BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 18 2003, pages 2477–2479
DOI: 10.1093/bioinformatics/btg334

gff2aplot: Plotting sequence comparisons

Josep F. Abril[1,*], *Roderic Guigó*[1] *and Thomas Wiehe*[2,†]

[1]Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica (IMIM) Universitat Pompeu Fabra (UPF)—Centre de Regulació Genòmica (CRG), Passeig Marítim de la Barceloneta 37–49, 08003 Barcelona, Catalonia, Spain and [2]Freie Universität Berlin, Arnimallee 22, 14195 Berlin, Germany

Received on March 4, 2003; revised on June 11, 2003; accepted on June 20, 2003

ABSTRACT

Summary: gff2aplot is a program to visualize the alignment of two sequences together with their annotations. Input for the program consists of single or multiple files in GFF-format which specify the alignment coordinates and annotation features of both sequences. Output is in PostScript format of any size. The features to be displayed are highly customizable to meet user specific needs. The program serves to generate print-quality images for comparative genome sequence analysis.

Availability: gff2aplot is freely available under the GNU software licence and can be downloaded from the address specified below.

Contact: jabril@imim.es

Supplementary information: http://genome.imim.es/software/gfftools/GFF2APLOT.html

An often occurring task in comparative sequence analysis is to suggestively display a pairwise alignment, possibly together with domain annotations for one or both sequences. Some well-known programs are Dotter (Sonnhammer and Durbin, 1995), PipMaker (Schwartz *et al.*, 2000), VISTA (Mayor *et al.*, 2000) or Laj (Wilson *et al.*, 2001). While the first tool is suited to interactively explore the site by site comparison of two sequences without annotations, the others produce a one-dimensional projection of a pairwise or a multiple alignment, along with the annotation features. In all these cases, however, the visualization tools are intrinsically tied to a specific underlying alignment algorithm. We have developed the program gff2aplot to generate two-dimensional annotated alignment plots in PostScript format. gff2aplot is not tied to a particular alignment algorithm, but rather can be used as a visualization filter after running some independent alignment tool. In this regard gff2aplot is related to Alfresco (Jareborg and Durbin, 2000), but while Alfresco is oriented towards highly interactive use and has limited printing capabilities, gff2aplot is intented for producing high quality printed images. The strategy used in gff2aplot is very similar to that employed in gff2ps (Abril and Guigó, 2000), a tool to visualize annotations of genomic sequences obtained from different sources. User may parse alignment segments from any of the current similarity search tools, and combine them if desired in the gff2aplot output. We provide several such filters from the gff2aplot website, to parse, for instance, NCBI-BLAST (Altschul *et al.*, 1997), WU-BLAST (W. Gish, 1996–2003, http://blast.wustl.edu), SIM (Huang and Miller, 1991), MUMMER (Delcher *et al.*, 1999) or BLAT (Kent, 2002). Integrating data will improve the information we obtain about pairs of genomic sequences. We distinguish records containing annotation features and those defining alignment segments. One or more ASCII input data files in GFF-format (see gff2aplot manual) can be processed in a single run. The image produced has a standard layout: it consists of two panels, placed above each other. The upper one displays the alignment of two sequences by means of a rectangular matrix. Sequence annotations are displayed along the top and left edges. The optional lower one contains vertical projections of the aligned fragments in the upper panel and displays their alignment scores or match percentages as in a PiPplot (see examples from Fig. 1). Sequence coordinate tags are shown on the lower and right edges of the panel frames. Projections of the annotated features can be shown under the alignment segments to highlight relationships between them. As in gff2ps, gff2aplot assumes that the input GFF records carry enough formatting information. Thus, in most cases, meaningful output can be obtained using the default settings. Nevertheless, gff2aplot allows for a high degree of customization. Almost any component of the plot can be configured, either through a very flexible customization file (several of such files can be processed for a single plot), or through command-line options (see gff2aplot manual). In particular, users can select any standard printing media size or define their own plot sizes.

gff2aplot is written in PERL and PostScript. It runs on UNIX or Linux platforms and it does not require any special compiler or additional software beyond the installation of Perl, version 5.5 or higher. The program generates a PostScript output file which can be viewed or printed with

[*]To whom correspondence should be addressed.

