Bioinformatics, 32(1), 2016, 447–449
doi: 10.1093/bioinformatics/btv579
Advance Access Publication Date: 10 October 2015
Applications Note

OXFORD

Sequence analysis

I-PV: a CIRCOS module for interactive protein sequence visualization

Ibrahim Tanyalcin[1,2,*], Carla Al Assaf[3], Alexander Gheldof[1], Katrien Stouffs[1,4], Willy Lissens[1,4] and Anna C. Jansen[5,2]

[1]Center for Medical Genetics, UZ Brussel, Brussels, Belgium, [2]Neurogenetics Research Group, Vrije Universiteit Brussel, Brussels, Belgium, [3]Center for Human Genetics, KU Leuven and University Hospitals Leuven, 3000 Leuven, Belgium, [4]Reproduction, Genetics and Regenerative Medicine, Vrije Universiteit Brussel, Brussels, Belgium and [5]Pediatric Neurology Unit, Department of Pediatrics, UZ Brussel, Brussels, Belgium

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on May 5, 2015; revised on September 14, 2015; accepted on October 2, 2015

Abstract

Summary: Today's genome browsers and protein databanks supply vast amounts of information about proteins. The challenge is to concisely bring together this information in an interactive and easy to generate format.

Availability and implementation: We have developed an interactive CIRCOS module called i-PV to visualize user supplied protein sequence, conservation and SNV data in a live presentable format. I-PV can be downloaded from http://www.i-pv.org.

Contact: ibrahim.tanyalcin@i-pv.org, itanyalc@vub.ac.be or support@i-pv.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Today's genome browsers and protein databanks supply vast amount of information about both the structural annotation and the single nucleotide variants (SNV) in genes. The challenge is to concisely bring together this information in an interactive and easy to generate format. Several freeware and commercial visualization tools are available nowadays. CIRCOS is one of these tools which is largely adopted and utilized in the field of genomics (Krzywinski *et al*., 2009). It generates clear and human readable data visualization in a circular layout. Moreover, due to its circular layout, it has a much higher compression rate of space required per datum compared to classic rectangular representations. However, images generated in CIRCOS are static and requires a specific input file format. We have developed an interactive CIRCOS module called i-PV to visualize user supplied protein sequence, conservation and SNV data while significantly easing and automating input file requirements and generation. All elements rendered in i-PV are interactive with mouse-over explanations and clickable buttons. It allows selective display/hiding of amino acid sequence, SNV, conservation and amino acid properties. This may be very beneficial for having novel insights into one's protein of interest (POI) and convey the information efficiently for researchers. I-PV can be downloaded from http://www.i-pv.org.

2 Methods

To use i-PV, only four text files (with '.txt' extension) have to be supplied to the software: conservation scores, protein and cDNA sequences and SNVs/Indels files (Supplementary Fig. S4). Protein and cDNA (or mRNA) sequence files are supplied in fasta format whereas SNP/Indel files are provided as annotated variant call format (vcf) file. The conservation scores are simply array of numbers separated by newline characters. The input files are supplied to i-PV, data are automatically checked for errors or duplicates and matched against the user provided fasta files, and then an interactive html file containing the graph is automatically generated as shown in Figure 1.

I-PV rendering works without problems in chrome 37.0. Minor browser side bugs about rendering SVG elements exist in firefox 32.0.3, explorer 11 and safari 5.1.7.

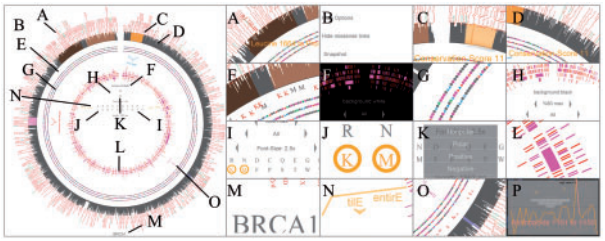

Fig. 1. Overview of i-PV features. (**A**) SNVs with mouse over explanation and automatic generated dbSNP links (red: Non-synonymous, green: Synonymous, gray: Not validated). (**B**) Console can be hidden for publication quality image. (**C**) Domains are colored based on user preference. (**D**) Conservation data from user generated alignment with mouse over information. (**E**) The user can define which amino acids to be shown on the sequence track. (**F**) Switch the color of the background to black. (**G**) Amino acids are plotted and split into 5 main categories (nonpolar: gray circle, polar: magenta circle, negative: blue triangle, positive: red triangle, aromatic: green hexagon). (**H**) Adjustable conservation score threshold to display regions above a certain percentage of maximum conservation score. (**I**) Font-size of chosen amino acids can be adjusted. (**J**) User selectable amino acids to be displayed. (**K**) Up to 17 different amino acid properties can be chosen to be displayed from drop-down menu. (**I**) Tile track showing SNVs and indels (red: SNVs, magenta: Indels, gray stroke: Not validated, black: collapsed due to over display). (**M**) Gene Name. (**N**) Buttons for mass selection of amino acids. (**O**) User defined regions are marked with custom name tag and mouse over information. (**P**) Meta-analysis of amino acid distributions. This information is only displayed in case of single amino acid comparisons. The log2 ratios are capped between −3 and 3. The maximum and the minimum blosum62 scores are −4 and 11. Since the blosum62 matrix is diagonally symmetric, the absolute value of the log ratios are mapped to this range and a p-value is indicated based on how close the two scores are

3 Results

In this article we demonstrate the graphs generated by i-PV for BRCA1, CALR and TUBA1A. BRCA1 (NM_007294, NP_009225) is a DNA repairing protein that plays role in type I breast cancer (Venkitaraman, 2011). In Supplementary Figure S1b, the outermost track shows reported SNVs from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). Users can click on the automatically generated links to the dbSNP website to have more information about these SNVs. To have an overview of the distribution of SNVs and Indels, users can refer to the innermost track.

