BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 19 2009, pages 2603–2604
doi:10.1093/bioinformatics/btp436

Genome analysis

DiProGB: the dinucleotide properties genome browser

Maik Friedel[1,*], Swetlana Nikolajewa[2], Jürgen Sühnel[1] and Thomas Wilhelm[3,*]

[1]Biocomputing Group, Leibniz Institute for Age Research – Fritz Lipmann Institute, Jena Centre for Bioinformatics, Beutenbergstr. 11, [2]Systems Biology/Bioinformatics Group, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena Centre for Bioinformatics, Beutenbergstr. 11a, 07745 Jena, Germany and [3]Theoretical Systems Biology, Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK

Received on April 7, 2009; revised on June 17, 2009; accepted on July 12, 2009

Advance Access publication July 15, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: DiProGB is an easy to use new genome browser that encodes the primary nucleotide sequence by thermodynamical and geometrical dinucleotide properties. The nucleotide sequence is thus converted into a sequence graph. This visualization, supported by different graph manipulation options, facilitates genome analyses, because the human brain can process visual information better than textual information. Also, DiProGB can identify genomic regions where certain physical properties are more conserved than the nucleotide sequence itself. Most of the DiProGB tools can be applied to both, the primary nucleotide sequence and the sequence graph. They include motif and repeat searches as well as statistical analyses. DiProGB adds a new dimension to the common genome analysis approaches by taking into account the physical properties of DNA and RNA.

Availability and Implementation: Source code and binaries are freely available for download at http://diprogb.fli-leibniz.de, implemented in C++ and supported on MS Windows and Linux (using e.g. WineHQ).

Contact: maikfr@fli-leibniz.de; thomas.wilhelm@bbsrc.ac.uk

Computational genome analysis aims at understanding the information encoded in the genomes. So far this is almost exclusively done by analyzing the character string of the primary sequence (Cline and Kent, 2009). However, the conservation of certain physical properties can be more important than conservation of the nucleotide sequence itself, especially for non-coding DNA (Babbitt and Kim, 2008). It has been shown that sequence-constraint algorithms often fail to identify non-coding functional elements because these methods neglect the 3D structure of DNA. By incorporating hydroxyl radical cleavage patterns which interrogate the solvent accessible surface area of DNA, it was recently found that 12% of the bases in the human genome are evolutionary constrained (Parker *et al.*, 2009). This is twice the fraction detected by common sequence-based algorithms.

We have developed the new genome browser DiProGB, which considers physico-chemical properties of nucleotide sequences. It helps to detect functionally relevant motifs that cannot be found by analyzing the primary nucleotide sequence alone. More specifically, DiProGB encodes a DNA or RNA sequence by thermodynamical or geometrical dinucleotide properties. In addition, character-based sequence information such as GC or purine content is also encoded on the dinucleotide level. Plotting the dinucleotide property values versus the sequence position leads to a graphical representation of the sequence that we call a sequence graph (Fig. 1). The rationale for exploiting dinucleotide properties is the widely accepted nearest neighbor model saying that thermodynamic properties of nucleic acids can be understood and predicted by considering dinucleotide contributions (Turner, 1996; Zimm and Bragg, 1959). This is also the basis for RNA secondary structure predictions (Yoon *et al.*, 1975). By default, DiProGB offers 10 dinucleotide property sets. Further properties can be downloaded from the dinucleotide property database DiProDB (http://diprodb.fli-leibniz.de) (Friedel *et al.*, 2009) containing >100 of such sets. A corresponding editor enables also manual input of new properties.

DiProGB allows loading sequence and annotated feature information as GenBank or FASTA files and from corresponding feature files (e.g. GFF or PTT) (Leonard *et al.*, 2007). It is also possible to open multiple FASTA files or raw sequence data, and there is an option for downloading sequence information from the NCBI web site http://www.ncbi.nlm.nih.gov/.

All annotated features such as genes, exons, introns or repeat regions and the corresponding qualifiers such as gene name, product and function can be separately addressed and specifically colored. All or parts of the annotated information can be displayed for either a single strand or for both strands together. Overlapping features are visualized by stacked bars in the so-called feature graph below the sequence graph (Fig. 1). The sequence graph can be smoothened by a shifting window technique. Using the mouse wheel, the shifting window size as well as the graph amplitude and the zoom status can be changed in real time.

DiProGB offers four lists for information handling: (i) the 'Sequence list' allows switching between pre-selected sequences; (ii) the 'DiPro list' contains the loaded dinucleotide property sets; (iii) the 'Feature list' allows searching for features and qualifiers and highlighting them in the sequence graph; and (iv) the 'Color list' displays all colors used for indicating the annotated information.

DiProGB combines sequence analyses based on both the primary sequence character string and the sequence graph representation.

*To whom correspondence should be addressed.

