Implementation of a genome browser using GBrowse and its application on microbial OMICS data

Thatchawanon Nantanaranon,[1] Jittisak Senachak,[2] Kalyanee Paithoonrangsarid,[2]
Supapon Cheevadhanarak,[3] Weerayuth Kittichotirat[4,*]

[1]*Bioinformatics and Systems Biology Program, King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok 10150, Thailand*
[2]*Biochemical Engineering and Pilot Plant Research and Development Unit, National Center for Genetic Engineering and Biotechnology at King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok 10150, Thailand*
[3]*School of Bioresources and Technology, Pilot Plant and Training Institute, King Mongkut's University of Technology Thonburi, Bangkok 10150, Thailand*
[4]*Systems Biology and Bioinformatics Research Group, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonbur (Bang Khun Thian Campus), Bangkok 10150, Thailand*
**e-mail: weerayuth.kit@kmutt.ac.th*

Abstract

The vast amount and complexity of omics data generated by high throughput technologies such as next generation sequencing and mass spectrometry presents a big challenge for scientists who are analyzing them. This is especially true when multiple types of omics data are available and integration is needed to make new discovery. In this study, we make use of Gbrowse, which is a genome browser tool, and created customized ways of displaying various omics data integratively. Our customized Gbrowse system was applied to *Arthrospira platensis* C1, which is a cyanobacterium that we have generated genomics and transcriptomics data for various conditions. We show that our system makes it very easy to search for genes of interest and by providing an integrative visualization of different types of omics data, gene of interest can easily be extracted without any deep computational skill. Besides *Arthrospira platensis* C1, the customized Gbrowse system developed in this study can easily be applied to other organisms where multiple types of omics data are available, which can greatly facilitate the discovery of new knowledge.

Keywords: *Arthrospira platensis* C1, genome browser, omics, transcriptome, proteome

Introduction

Nowadays, technologies for sequencing the genome of a living organism are developing faster and costs are becoming cheaper [Stranneheim and Lundeberg, 2012; Liu et al., 2012; Soon et al., 2013]. This has resulted in an exponential increase in the amount of organisms whose genomes were sequenced [Soon et al., 2013; Stein et al., 2002] and have generated a large amount of new sequence data [Soon et al., 2013]. The size and complexity of sequence data often makes it difficult for biologists to analyze, especially when the research group has limited number of human resource to process the data [Brusic and Ranganathan, 2008; Tao et al., 2004]. This hinders the ability for biologists to get a full usage of information [Pook et al., 1998]. To solve this problem, visualization techniques such as genome browser become a tool that can help biologists to organize and analyze a huge amount of sequence data. A genome browser can help to simplify the analysis of genomic data because it allows visualization of the sequence and therefore makes it easy to understand. A genome browser can display the whole genome sequence or zoom into a sub region and display features associated with that area of an organism genome, something which can be helpful for

understanding of overall sequence data of an organism [Tao et al., 2004; Gibson and Smith, 2003].

Researchers at King Mongkut's University of Technology Thonburi and other collaborators have generated and published a draft genome sequence of *Arthrospira platensis* C1 [Cheevadhanarak et al., 2012]. In addition, other omics data such as transcriptomics and proteomics data for various growth conditions/treatments for *Arthrospira platensis* C1 have also been generated. The availability of multiple omics data motivates us to create a computational tool that can integratively visualize these data and facilitate new discovery. In this study, we therefore make use of the GBrowse software to create a Genome Browser tool to visualize the genome sequence and annotation of *Arthrospira platensis* C1 as well as other omics data in an integrative fashion. We show that our system makes it very easy to search for genes of interest and by providing an integrative visualization of different types of omics data, target genes can easily be extracted without any deep computational skill.

Methodology

1. Implementation

The Gbrowse software was downloaded as a virtual machine package from the GMOD webpage (http://gmod.org/wiki/GBrowse2_VMs). The downloaded package was then installed and ran on a MacBook Pro-Mid 2012 with CPU: Intel®Core i5 CPU@2.5 GHz, Hard disk 500 GB and memory 4GB via a virtual machine software called VirtualBox (https://www.virtualbox.org/). Other customized search and informational pages were created using Perl and Hyper Text Markup Language.