[†]Current Address: Universität zu Köln, Institut für Genetik, Weyertal 121, 50931 Köln, Germany.

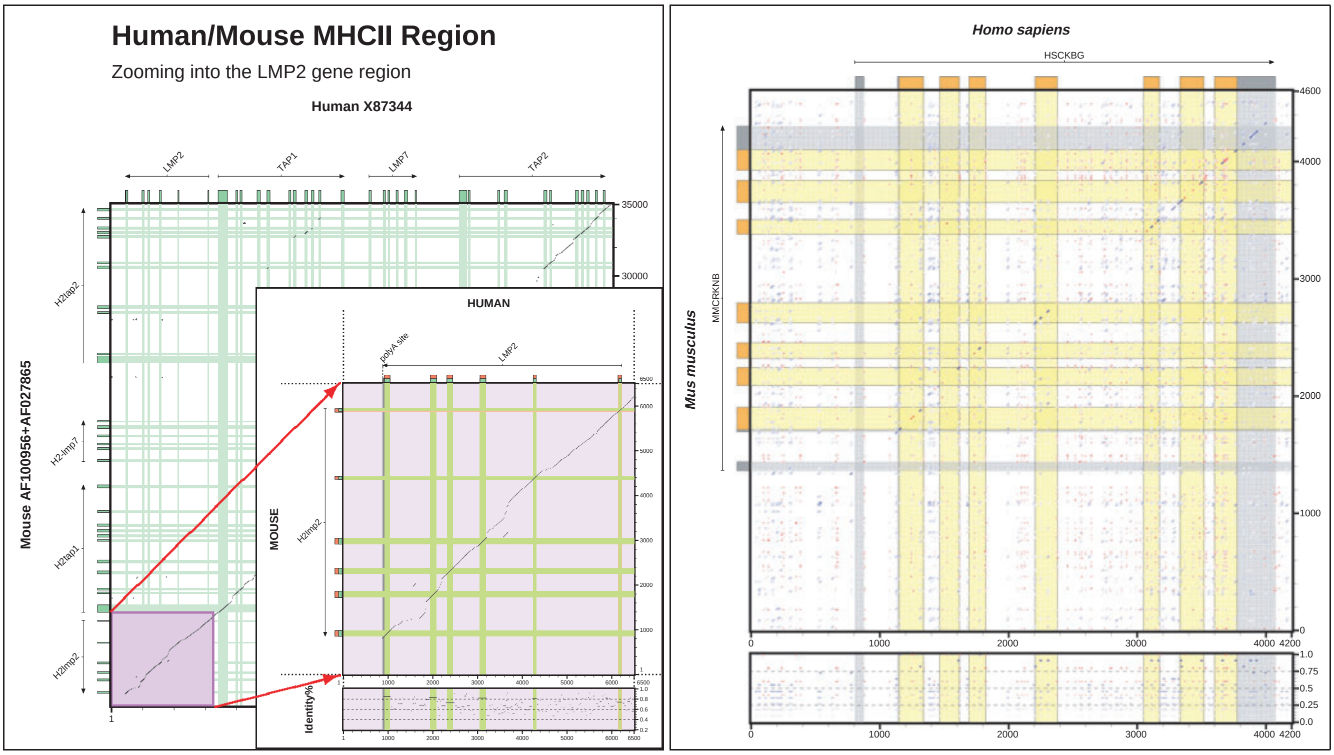

Fig. 1. (Left panel) Comparative analysis of a syntenic genomic region between human and mouse, extending across several genes (MHC II region, Accession Numbers shown on annotation axes main labels). Alignment was obtained using SIM (Huang and Miller, 1991). The pale violet box highlights the region being expanded in the bottom right plot, while red arrows show the corresponding areas on both figures. The region being expanded is the LMP2 human gene region and its counterpart in mouse. Green boxes and projections correspond to CDSs, as annotated in GenBank, while red boxes conform the predicted gene structure by program SGP-1 (Wiehe *et al*., 2001). (Right panel) All possible pairs of potential splice sites on human and mouse homologous sequences were analyzed against the corresponding gene structures, in this case for *creatine kinase B* gene (Accession Numbers X15334 and M74149, for human and mouse, respectively). `gff2aplot` combines here results from two different analysis, red bars correspond to putative donor sites and blue bars to acceptor site ones. All input and parameter files which were used to generate the examples in the figure are accessible from the `gff2aplot` website. Additional examples, as well as a detailed user manual, can also be found there.

