FeatureStack: Perl module for comparative visualization of gene features

Christian Frech, Caleb Choo and Nansheng Chen*

Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6

Associate Editor: Alex Bateman

ABSTRACT

Summary: FeatureStack is a Perl module for the automatic generation of multi-gene images. FeatureStack takes BioPerl-compliant gene or transcript features as input and renders them side by side using a user-defined BioPerl glyph. Output images can be generated in SVG or PNG format. FeatureStack comes with a new BioPerl glyph, decorated_gene, which can highlight protein features on top of gene models. Used in combination, FeatureStack and decorated_gene enable rapid and automated generation of annotation-rich images of stacked gene models that greatly facilitate evolutionary studies of related gene structures and gene families.

Availability and implementation: Bio-Draw-FeatureStack and Bio-Graphics-glyph-decorated_gene are freely available at the Comprehensive Perl Archive Network (CPAN) and GitHub.

Contact: chenn@sfu.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on June 28, 2012; revised on August 29, 2012; accepted on September 17, 2012

1 INTRODUCTION

Comparative analysis of gene structures is important for understanding gene function and evolution. To facilitate gene structure comparison, multiple related gene models need to be shown side by side in a single compact image. In addition, sequence features such as protein domains should be highlighted for functional annotation and to provide reference points for comparison (Pain et al., 2008).

Few specialized tools have been developed for the comparative visualization of gene structures, including FancyGene (Rambaldi and Ciccarelli, 2009), GECA (Fawal et al., 2012) and GSDS (Guo et al., 2007). FancyGene provides rich annotation options but is limited to the display of a single genomic locus and image generation cannot be automated. Conversely, GECA and GSDS are freely available at the Comprehensive Perl Archive Network (CPAN) and GitHub.

Here, we present two Perl modules, Bio::Draw::FeatureStack and Bio::Graphics::glyph-decorated_gene, which build upon existing BioPerl (Stajich et al., 2002) and BioGraphics (Stein et al., 2002) functionality for the highly generic and versatile visualization of multiple gene structures. When used in combination, these two modules allow for fully automated and yet highly configurable image generation, which greatly facilitates comparisons of many gene structures.

2 METHODOLOGY

FeatureStack takes an array of BioPerl feature objects as input; projects them onto a common coordinate space; flips features on the negative strand (option—flip_minus), removes untranslated regions (option—ignore_utr); left-aligns them by start codon, protein domain or a user-defined offset (option—feature_offsets); sets a fixed intron size (option—intron_size); removes unwanted transcripts (option—transcripts_to_skip) and then draws a SVG or PNG image of the so transformed features using a user-specified glyph (option—glyph). Below is a brief synopsis for the use of FeatureStack:

```perl
$feature_stack = new Bio::Draw::FeatureStack
{   -features => \@features # feature array-ref
    -glyph => 'gene',
    -flip_minus => 1,
    -ignore_utr => 1,
    -panel_params => {   # passed on to panel
        -width => 1024,
        -pad_left => 80,
        -pad_right => 20,
        -grid => 1
    },
    -glyph_params => {   # passed on to glyph
        -utr_color => 'white',
        -label_transcripts => 1,
        -description => 1
    } );

$png =$feature_stack->png; # or ->svg
```

Input features can represent BioPerl genes or transcripts with a three-tier (gene→mRNA→CDS/UTR) or two-tier (mRNA→CDS/UTR) level structure, respectively. The way features are retrieved is FeatureStack-independent and can, for example, be achieved using Bio::DB::SeqFeature::Store or Bio::DB::GenBank, both BioPerl modules.

FeatureStack was designed with the goal of providing maximum flexibility in image generation. As such, the user can control the output both via FeatureStack’s own options and by providing panel- and glyph-specific parameters to fine-control all aspects of the rendering process. FeatureStack can be used with
any BioPerl glyph that is compatible with the input features’ structure and is particularly powerful when used in combination with our newly implemented decorated_gene glyph, which installs together with FeatureStack as Comprehensive Perl Archive Network (CPAN) dependency. decorated_gene allows the highlighting and labeling of protein motifs such as signal peptides, transmembrane domains or protein domains on top of gene models, which greatly facilitates the comparison of gene structures. Protein features can be specified in amino acid coordinates and will be automatically mapped to nucleotide coordinates. Please refer to the CPAN module description of decorated_gene for a detailed documentation of glyph options.

Figure 1 showcases the functionality of FeatureStack and decorated_gene on the example of the regulatory factor X (RFX) transcription factor gene family (Chu et al., 2010). Genes were ordered by their phylogenetic distance and automatically aligned horizontally by the start of the DNA-binding domain (shown in black), which represents their most conserved feature. Note that exons were drawn to scale, whereas introns were displayed with a fixed size of 50 bp to accommodate for the large intron size differences between species. By default, FeatureStack draws both exons and introns to scale. Differences in gene structure and features become evident once gene models are displayed with FeatureStack as shown in Figure 1. For example, the DNA-binding domain can be encoded by one to three exons, and the transcription activation domain is only conserved in some human and fly genes.

FeatureStack can also be used (option—alt_feature_type) to display various types of features associated with gene models, such as cis-regulatory elements or genomic variations. Supplementary Figure S1 shows RFX target genes in Caenorhabditis elegans next to their associated X-box motifs. X-box motifs are cis-regulatory elements bound by RFX and are found in the promoters of almost all C.elegans ciliary genes. Typically, X-boxes locate ~50–200 bp upstream of translation start sites. Outliers like nud-1 and dyf-5 that have their X-box motif farther upstream are easily identified from the image.

Finally, we want to emphasize FeatureStack’s usefulness in identifying atypical members of a gene family, pointing towards biologically interesting family members or gene prediction errors (Supplementary Fig. S2).

Additional documentation as well as source code and data files used to produce the three figures in this article are available online at CPAN.

ACKNOWLEDGEMENTS

We thank Western Canada Research Grid and Ata Roudgar for providing computational infrastructure and support. N.C. is a Michael Smith Foundation for Health Research (MSFHR) Scholar and a Canadian Institutes of Health Research (CIHR) New Investigator.

Funding: NSERC of Canada (to N.C.) and the Simon Fraser University Community Trust Endowment Fund through the BCID Project. C.F. was supported by Weyerhaeuser and Simon Fraser University graduate scholarships.

Conflict of Interest: none declared.

REFERENCES