preprint
preprint
Applications Note

Sequence Analysis

FlexiDot: Highly customizable, ambiguity-aware dotplots for visual sequence analyses

Kathrin M. Seibt, Thomas Schmidt * and Tony Heitkam *

Institute of Botany, TU Dresden, Dresden, 01277, Germany.

*To whom correspondence should be addressed.

Abstract

Summary: FlexiDot is a cross-platform dotplot suite generating high quality self, pairwise and all-against-all visualizations. To improve dotplot suitability for comparison of consensus and error-prone sequences, FlexiDot harbors routines for strict and relaxed handling of ambiguities and substitutions. Our shading modules facilitate dotplot interpretation and motif identification by adding information on sequence annotations and sequence similarities. Combined with collage-like outputs, FlexiDot supports simultaneous visual screening of large sequence sets, enabling dotplot use for routine analyses.
Availability and implementation: FlexiDot is implemented in Python 2.7. Software and documentation are freely available at http://github.com/molbio-dresden/flexidot.
Contact: thomas.schmidt@tu-dresden.de, tony.heitkam@tu-dresden.de
Supplementary information: Using test sequences and application use cases, we illustrate all FlexiDot features in the Suppl. Data.

1 Introduction

First described five decades ago (Gibbs and McIntyre, 1970), dotplots remain among the most powerful and effective tools for explorative sequence investigation, still conveying key messages in current research (Hosaka *et al*., 2017). Dotplots allow characterization of complex or repetitive sequences, enable visual detection of DNA structural motifs, and support the identification of modular similarities between sequences.

Despite advances in dotplot algorithms and availability of different software tools, essential features are missing and, if available, scattered across various tools (Table 1, Suppl. Data). This prompted us to combine established functionalities (e.g. all-against-all modes) with new features for dotplot improvement (base ambiguity handling, shading, integration of annotations), while retaining usability and customizability.

2 Features and implementation

FlexiDot is a multi-purpose dotplot suite for publication-ready dotplots, handling self, pairwise and all-against-all comparisons with individual and combined visualizations (Fig. 1 A-C, see Suppl. Data for details). We want to highlight that (1) our mismatch and ambiguity handling enables analyses of degenerate consensus sequences and error-prone long reads (Fig. 1 B), and that (2) our sequence similarity and annotation-based shadings for self and all-against-all representations (Fig. 1 A and C, respectively) convey descriptive information to facilitate sequence interpretation.

The FlexiDot algorithm identifies matches, transforms them into diagonals and creates clear vector images (pdf, svg) or standard raster graphics (png). Less stringent matching is possible by addressing

Table 1. Feature list of commonly used dotplot tools.

Tool	Ambiguity handling	Annotation shading	All-against-all mode	Batch analyses	Interactive GUI	Input: DNA-DNA	Input: DNA-protein	Input: protein-protein	Multiple output formats	Reverse complement	Self/pairwise collages	Similarity shading	Strict/relaxed matching	Citation
FlexiDot	+	+	+	+	-	+	-	+	+	+	+	+	+/+	here
Dotmatcher	-	-	-	+	-	+	-	+	+	-	-	-	-/+	[1]
Dotter	-	-	+	-	+	+	+	+	-	+	-	-	+/+	[2]
Dottup	-	-	-	+	-	+	-	+	+	-	-	-	+/-	[1]
Gepard	-	+	+	-	+	+	-	+	-	+	-	-	+/-	[3]
PolyDot	-	-	+	+	-	+	-	+	+	-	-	-	+/-	[1]
YASS webserver	+	-	-	-	+	+	-	+	-	+	-	-	+/+	[4]

GUI: Graphical user interface, [1]: Rice *et al*. (2000), [2]: Sonnhammer and Durbin (1995), [3]: Krumsiek *et al*. (2007), [4]: Noé *et al*. (2005)

ambiguous residues specifically or by allowing a defined number of substitutions. A tabular output with lengths of the longest match (longest common subsequence, LCS) of all sequence pairs is provided.

FlexiDot integrates highly customizable shadings: (1) Self dotplot regions can be highlighted according to their sequence annotation provided as general feature file (Fig. 1 A). (2) All-against-all comparisons can be shaded according to the LCS length in forward, reverse or both directions (Fig. 1 C). (3) The user can provide a matrix with numerical values (e.g. identities) to guide shading. Matrix values can be displayed in the dotplot.

FlexiDot uses Python 2.7 with numpy, matplotlib, biopython, regex, colormap and colour libraries. It is operated from the command line under Windows, Linux, and Mac. Input sequences are either specified as single or multi-fasta, or automatically detected in the working directory.