2. Data Materials and Data Preparation

The genome sequence *Arthrospira platensis* C1 (PCC 9438) and its corresponding annotation data used in this was downloaded from NCBI in GenBank format using an accession number AFXD00000000 (http://www.ncbi.nlm.nih.gov/Traces/wgs/?download=AFXD01.gbff.1.gz). The genome data is incomplete and contains a total of 63 contigs. The transcriptome data for *Arthrospira platensis* C1 was kindly provided ahead of publication by Dr. Kalyanee Paithoonrangsarid from the Algae Biotechnology laboratory, King Mongkut's University of Technology Thonburi. The expression value of each gene in *Arthrospira platensis* C1 came from the experiment where cells were cultured at 35°C and then shifted to 42°C (high temperature stress). The samples were collected for 5 intervals e.g. 20, 60, 120, 240 and 480 minutes. Two-color microarray method was used to detect gene expressions and log ratio values were calculated in 3 replicates. Perl scripts were created to convert GenBank and Transcriptome data into file format that is compatible with Gbrowse program (GFF3 format).

3. System Implementation

After genome and transcriptome data were converted into GFF3 format, all data were uploaded into a MySQL database. After that, the Gbrowse configuration files that are found in directory /opt/gbrowse/etc were edited accordingly. This allows the details of new database, tracks, and plugins to be available when the genome browser page is displayed on a web browser. Finally, the description and path of configuration files were added into the Gbrowse main setting file (GBrowse.conf). Once the web server is restarted, the new genome browser containing *Arthrospira platensis* C1 (PCC 9438) omics data can be displayed by

going to http://localhost/fgb2/gbrowse/new_arthrospira. For gene expression search tool, it is created using HTML, CSS, Javascript and Perl computer language.

Results

1. Genomics and transcriptomes data

Genome sequence data of *Arthrospira platensis* C1 (PCC 9438) in GenBank file format that was downloaded from the GenBank database and the GFF3 format that was generated using a Perl script are shown in Figure 1.

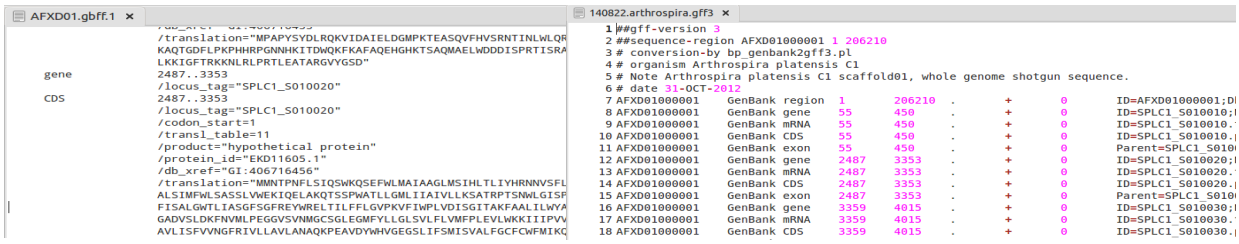

Figure 1: A GenBank file containing the genome sequence and annotation data of *Arthrospira platensis* C1 (left) and the corresponding GFF3 file that was created using a Perl script (right)

The transcriptome data, which is a tab-delimited file, was read and converted to GFF3 file format by using a custom made Perl script. The resulting files are showed in Figure 2 and Figure 3.

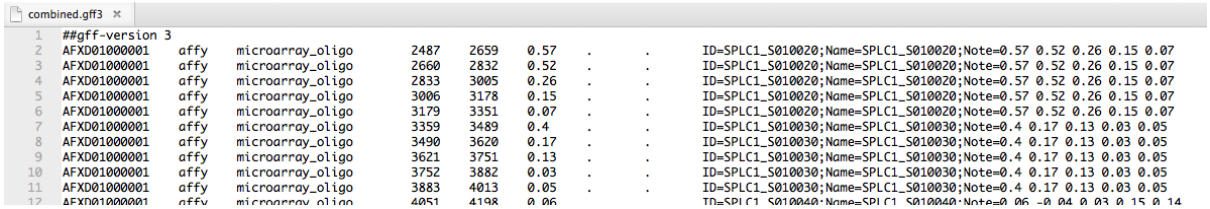

Figure 2: Transcriptome data in GFF3 file format that is used in "Transcriptome_Gene_Expression" track

