Bioinformatics, 32(13), 2016, 2024–2025
doi: 10.1093/bioinformatics/btw095
Advance Access Publication Date: 22 February 2016
Applications Note

OXFORD

Genome analysis

Transcript structure and domain display: a customizable transcript visualization tool

Kenneth A. Watanabe[1,†], Kaiwang Ma[1,2,†], Arielle Homayouni[1], Paul J. Rushton[3] and Qingxi J. Shen[1,*]

[1]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA, [2]School of Medical Technology & Engineering, Henan University of Science & Technology, Luoyang, 471003, China and [3]Texas A&M AgriLife Research, Dallas, TX 75252, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on October 1, 2015; revised on January 21, 2016; accepted on February 12, 2016

Abstract

Summary: Transcript Structure and Domain Display (TSDD) is a publicly available, web-based program that provides publication quality images of transcript structures and domains. TSDD is capable of producing transcript structures from GFF/GFF3 and BED files. Alternatively, the GFF files of several model organisms have been pre-loaded so that users only needs to enter the locus IDs of the transcripts to be displayed. Visualization of transcripts provides many benefits to researchers, ranging from evolutionary analysis of DNA-binding domains to predictive function modeling.
Availability and implementation: TSDD is freely available for non-commercial users at http://shenlab.sols.unlv.edu/shenlab/software/TSD/transcript_display.html.
Contact: jeffery.shen@unlv.nevada.edu

1 Introduction

Due to high demand for publication quality transcript visualization software, there have been several software solutions available for researchers. Each of these solutions has advantages and drawbacks. For instance, GECA (Fawal *et al.*, 2012), which can be used online or downloaded to a server for offline usage, has a relatively quick runtime, and has an appealing user interface. However, it is difficult to use and lacks customizable features and custom motifs. GECA also requires the DNA sequence, protein sequence and a GFF3 file which could pose a problem for those who may not have this information available. FeatureStack's (Frech *et al.*, 2012) advantages include relatively high-quality images and the capacity to display custom domains. However, FeatureStack lacks an online version, requiring users to download and install a program onto their server. FancyGene (Rambaldi and Ciccarelli, 2009) is highly customizable, allowing selection of feature colors, feature sizes and custom motifs. However, it can only display a single gene at a time. GSDraw (Wang *et al.*, 2013) produces customizable, high-quality images with the capability of custom motifs. In addition, GSDraw provides users with a phylogenetic tree. However, GSDraw has a long run time, requires both the genomic and CDS sequences, and cannot utilize a GFF3 file. GSDS 2.0 (Hu *et al.*, 2015), is a fast, easy to use, web-based program that requires only a GFF3 file and produces customizable results that include custom motifs and a phylogenetic tree. However, GSDS has some limitations. GSDS can only process a maximum of 50 genes at a time, which may pose a problem when studying large gene families. For example, the WRKY gene family contains well over 100 genes in many species (Eulgem *et al.*, 2000; Zhang and Wang, 2005). Other limitations of GSDS include the inability to select a font or font size. GSDS also lacks the capability to vertically space genes so when the size of gene features is enlarged, the genes overlap. Herein, we report an alternative and enhanced transcript display software that is fast, easy to use, web-based and resolves issues with other transcript visualization software. TSDD is capable of providing researchers with publication quality images that can be customized and downloaded.

2 Usage and implementation

Transcript Structure and Domain Display (TSDD) can be accessed from our website at: http://shenlab.sols.unlv.edu/shenlab/software/TSD/transcript_display.html. TSDD relies on GFF/GFF3 files to draw the transcript structures and domains. GFF/GFF3 files were used as a data source to produce the transcript structures since almost all annotated genomes use this file format. The GFF files and protein sequences of 13 model organisms were preloaded into the TSDD database. To generate transcript structures of one of these organisms, users only need to select the organism from the pull-down menu and then enter the locus IDs of the genes they wish to display into the text box. Alternatively, users can load a '.txt' file of the locus IDs from their PC. For other organisms, users can select custom GFF3 or BED file from the pull-down menu and then can either paste the gene structure data into the textbox or load a file from their PC with the extension '.gff' or '.bed', respectively. Users can click the 'Demo Data' button to load example data of the selected organism or data file so the format of the data can be viewed.

TSDD will then parse the data and generate the transcript structures (Fig. 1A). TSDD can display at least 300 transcripts in a single run depending on the memory of users' PC.