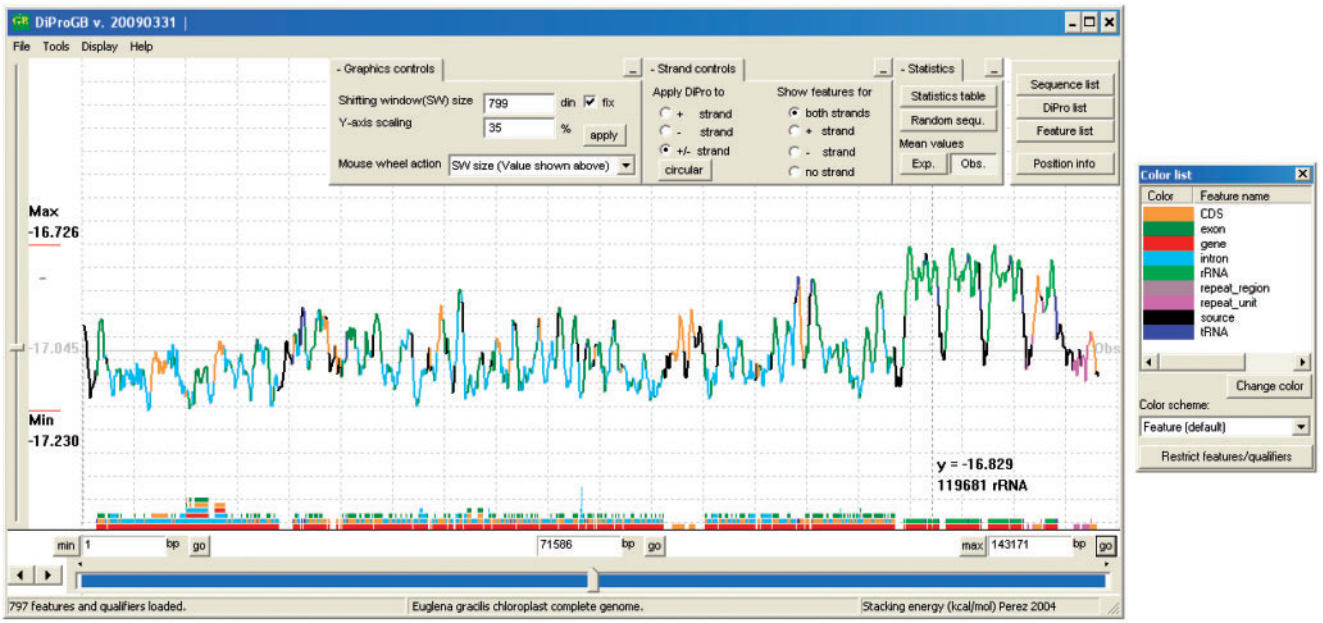

Fig. 1. The full genome of the chloroplast of *Euglena gracilis* is shown in the main window of DiProGB. The sequence graph in the middle is encoded by the physicochemical dinucleotide property stacking energy (Pérez *et al*., 2004) and smoothened applying a shifting window of size 800 nt. The color coding of annotated features on both strands is explained in the color list (right window). Note the significantly different graph shape for the rRNA genes in green. The bottom panel shows the feature graph.

The latter often allows pattern recognition by visual inspection (e.g. identification of large repeats or of certain nucleotide distributions). In addition, DiProGB offers tools for a systematic motif and repeat search for both, the character string and the sequence graph. The motif search algorithms are based on Gusfield's Z-boxes (Gusfield, 1997). The repeat finder searches for maximal and supermaximal repeats. It is based on suffix arrays which is one of the most efficient methods for repeat search (Abouelhoda *et al.*, 2004). DiProGB also offers a fast Fourier transformation for the identification of sequence periodicities.

DiProGB provides two types of statistical analyses. First, the user can calculate mean values for either a partial or the complete sequence. This allows, for example, comparing mean values of physical properties (e.g. entropy) for a given feature (e.g. gene) with the corresponding mean value of the whole genome. Second, there is a so-called position-specific statistics. Here, selected sub-sequences, for example coding sequences, are aligned relative to a specific sequence position (e.g. 100 nt upstream of start) and mean values are calculated for each position in the alignment. The position-specific statistics is a powerful tool for detecting common motifs in annotated features.

DiProGB is a standalone computer program written in VC++. It has been optimized to cope with large genomes. The program has been developed under the Microsoft Windows operating system. It can, however, also be used under Linux, Mac, BSD and Solaris after installing the program WineHQ (http://winehq.org), for example. A more detailed description of DiProGB is available at http://diprogb.fli-leibniz.de .

In summary, DiProGB is a new genome browser for enhanced genome analysis. Its application will lead to deeper insight into organization and functioning of the genome.

ACKNOWLEDGEMENTS

We thank Roman Siddiqui and Vladimir Shelest for testing DiProGB and for important comments.

Funding: Leibniz Graduate School for Aging and Age-related Diseases - LGSA (to M.F.); BBSRC Core Strategic Grant for IFR (to T.W.).

Conflict of Interest: none declared.

REFERENCES

Abouelhoda,M.I. *et al.* (2004) Replacing suffix trees with enhanced suffix arrays. *J. Discrete Alg.*, **2**, 53–86.

Babbitt,G.A. and Kim,Y. (2008) Inferring natural selection on fine-scale chromatin organization in yeast. *Mol. Biol. Evol.*, **25**, 1714–1727.

Cline,M.S. and Kent,W.J. (2009) Understanding genome browsing. *Nature Biotech.*, **27**, 153–155.

Friedel,M. *et al.* (2009) DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.*, **37**, D37–D40.

Gusfield,D. (1997) Algorithms on strings, trees, and sequences. In *Computer Science and Computational Biology.* Cambridge Unvesity Press, pp. 7.

Leonard,S.A. *et al.* (2007) Common file formats. *Curr. Protoc. Bioinformatics*, Appendix 1: Appendix 1B.

Parker,S.C.J. *et al.* (2009) Local DNA topography correlates with functional non-coding regions of the human genome. *Science*, **324**, 389–392.

Pérez,A. *et al.* (2004) The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res.*, **32**, 6144–6151.

Turner,D.H. (1996) Thermodynamics of base pairing. *Curr. Opin. Struct. Biol.*, **6**, 299–304.

Yoon,K. *et al.* (1975) The kinetics of codon-anticodon interaction in yeast phenylalanine transfer RNA. *J. Mol. Biol.*, **99**, 507–518.

Zimm,B.H. and Bragg,J.K. (1959) Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.*, **31**, 526–535.