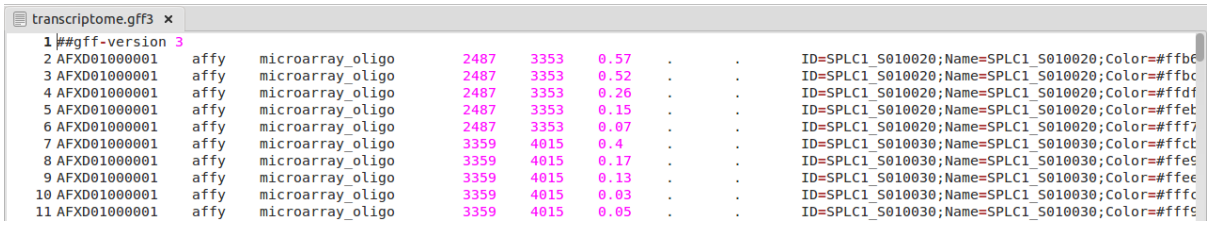

Figure 3: Transcriptome data in GFF3 file format that is used in "Transcriptome" track

2. System Implementation

The genomics and transcriptomics data of *Arthrospira platensis* C1 (PCC 9438), which are in GFF3 format, were uploaded into MySQL database. After that, the configuration file was edited by adding in the details of the databases, track configuration and plugins. Finally, the

description and path of the configuration files were added into the Gbrowse main setting file (GBrowse.conf) using standza shown below.

```
[new_arthrospira]
 description = new arthrospira platensis C1
 path        = new_arthrospira.conf
```

The new Genome browser system containing genomics and transcriptomics data of *Arthrospira platensis* C1 (PCC 9438) can then be displayed by going to http://localhost/fgb2/gbrowse/new_arthrospira. The Genome browser page will be visualized as shown in Figure 4.

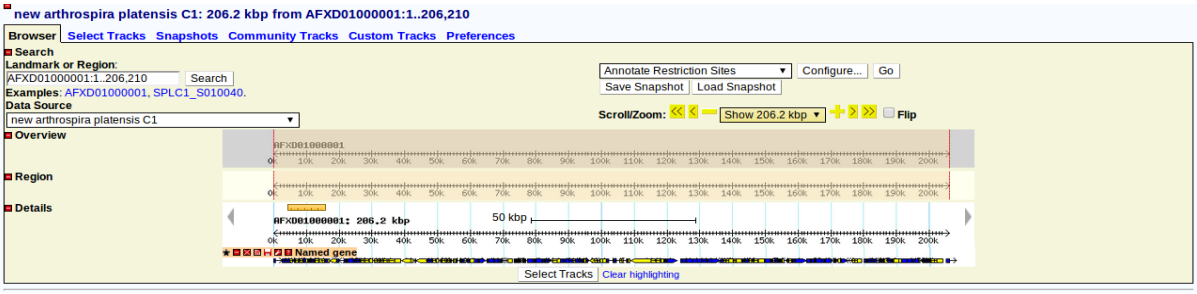

Figure 4: A genome browser tool visualizing *Arthrospira platensis* C1 (PCC 9438) genomics and transcriptomics data.

3. Using the genome browser

This section shows an example usage of our customized genome browser tool. When a user enter our tool, that user will see the page as shown in Figure 4, which displays a tracked called “Named gene”. This track shows all genes that are found in *Arthrospira platensis* C1 genome and is displayed by default. On this page, users can search for gene of interest by using the Reference Sequence <AFXDxxxxxxxx> or Name <SPLC1_xxxxxxx> as a keyword into the red box that is shown in Figure 5. After choosing the search result, the selected genomic feature will be displayed on the next page.

