analysis flowchart: A) Visualization of genomic region. B) Analysis of user uploaded gene list.

In addition to the Shiny package, we utilized some of the other R packages to enrich SynRio suite. All necessary packages are installed from CRAN **[5]** and Bioconductor **[6]** sites like: Circlize for constructing circular genome plot; Biostrings for sequence manipulations; Rgraphviz for network analysis and seqLogo for PWM ploting. In addition ShinyBS package is utilize for web interface betterment. The genome sequence and gene based annotations are downloaded from NCBI site. Additional gene based and sequence based annotations are collected from various published resources and freely usable datasets.

Installation of standalone version:
To run SynRio from Github, install shiny package in R with additional packages like graph, Rgraphviz, shinyBS, circlize, Biostrings, shinyIncubator, seqLogo. Then run a tar or zip file directly like below:
runUrl("https://github.com/NFMC/SynRio/archive/master.tar.gz")
runUrl("https://github.com/NFMC/SynRio/archive/master.zip")
Or you can clone the git repository, then use runApp()
Alternatively, from here:
runUrl("http://www.nfmc.res.in/SynRio/archive/master.tar.gz")
runUrl("http://www.nfmc.res.in/SynRio/archive/master.zip")

Methodology:
SynRio is developed using *Synechocystis* PCC 6803 chromosomal sequence and gene based annotations. The web-browser based interface is almost completely designed using R programming language **[7]** with Shiny package **[8].** As the first step, the genome coordinates with start and end positions should be considered to draw genomic visualization using genome browser option button and then selection of the gene(s) or extraction of genome sequence is possible using utility features. The selected list of genes can also be made as a subset to be used for different analysis, like: gene clustering based on COG, KEGG; network construction based on protein-protein interactions; gene upstream region analysis etc (**Figure 1).**

Program input:
There are two different ways in which the genomic data can be accessed. Firstly, by specifying the user start and end genomic positions, a simple genome browser can be launched to visualize genomic region and further genes can be extracted from the selected genomic range **(Figure 1A).** Secondly, the user gene list (*Synechocystis* gene ids) can be uploaded directly into the tool for getting the annotation or for performing other advanced analyses **(Figure 1B).**

Utility:
In addition to its genome visualization capability, SynRio is also helpful in annotating and analysing the gene ids which are uploaded into it, using data upload feature. As an example, with a minimal list of genes associated with carotene biosynthesis **[9]**, **Table 1 (see supplementary material),** we tried to find a set of common features present among them. Abiotic stress based clustering feature along with COG clustering showed some similarity among the genes for salt/osmotic and high-light stress, when compared to other stress types **(Figure 2E, F).** Also, the -500 gene upstream region of all these genes was searched for high light-responsive element 1 (HLR1 element "KTTACAWW") **[10]** using pattern search utility. Interestingly, the HLR1 pattern is found partially, in almost 7 genes out of 8 selected **Table 2 (see supplementary material)** which needs more investigation. In addition, a pattern: "GGCGATCGCC" - a known and highly repetitive region of the genome, is also enriched in upstream region of most of these carotene biosynthesis genes.

Program output:
All the data/result tables from different panels are downloadable as .txt or .csv.

Features:
A Genome Browser with detailed navigation and zooming options. The selected genomic region could be highlighted with gene features and further colour coded using Cluster of Orthologous Groups (COGS) and Kyoto Encyclopedia of Genes and Genomes (KEGG) based gene annotation specifications **(Figure 2A)**. A simple pattern search tool is incorporated to perform genome wide DNA pattern identification. In addition, the gene set selected based on user specification can be clustered based on COG, Gene Ontology or grouped as network based on protein-protein interactions **(Figure 2B, C).** Finally, a gene cart to choose specific set of genes from different panels including browser or data upload panel can be used to download the selected data with mapped annotations or can be

reused to perform chosen gene set analysis. There are accessory tools like circular plot, which is helpful in advanced

