BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2620–2621
doi:10.1093/bioinformatics/btq484

Sequence analysis

Advance Access publication August 24, 2010

Circoletto: visualizing sequence similarity with Circos

Nikos Darzentas

Institute of Agrobiotechnology, Centre for Research and Technology Hellas, Thessaloniki 57001, Greece

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: We present Circoletto, an online visualization tool based on Circos, which provides a fast, aesthetically pleasing and informative overview of sequence similarity search results.

Availability and implementation: Online version and downloadable software package for offline use (source code in PERL) freely available at http://bat.ina.certh.gr/tools/circoletto/

Contact: ndarz@certh.gr

Received on July 21, 2010; revised on August 16, 2010; accepted on August 17, 2010

1 INTRODUCTION

The calculation of sequence similarity between two biological sequences is as old a procedure as the field of bioinformatics. The Basic Local Alignment Search Tool, or BLAST (Altschul *et al*., 1990), is a celebrated and highly popular method for the task. However, and partly because the heritage of such tools goes back to very basic visualization technologies and monochromatic terminals, partly because speed and efficiency are still paramount, these tools do not come with visualization capabilities. Yet, there is great need for intuitive ways to quickly understand their output, and interactive solutions have so far included linearly stacked pairs of ideograms (Gollapudi *et al*., 2008), and highly efficient graph layouts of large numbers of nodes representing sequences and edges representing similarity (Frickey and Lupas, 2004). In this context, Circoletto has been developed to complement such methods in a novel way.

2 METHODS

Circoletto is innovative in bringing together BLAST and Circos (Krzywinski *et al*., 2009), a rich, flexible and aesthetically pleasing visualization suite written in Perl, which, as the name suggests, is based on a circular organization of information, resembling the way bacterial genomes have been depicted for many years. Circular representations have desirable traits for data visualization, discussed in (Krzywinski *et al*., 2009) and the Circos site (http://mkweb.bcgsc.ca/circos/intro/circular_approach/), and which include circumventing link crossover in stacked or linearly placed ideograms, avoiding potentially very dense and complex graphs (the so-called 'hairball' effect), and better use of more available space.

Circoletto has been developed to be user-friendly and efficient, making the procedure as straightforward as possible—to that effect, it is also served by a small-scale BLAST server. It accepts either two FASTA-formatted sequence files for the BLAST run, one containing the query sequences and the other the database sequences, or a pre-computed BLAST output in the default pairwise format. It allows the user to choose between preset *E*-values (from relaxed to strict, and only applicable if the user has provided sequence files), the Circos output format (between the fast, lightweight, lower quality PNG and its contrary, the SVG) and size. Additionally, the user can decide whether only the best local alignment per query will be shown, and whether Circos should attempt to untangle the ribbons by potentially re-ordering queries and database sequences. Importantly, for the better understanding of the sequence relations to be shown, the user can optionally load a custom annotation file which can contain information to colour the entries (e.g. into functional groups), and mark coordinate-based domains on the ideograms.

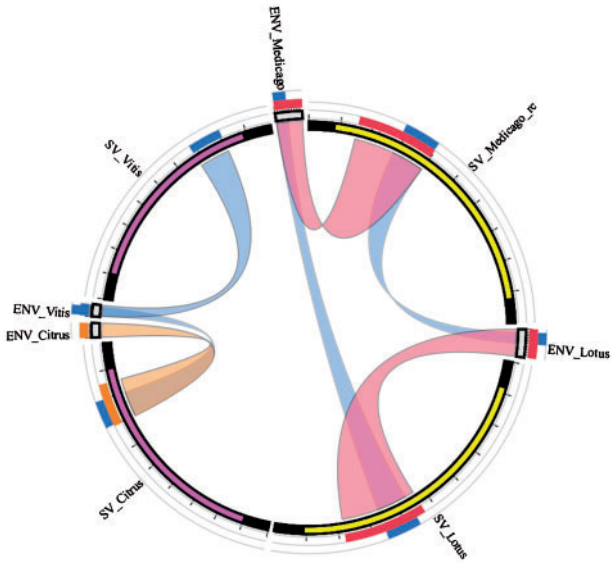

Fig. 1. A relatively simple but all-inclusive example of output from Circoletto. Queries [Envelope (ENV) genes from Sirevirus (SV) retrotransposons] are protruding short grey ideograms, while database entries (full-length Sireviruses) have been divided into two groups (yellow and purple) and their long terminal repeats (LTRs) marked black with the optional annotation file. Note the twist of the two ribbons reaching the *Medicago* (artificially) reverse complementary sequence. The relatedness of *Lotus* and *Medicago* on one side (the left), and *Citrus* and *Vitis* (less so) on the other, are clear to see.

Upon submission, the provided data are processed, all the necessary files for Circos are produced, and Circos is ran, while the user is being kept informed. Finally, links to the visualization and either to two versions of the Circoletto-calculated BLAST results, in tabulated and in HTML formats, or to the user-provided BLAST output, are produced.

Several checks and thresholds are in place to control the complexity of the output, since too much information can render the process too slow and uninformative. These include limits to the number and size of sequences, and the number of local alignments produced—the latter produces an alternative path, with Circoletto attempting to reduce the number of links (ribbons in this case) it will be asked to visualize by keeping top-scoring local alignments only.

3 RESULTS

Figure 1 contains a sample Circoletto output (more can be found here: http://tools.bat.ina.certh.gr/circoletto/examples/). Sequences are placed around a circle, read clockwise, starting at 12 o'clock. Circos has the ability to place the ideograms in such an order so that the ribbons are maximally untangled, which is why queries and database entries can be intertwined (Fig. 1). The ribbons represent the local alignments produced by BLAST, their width the alignment length, and the colours the alignment bitscore in four quartiles: blue for the first (i.e. worst) 25% of the maximum bitscore, green for the next 25%, orange for the third, and finally red for the top (i.e. best) bitscores of between 75% and 100% of the maximum bitscore. Black outlined ribbons, preferentially placed on top of others, indicate the best scoring local alignment for the corresponding query against the database, while a twisted ribbon means that the local alignment is inverted (one sequence is reverse complementary of the other). Since the output can be rather complex, we provide stacked histograms on top of the ideograms, representing the frequency (simple count) and score (by aforementioned colour) of the ribbons at each point. Finally, an optional annotation file can divide the sequences into groups by user-defined colours, and guide the marking of domains (Fig. 1).

4 CONCLUSIONS AND OUTLOOK

To the best of our knowledge, Circoletto fills a gap in the visualization of sequence similarity results, and has been warmly received by laboratory scientists in our Institute. It by no means cancels the need for a careful look at the raw data, but we believe it provides an essential first glimpse at sequence relationships. We will be expanding its capabilities as its usage provides more feedback, ensuring it continues to support research.

ACKNOWLEDGEMENTS

The author wishes to thank the Bioinformatics Analysis Team and other scientists at the Institute of Agrobiotechnology for the motivation, their extensive feedback and ideas.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics,* **20**, 3702–3704.

Gollapudi,R. *et al*. (2008) BOV—a web-based BLAST output visualization tool. *BMC Genomics*, **9**, 414.

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