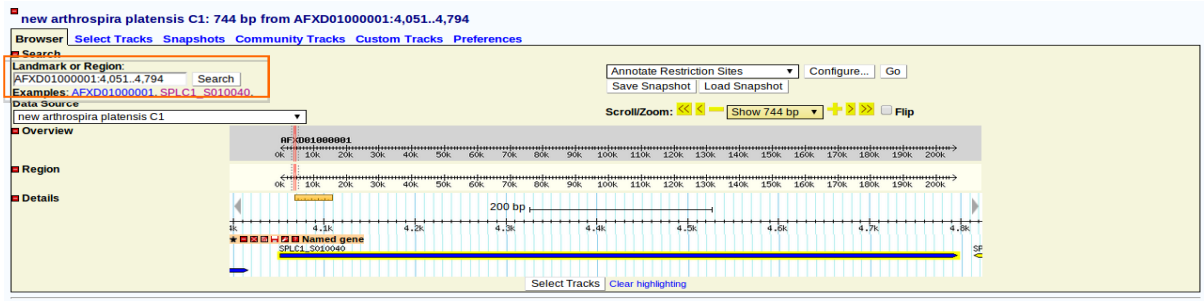

Figure 5: Search by gene name

On the track “Named gene”, if the users do a mouse hover over the feature, a popup balloon that contains a description of the gene will be displayed. In addition, when the user clicks on the feature will display a details of the gene and a link to download nucleotide sequence and amino acid sequence as shown in Figure 6.

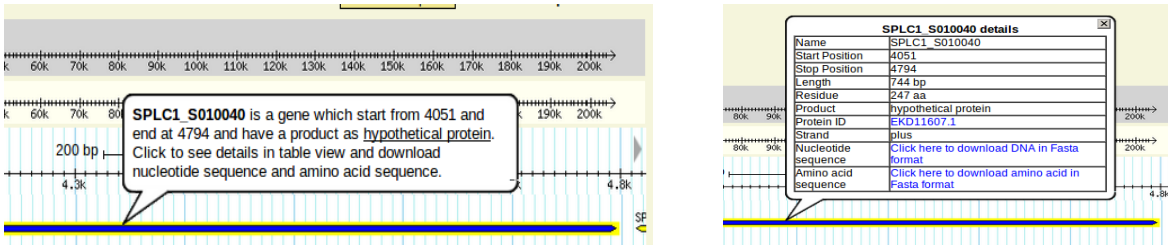

Figure 6: A balloon hover displaying description (left) and Balloon click display the details (right)

Next, if the user is focus to visualize the transcriptome data of *Arthrospira platensis* C1 on the genome browser tool. The users have to go to the second tabs called "Select tracks", this tab menu contains a lot of features that users can check to make it display on the genome browser page. To visual the transcriptome data as xyplot, the users must check on "Transcriptome_Gene_Expression", which is shown in the red box in Figure 7. After this box is checked, track "Transcriptome_Gene_Expression" will be displayed on the main page as shown in Figure 8. This track represents the expression value of each gene in each condition by using xy plot where X-axis shows each time condition and Y-axis shows the gene expression value. Expression values that are above zero represent up-regulation, which is displayed in the red color. On the other hand, expression values that are below zero represent down-regulation, which is displayed in the green color.

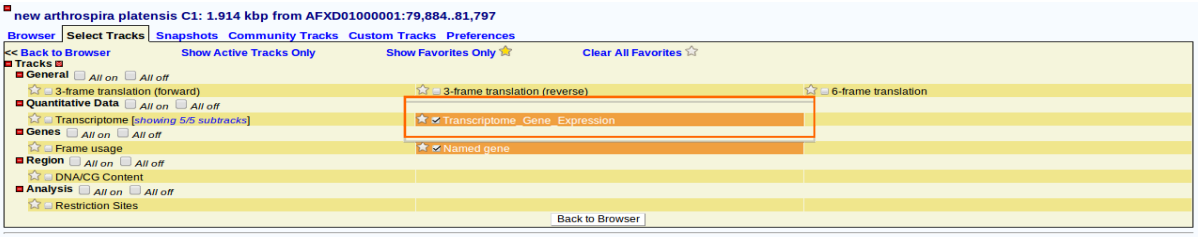

Figure 7: Selection of "Transcriptome_Gene_Expression" to open the track on the main page