CALR is a protein (NM_004343, NP_004334), which is mutated in myeloproliferative neoplasms (Trifa *et al.*, 2015). Supplementary Figure S1a shows a partial screen shot of CALR with the minimum and maximum range of conservation scores dynamically adjusted and plotted so that even when alignments from closely related species are used, small differences in conservation scores can be observed. Supplementary Figure S1a shows the functionally important C-terminal region of CALR.

TUBA1A is a protein (NM_006009.3, NP_006000.2) that, if mutated, can lead to type I lissencephaly (Romaniello *et al.*, 2014). Supplementary Figure S2 shows a partial screen shot of TUBA1A and the highly conserved GTP binding region amongst the species. This region is enriched in glycine, serine and threonine residues (sequence track, colored letters) with few surrounding amino acids with aromatic side chains (green hexagons).

It is often necessary to know the distributions of different amino acids along a protein of interest. For instance, you may want to see distribution of specific amino acid(s) against some other set of user defined amino acid(s). Moreover, users can choose to display the amino acid information with window sizes varying from 10 to 100 residues. All data points belonging to these windows are highlighted with values upon mouse over events. In addition to amino acid distributions, in case of single amino acid comparisons, a blosum62 agreement graph is also shown telling the user if the current distribution is in agreement with the matrix score of that amino acid pair (Wei *et al.*, 1997). Supplementary Figure S3 illustrates such example taken from TUBA1A.

I-PV also comes with tools to extract sequence, conservation, specific features (user selected amino acids with different properties) and SNVs between user defined regions in a formatted text layout with user defined number of items per line. Feature extraction is done simply by clicking on the conservation track and selecting the extract option.

4 Discussion

Many sequence visualization tools focus on certain aspects of proteins such as conservation, variations, sequence alignments or topology. For instance I-COMS (Iserte *et al.*, 2015) focus on multiple sequence alignments and recruits links with 3 outer tracks. Likewise, Circoletto (Darzentas, 2010) parses blast results and visualizes sequence similarity with interactivity at the level of web server. Another software package, ggbio (Yin *et al.*, 2012) implements circular visualization for genomics and high throughput data using R programming language. Following a different approach, RCircos (Zhang *et al.*, 2013) directly integrates Circos into R for 2D plots. Although more flexible than i-PV, both tools are limited in providing user interaction. While all aforementioned tools are very useful in their own right, we pursued a more interactivity based design. Therefore, i-PV is not solely designed for visualization but also for live presentable graphs and information that can selectively be displayed and customized. I-PV combines major sources of information under one html file that is easy to generate and share on both desktop and mobile environments. Some visualization tools come with mouse over or on-click explanations; however, since they are based on image files, there are limited user-driven changes that can be performed on the image. Due to the same reason, images from many sequence visualization tools are limited in diversity whereas in i-PV, the user can entirely customize the image since she/he is in charge of the input files, even after output generation. I-PV does not use an image map like most conventional visualization tools; rather it links data to scalable vector graphics (svg) elements which makes sequence and feature extraction possible with a few mouse clicks. The software packages that are used in the making of i-PV, have their own active community such as CIRCOS and D3 (Bostock *et al.*, 2011) which also makes i-PV a live framework where new features can be added. Last but not least, many visualization tools are based on rectangular-scroll based representation of information which does not deliver a 'wide angle' view of the sequence data unlike circular visualization. However, as like all other types of visualizations, there are also limitations for circular graphs when it comes to conveniently zoom in to a particular region or visually align tracks with different radii. We intend to further develop this software with several other features based on end user needs. The current version of i-PV can be downloaded from http://www.i-pv.org.

Funding

I.T received funding from the Scientific Fund Willy Gepts (grant number: 71074) and the Stichting Marie Marguerite Delacroix.

Conflict of Interest: none declared.

References

Bostock,M. *et al.* (2011) D3: Data-Driven Documents. *IEEE Trans. Vis. Comp. Graph. (Proc. InfoVis)*, **17**, 2301–2309.

Darzentas,N. (2010) Circoletto: visualizing sequence similarity with Circos. *Bioinformatics*, **26**, 2620–2621.

Iserte,J., *et al*. (2015) I-COMS: Interprotein-COrrelated Mutations Server. *Nucleic Acids Res*., **43**, W320–W325.

Krzywinski,M., *et al*. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*., **19**, 1639–1645.

Romaniello,R. *et al*. (2014) Brain malformations and mutations in α- and β-tubulin genes: a review of the literature and description of two new cases. *Dev Med Child Neurol*, **56**, 354–360.

Trifa,A.P. *et al*. (2015) CALR versus JAK2 mutated essential thrombocythaemia—a report on 141 patients. *Br. J. Haematol*., **168**, 151–153.

Venkitaraman,A.R. (2011) Cancer: Let sleeping DNA lie. *Nature*, **477**, 169–170.

Wei,L. *et al*. (1997) Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences. *Pac. Symp. Biocomput*., 465–476.

Yin,T. *et al*. (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol*., **13**, R77.

Zhang,H. *et al*. (2013) RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics*, **14**, 244.