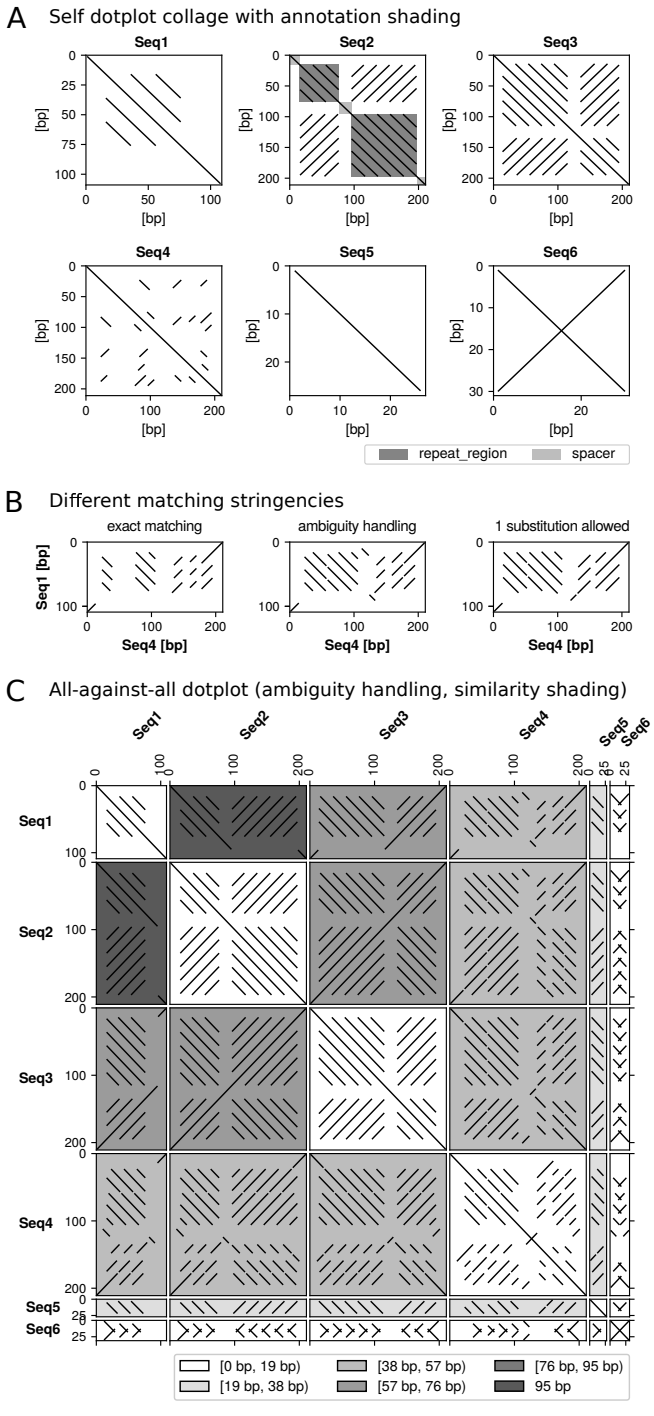

Fig. 1. Visual sequence comparison by FlexiDot with window size 10 using six artificial test sequences. (A) Self dotplot collage. The Seq2 dotplot is shaded with custom annotations. (B) Influence of ambiguity and mismatch handling on pairwise dotplots. (C) All-against-all dotplot of the six sequences with ambiguity handling and similarity shading.

3 Application

As demonstrated for a variety of use cases in the Suppl. Data, FlexiDot creates publication-ready figures for complex sequences. This facilitates:

- evaluation of tandem repeat higher order structures of error-prone long reads, e.g. as seen in Sevim *et al*. (2016); Symonova *et al*. (2017),
- combined depiction of sequence structure and functional annotations,
- identification of conserved motifs in related sequences,
- gene or repeat comparisons using degenerated consensus sequences (Weber *et al*., 2013; Schwichtenberg *et al*., 2016),
- analysis of terminal or internal inverted or direct repeats, e.g. for transposable element annotation (Hosaka *et al*., 2017).

Acknowledgements

We sincerely thank Michael Standke for help with the algorithm, as well as Beatrice Weber and Björn Langer for code testing and valuable feedback.

Funding

This work has been supported by the German Federal Ministry of Education and Research (KMU-innovativ-18 grant 031B0224B).

References

Gibbs,A.J. and McIntyre,G.A. (1970) The diagram: a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.*, **16**, 1-11

Hosaka,A., Saito,R., Takashima,K., Sasaki,T., Fu,Y., Kawabe,A., Ito,T., Toyoda,A., Fujiyama,A., Tarutani,Y., Kakutani,T. (2017) Evolution of sequence-specific anti-silencing systems in *Arabidopsis*. *Nature Comm.*, **8**

Krumsiek,J., Arnold,R., Rattei,T. (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, **23**, 1026-1028

Noé,L., Kucherov,G. (2005) YASS: enhancing the sensitivity of DNA similarity search, *Nuc. Acids. Res.*, **33**, W540-W543

Rice,P., Longden,I., Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16** (6), 276-277

Schwichtenberg,K., Wenke,T., Zakrzewski,F., Seibt,K.M., Minoche,A.E., Dohm,J.C., Weisshaar,B., Himmelbauer,H., and Schmidt,T. (2016) Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species. *Plant J.*, **85**, 229-244

Sevim,V., Bashir,A., Chin,C.S. and Miga,K.H. (2016) Alpha-CENTAURI: Assess-ing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics*, **32**(13):1921-1924

Sonnhammer,E.L. and Durbin,R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1-GC10

Symonova,R., Ocalewicz,K., Kirtiklis,L., Delmastro,G.B., Pelikanova,S., GarciaS., and Kovarik,A. (2017) Higher-order organisation of extremely amplified, potentially functional and massively methylated 5S rDNA in European pikes (*Esox* sp.). *BMC Genomics*, **18**, 391

Weber,B., Heitkam,T., Holtgräwe,D., Weisshaar,B., Minoche,A.E., Dohm,J.C., Himmelbauer,H., and Schmidt,T. (2013) Highly diverse

chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration. *Mob. DNA*, **4**, 8