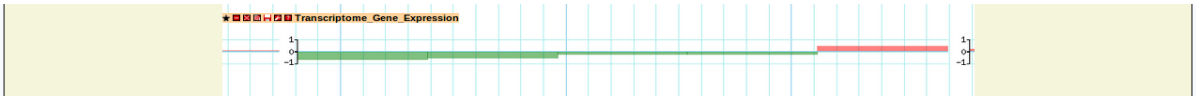

Figure 8: "Transcriptome_Gene_expression" track on the page

Similar to the track "Named gene", if the user do a mouse hover over or click on the feature "Transcriptome_Gene_Expression", a popup balloon that describes about the feature will be displayed. On the other hand, when the user clicks on the feature a table of expression values of the gene depend on the time condition will be displayed as show in Figure 9.

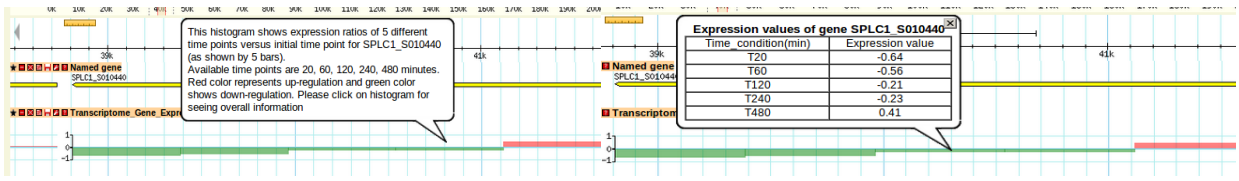

Figure 9: Balloon hover showing a description of a gene expression feature (left) and Balloon click displaying an expression value of the gene for each condition (right)

Users who are interested to see the transcriptome data of *Arthrospira platensis* C1 in expression profile on the page and follow these steps. In "Select tracks" menu, the user must choose "Transcriptome" as shown in the red box in Figure 10. After this box is checked, track "Transcriptome" will be displayed on the main page as shown in Figure 11. This track represents gene expression values using different color depend on the score of expression.

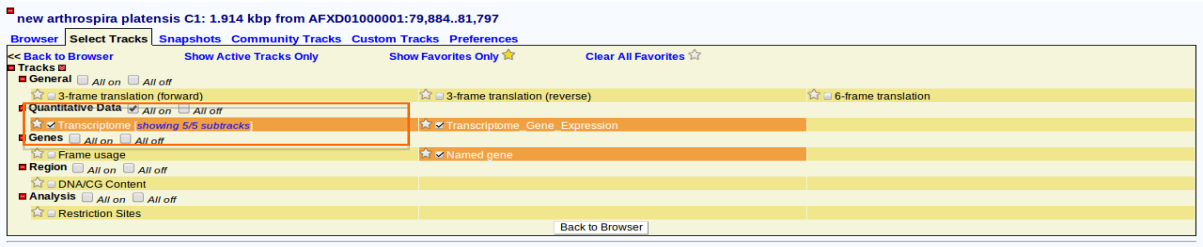

Figure 10: Select "Transcriptome" to open the track on the page

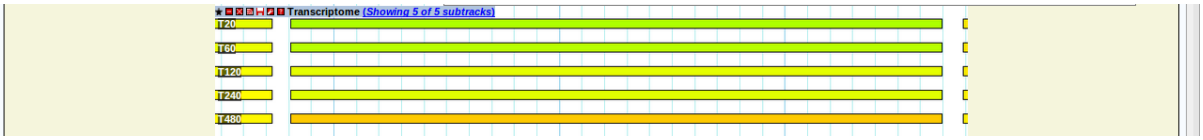

Figure 11: Display track "Transcriptome" on the page

When the users do a mouse hover over or click on the feature "Transcriptome", the tool will display a popup balloon that describes the feature. When the user clicks on the feature, the tool will display an expression value of the gene for each condition as show in Figure 12. Also in this track, the users can choose which time condition that they would like to show on the main display by clicking on "Showing 5 of 5 subtracks" near the title bar as show in the red box in Figure 13. Tracks will appear according to the box that users have selected.