any PostScript capable output device. Although PostScript lacks user-interactivity and hyper-link capability, for high-quality images the page description language PostScript has several advantages over bitmap graphics programs. Among these are the free scalability of all plots, the embed-ability of PostScript picture files into text documents (specially those written in LaTeX), the graphics device independence and the robustness with respect to handling large amounts of data. These properties have made `gff2ps` the tool of election to produce, among other applications, the gene content maps of the fly (Adams *et al*., 2000), human (Venter *et al*., 2001), and mosquito (Holt *et al*., 2002) genomes. Like `gff2ps`, the `gff2aplot` program described here is suitable as a filter for high-throughput analysis pipelines. The program has already been applied as a drawing tool for human/mouse comparisons in recent publications (e.g. Parra *et al*., 2003); development versions of `gff2aplot` have already been used in other papers (e.g. Reichwald *et al*., 2000; Wiehe *et al*., 2001). Although initially developed to display sequence similarity relationships, the simplicity and generality of the GFF standard may make `gff2aplot`, through its high customization capabilities, useful to display other matrix-like generic relationships between sequences, for example the splice sites analysis shown in the right panel of Figure 1.

ACKNOWLEDGEMENTS

We would like to thank Matthias Platzer, Jena, for helpful comments while developing the program prototype, and the colleagues at the Genome BioInformatics Laboratory for their extensive testing of `gff2aplot`. J.F.A. is supported by a predoctoral fellowship from the ‘Instituto de Salud Carlos III (Spain)’, 99/9345. This work was also supported by a joint grant from the ‘German Academic Exchange Service (DAAD)’ to TW and the ‘Ministerio de Educación y Ciencia (Spain)’ to RG. Research at the Genome BioInformatics

Laboratory is supported by grant from 'Plan Nacional de I+D (Spain)' to RG, BIO2000-1358-C02-02.

REFERENCES

Abril,J.F. and Guigó,R. (2000) gff2ps: Visualizing genomic annotations. *Bioinformatics*, **16**, 743–744.

Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Altschul,S.F., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman.D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Delcher,A.L., Kasif,S., Fleischmann,R.D., Peterson,J., White,O. and Salzberg,S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **27**, 2369–2376.

Holt,R.A., Subramanian,G.M., Halpern,A., Sutton,G.G., Charlab,R., Nusskern,D.R., Wincker,P., Clark,A.G., Ribeiro,J.M.C., Wides,R. *et al*. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.

Huang,X. and Miller,W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*,**12**, 337–357.

Jareborg,N. and Durbin,R. (2000) Alfresco—A workbench for comparative genomic sequence analysis. *Genome Res.*, **10**, 1148–1157.

Kent,W.J. (2002) Blat—the blast-like alignment tool. *Genome Res.*, **12**, 656–664.

Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S. and Dubchak,I. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046–1047.

Parra,G., Agarwal,P., Abril,J.F., Wiehe,T., Fickett,J.W. and Guigó,R. (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.

Reichwald,K., Thiesen,J., Wiehe,T., Weitzel,J., Strätling,W.H., Kioschis,P., Poustka,A., Rosenthal,A. and Platzer,M. (2000) Comparative sequence analysis of the MECP2-locus in human and mouse reveals new transcribed regions. *Mammalian Genome*, **11**, 182–190.

Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMaker—A web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.

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**, GC1–10.

Venter,C.J., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wiehe,T., Gebauer-Jung,S., Mitchell-Olds,T. and Guigó,R. (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res.*, **11**, 1574–1583.

Wilson,M.D., Riemer,C., Martindale,D.W., Schnupf,P., Boright,A.P., Cheung,T.L., Hardy,D.M., Schwartz,S., Scherer,S.W., Tsui,L.-C., Miller,W. and Koop,B.F. (2001) Comparative analysis of the gene-dense *ACHE/TFR2* region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acid Res.*, **29**, 1352–1365.