genomic visualization **(Figure 2D).** A detailed help section is included to guide the navigation and analysis steps.
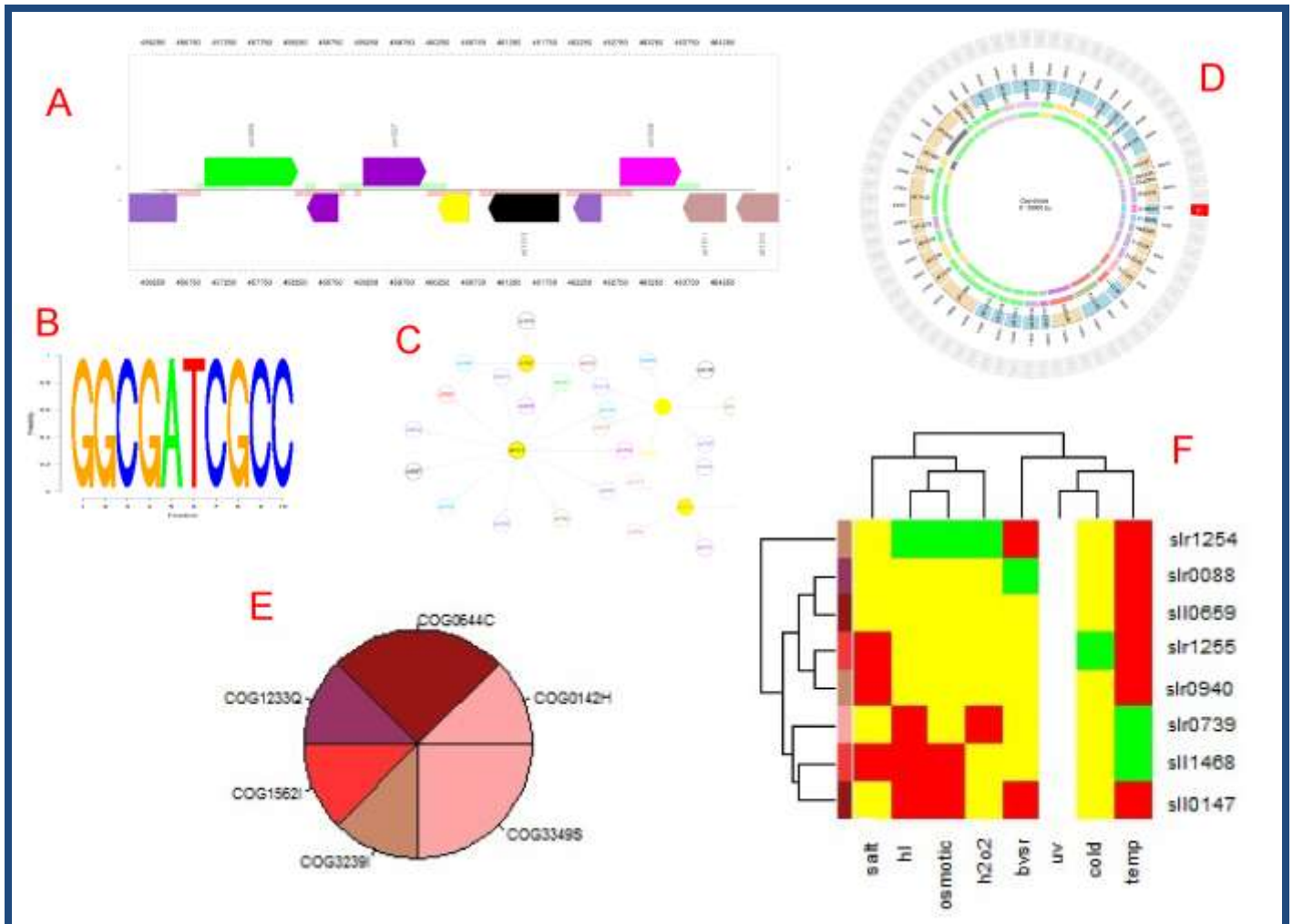
Figure 2: SynRio-web interface: A typical session starts with the selection of browser from the top navigation: **A)** The Synechocystis PCC 6803 genome browser with navigation tools. Additional panels to view annotations tracts are available; **B)** Pattern analysis with user defined DNA sequence motifs; **C)** Protein interaction and network view panel; **D)** Circular representation of a genomic region; **E)** COG based clustering of carotene biosynthesis genes; **F)** Abiotic stress clustering of carotene biosynthesis genes.

Conclusion:
SynRio helps to visualize cyanobacterial genome with known gene dataannotations. It also helps to perform gene or sequence based analysisdeveloped using R programming and Shiny package.

References:
[1] Shastri AA & Morgan JA, *Biotechnol Prog*. 2005 **21:** 1617 [PMID: 16321043]
[2] Tan LT *et al. Phytochemistry*. 2007 **68:** 954 [PMID: 17336349]
[3] Nakamura Y *et al. Nucleic Acids Res*. 1998 **26:** 63 [PMID: 9399802]
[4] Ogata H *et al. Nucleic Acids Res*.1999 **27:** 29 [PMID: 9847135]
[5] cran.r-project.org/
[6] Gentleman RC *et al. Genome Biol.* 2004 **5:** R80 [PMID: 15461798]
[7] http://www.R-project.org.
[8] http://www.rstudio.com/shiny/
[9] Liang C *et al. Int J Biol Sci*. 2006 **18**: 197 [PMID: 16967101]
[10] Takahashi T *et al. J Bacteriol.* 2010 **192:** 4031 [PMID: 20511509]

Supplementary material:

Table 1: Carotene biosynthesis genes in *Synechocystis* PCC 6803 genome

id	start	end	strand	name
sll0147	2188230	2190236	-	hypothetical protein
sll0659	3119608	3121176	-	hypothetical protein
sll1468	981691	982629	-	beta-carotene hydroxylase
slr0088	2894946	2896574	+	beta-carotene ketolase
slr0739	128802	129710	+	geranylgeranyl pyrophosphate synthase
slr0940	2004626	2006095	+	zeta-carotene desaturase
slr1254	1397819	1399237	+	phytoene dehydrogenase (phytoene desaturase)
slr1255	1399298	1400311	+	phytoene synthase

Table 2: HLR1 pattern found in Carotene biosynthesis genes

HLR1 element				KTTACAWWNNKTTACAWW
geneid	start	end	width	Pattern found in the sequence
slr0739	436	453	18	ATTANAAWNNNTWNCAGW
sll1468	444	460	17	KWWWNWWWNNTWWCANW
slr0739	437	453	17	KWWWNWWWNNTWWCANW
sll1468	396	413	18	AWTACAWWNNKNNNNWNW
slr1255	174	190	17	KTWACAAANNNWANANW
slr0088	292	298	7	KTWACNW
slr0940	365	371	7	KTWACNW
slr1254	90	96	7	KTWACNW
slr1254	373	379	7	KTWACNW
slr1255	16	22	7	KTWACNW
slr1255	174	180	7	KTWACNW
sll0659	233	240	8	KTTANAWW
slr1254	124	131	8	KTTANAWW

K = G or T, W = A or T, N = any nucleotide