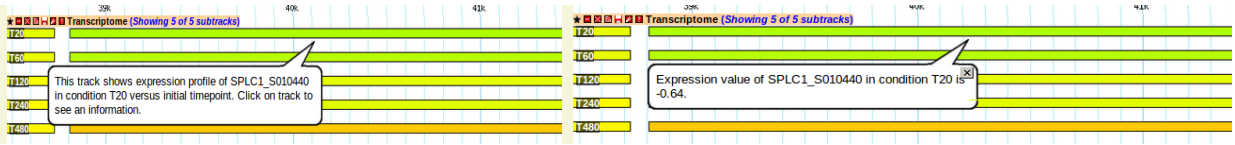

Figure 12: Balloon hover displaying description (left) and Balloon click displaying the details (right)

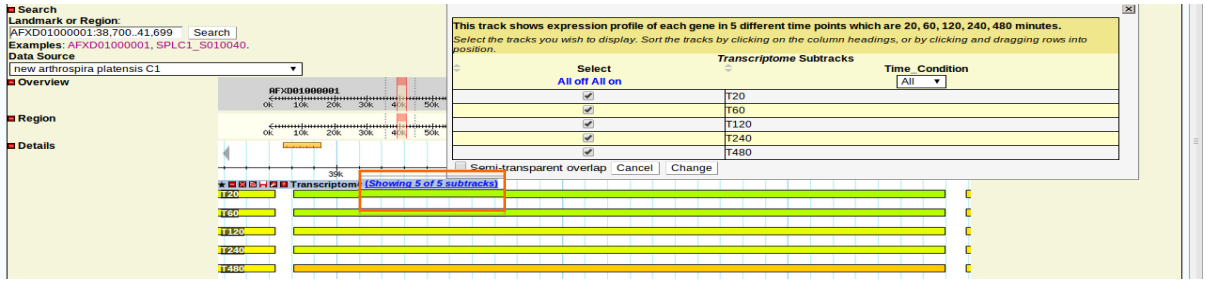

Figure 13: Transcriptome Subtracks

The last tool that was created in this work is the gene expression search tool as show in the Figure 14. When the users come to this page, the users have to do 2 steps. In the first step, the

users have to choose the mode between “absolute value” or “range of values”. In mode absolute value, the user can input the value into the form and then choose the criteria which can be greater than, equal or lower than. In contrast, the “range of values”, the users have to input 2 values for lower value and upper value. The second step, the users choose the condition that the user is interested and the logic which is “and” for searching the result which occur in every condition or the users can choose logic “or” for searching the result which occur in at least 1 condition. The result from search is shown in Figure 14 (right), which displayed the gene name, condition and expression value. The gene expression search result contains links that can bring users from the result page to the genome browser page by clicking on gene name.

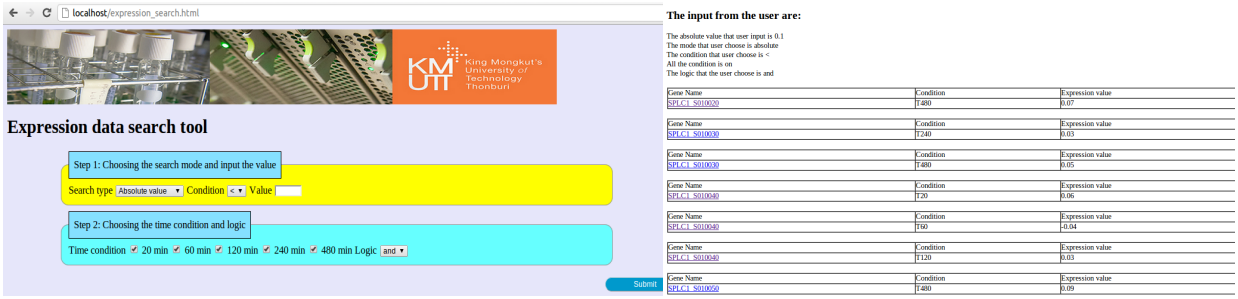

Figure 14: Expression data search tool interface (left) and the result from search (right)