TSDD is highly customizable to meet the needs of users. They can select from a variety of fonts and font sizes, and can also display an arrow indicating the orientation of the transcript (Fig. 1B). The vertical spacing between transcripts can also be adjusted, as well as the color and height of the genomic features: UTR, CDS or Introns. Users can also select the color and height of up to 20 custom domains. If the user selects one of the pre-loaded organisms, they can specify the pattern of the domains they wish to display. In addition, they may choose to display either all of the domains that fit the pattern, only the first or last, or ones within a specified distance from the N- or C-terminus. Since some transcripts have long introns making the transcript difficult to view, TSDD also has the option to compress the introns by a specified percentage or even omit the introns or UTRs for easier viewing (Fig. 1C). TSDD also has a preview feature on the main input screen which allows users to preview any changes to the default settings prior to generating the transcript structures. This feature will save users time if a large number of transcripts are being generated. TSDD also gives users the option to save their current configuration to a text file. This will allow users to quickly reload their data so that results can easily be replicated at a future time. The generated transcript structures can be saved as one of several file formats (PNG, PDF, JPG, GIF, BMP, TIFF), which can be downloaded to users' PC for viewing and editing. TSSD proves easy to use, fast, and customizable, in addition to providing features that none of the current alternatives provide.

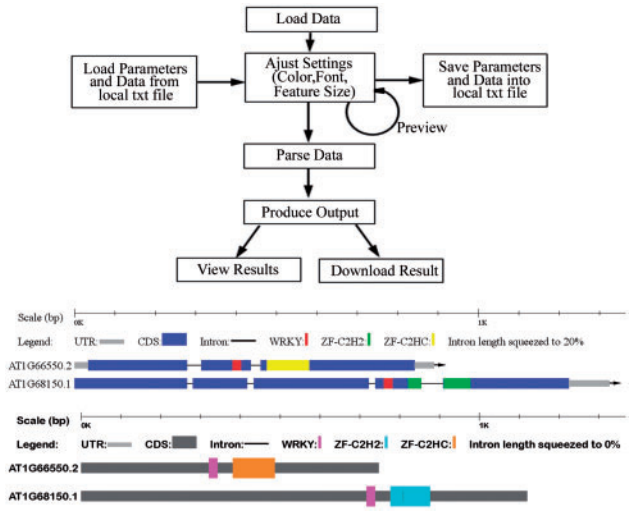

Fig. 1. (A) The workflow of TSDD for producing transcript structures. **(B)** Example output showing transcript structures of two Arabidopsis genes using default settings. **(C)** Alternative output showing customized output of the same two Arabidopsis genes

3 Future direction

We plan on building a database of domains and their consensus sequences using popular protein domain databases, such as Pfam (Finn *et al.*, 2014), so that users do not need to populate the pattern field manually. We also plan on having TSDD scan the entered loci for any domains within our database. Users will then be able to select which of the identified domains to display. The list of preloaded organisms will grow over time as users request them to be added to our database. We will listen to the suggestions and feedback from the users so that we can continuously improve TSDD to meet the growing needs of the scientific community.

4 Availability of software

Project name: Transcript Structure and Domain Display
Project home page: http://shenlab.sols.unlv.edu/shenlab/software/TSD/transcript_display.html
Programming language: HTML, Javascript, PHP
License: Open Source license GNU General Public License version 2.0
Restrictions to use by non-academics: license needed

Funding

This work was supported by grant 2008-35100-04519 from the United States Department of Agriculture. K. Ma received support from the China Scholarship Council (Grant No.:201308410108).

Conflict of Interest: none declared.

References

Eulgem,T. *et al.* (2000) The WRKY superfamily of plant transcription factors. *Trends Plant Sci.*, **5**, 199–206.
Fawal,N. *et al.* (2012) GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families. *Bioinformatics*, **28**, 1398–1399.
Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
Frech,C. *et al.* (2012) FeatureStack: Perl module for comparative visualization of gene features. *Bioinformatics*, **28**, 3137–3138.
Hu,B. *et al.* (2015) GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics*, **31**, 1296–1297.
Rambaldi,D. and Ciccarelli,F.D. (2009) FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics*, **25**, 2281–2282.
Wang,Y. *et al.* (2013) PIECE: a database for plant gene structure comparison and evolution. *Nucleic Acids Res.*, **41**, D1159–D1166.
Zhang,Y. and Wang,L. (2005) The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC EvolBiol.*, **5**, 1.