Discussion

Arthrospira platensis C1 genome browser, which is implemented under GBrowse program visualizes the position of the gene in *Arthrospira platensis* C1 and gene expression for various time points to provide the biologist to understand how the characteristic of the gene is. In addition, visualized the position and transcription gene expression the user also can be visualized GC content, 3-frame and 6-frame translation. When the user hover on the gene will pop-up a short details of the gene and click on the gene the fully details and the link to download DNA and amino acid sequence will appear. In genome browser, user can submit selection for doing blast against NCBI, also allow zooming and panning or user can search into the specific gene that is interested and the user can using the plugin such as annotated restriction sites. In the future, we further to integrate the proteomics data of *Arthrospira platensis* C1 into our genome browser to provide the biologist can be analyzed which could help more understand on the gene content of *Arthrospira platensis* C1 and also provide some useful plugins and search tools to make the user is more convenience.

Conclusion

Arthrospira platensis C1 genome browser is a genome browser that is implemented to visualize the data, which related to *Arthrospira platensis* C1. The genome browser is implemented under GBrowse program by using Bioperl and hypertext markup language. *Arthrospira platensis* C1 genome browser provides the visualize of the position of the gene, transcriptome gene expression, GC content, 3-frame and 6-frame translation and having a features or tools that useful for Arthrospira researchers.

Acknowledgements

The authors would like to thank Dr. Weerayuth Kittichotirat for all the advice and encouragement, Assoc. Prof. Supapon Cheevadhanarak for a lot of advices and useful suggestions, Dr. Kalyanee Paithoonrangsarid and Dr. Jittisak Senachak for transcriptome gene expression data, proteome data respectively and also their valuable comment which is helpful to make the work complete. This research work was supported by Bioinformatics and Systems Biology Program at King Mongkut's University of Technology Thonburi and National Center for Genetic Engineering and Biotechnology (BIOTEC), NSTDA, Thailand.

References

Brusic V., Ranganathan S. (2008) Critical technologies for bioinformatics. Briefings in bioinformatics 9:261-262.

Gibson R., Smith D. R. (2003) Genome visualization made fast and simple. Bioinformatics 19:1449-1450.

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L., Law M. (2012) Comparison of next-generation sequencing systems. BioMed Research International 2012:1-11

Pook S., Vaysseix G., Barillot E. (1998) Zomit: biological data visualization and browsing. Bioinformatics 14:807-814.

Soon W. W., Hariharan M., Snyder M. P. (2013) High-throughput sequencing for biology and medicine. Molecular systems biology 9:1-14.

Stein L. D., Mungall C., Shu S., Caudy M., Mangone M., Day A., Nickerson E., Stajich E., Harris T.W., Arva A., Lewis S. (2002) The generic genome browser: a building block for a model organism system database. Genome research 12:1599–1610.

Stranneheim H., Lundeberg J. (2012) Stepping stones in DNA sequencing Biotechnology journal 7:1063–1073.

Cheevadhanarak S., Paithoonrangsarid K., Prommeenate P., Kaewngam W., Musigkain A., Tragoonrung S., Tabata S., Kaneko T., Chaijaruwanich J., Sangsrakru D., Tangphatsornruang T., Champrasert J., Tongsima S., Kusonmano K., Jeamton W., Dulsawat S., Klanchui A., Vorapreeda T., Chumchua V., Khannapho C., Thammarongtham C., Plengvidhya V., Subudhi S., Hongsthong A., Ruengjitchatchawalya M., Meechai A., Senachak J., Tanticharoen M. (2012) Draft genome sequence of *Arthrospira platensis* C1 (PCC9438). Standards in genomic sciences 6:43-53.

Tao Y., Liu Y., Friedman C., Lussier Y.A. (2004) Information Visualization Techniques in Bioinformatics during the Postgenomic Era. Drug discovery today: Biosilico 2:237–245.

http://www.ncbi.nlm.nih.gov/Traces/wgs/?download=AFXD01.gbff.1.gz.