BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 10 2012, pages 1398–1399
doi:10.1093/bioinformatics/bts153

Genome analysis

Advance Access publication March 29, 2012

GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families

Nizar Fawal[1,2], Bruno Savelli[1,2], Christophe Dunand[1,2] and Catherine Mathé[1,2,*]

[1]Université de Toulouse, UPS, UMR 5546, Laboratoire de Recherche en Sciences Végétales, BP 42617 Auze ville, F-31326 Castanet-Tolosan, France and [2]CNRS, UMR 5546, BP 42617, F-31326 Castanet-Tolosan, France

Associate Editor: Martin Bishop

ABSTRACT

Summary: GECA is a fast, user-friendly and freely-available tool for representing gene exon/intron organization and highlighting changes in gene structure among members of a gene family. It relies on protein alignment, completed with the identification of common introns in the corresponding genes using CIWOG. GECA produces a main graphical representation showing the resulting aligned set of gene structures, where exons are to scale. The important and original feature of GECA is that it combines these gene structures with a symbolic display highlighting sequence similarity between subsequent genes. It is worth noting that this combination of gene structure with the indications of similarities between related genes allows rapid identification of possible events of gain or loss of introns, or points to erroneous structural annotations. The output image is generated in a portable network graphics format which can be used for scientific publications.

Availability and implementation: Web-implemented version and source code are freely available at https://peroxibase.toulouse.inra.fr/geca_input_demo.php and a detailed example can be found at https://peroxibase.toulouse.inra.fr/geca_instructions.php

Contact: mathe@lrsv.ups-tlse.fr

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

Received on February 6, 2012; revised on March 9, 2012; accepted on March 23, 2012

1 INTRODUCTION

The increasing pace of genome sequencing is providing us with a growing number of sequences, containing a wealth of information that will fuel studies for years to come. One of the numerous resulting fields of investigation is the comparison of annotations and subsequent genic organizations in eukaryotic genomes. The amount of structurally annotated protein genes available has incited the development of web tools such as GSDS (Chuan *et al.*, 2007), FancyGene (Rambaldi and Ciccarelli, 2009). The purpose of these programs is to represent the exon/intron structure of several genes in a single image in order to perform global gene structure comparison.

However, these resources display the gene structures independently, rendering gene comparison difficult and leaving the individual gene structures lacking in information concerning sequence conservation.

In order to accurately compare gene structures, we came to the conclusion that the following data is necessary. The position and level of similarity within a set of sequences is needed to highlight conserved regions. Once identified, a special focus to the conservation degree around introns is essential in order to determine whether they can be considered as conserved introns between paralogs/orthologs.

Since, to our best knowledge, no tool currently exists that combines this information into a single output, we decided to develop our own tool, GECA. Its strategy relies on a simple observation: by aligning genes using the position of common introns shared by related sequences, we align the surrounding exons of the respective genes. Thus, GECA fully relies on the output of CIWOG (Wilkerson *et al.,* 2009), a freely available software that detects common introns (named *cintrons*) in a set of related protein-coding genes. In order to provide the protein alignment file required by CIWOG, GECA currently launches MAFFT (Katoh *et al.,* 2002) but any multiple alignment program with a ClustalW-like output is suitable.

Once the protein sequences are aligned and the common introns identified, GECA is able to produce a main graphical representation: schematic gene structures, anchored by their first common intron position, overlaid with indications highlighting similar content between pairs of genes.

2 FEATURES

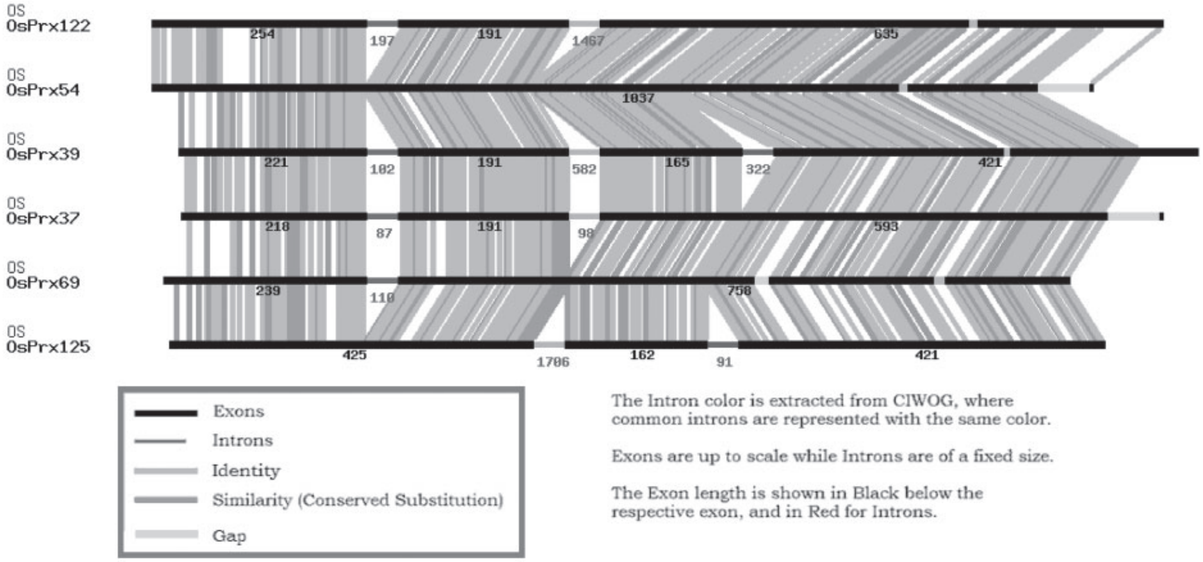
2.1 Input data format

Protein and genomic sequences in FASTA format together with gene structures in GenBank feature format are required to execute GECA. A specific FASTA header is needed and must be identical for all three. It should be in the form of '>AccessionID | sequence name'.

2.2 GECA representation of aligned gene structures

Once the user-supplied data is uploaded, GECA uses PERL's GD and GD::Text::Align libraries to draw the gene structures. The intron/exon organizations of subsequent genes are aligned using their first common intron. Exons are to scale, whereas information concerning actual length of introns, that all share the same size in this representation, is given by a color code used by CIWOG. To display the similarities between the sequences, GECA uses the protein alignment produced by MAFFT. The similarities between subsequent sequences in the alignment are represented at the level of amino acids in the translated exons. Two amino acids are linked by a blue line if they are identical and a purple line indicates conservative

*To whom correspondence should be addressed.



The Intron color is extracted from CIWOG, where common introns are represented with the same color.

Exons are up to scale while Introns are of a fixed size.

The Exon length is shown in Black below the respective exon, and in Red for Introns.

Fig. 1. Screen shot of a result of GECA.

substitutions (default matrix is BLOSUM62 but BLOSUM45 and 80 are also available). Gaps longer than five amino acids in the alignments are displayed in gray. Finally, the exon and intron sizes are calculated from the genomic coordinates and are displayed under the exons and introns in black and red, respectively (Fig. 1).

Alongside its main and original graphical output, GECA also simultaneously provides two complementary outputs: (i) a classical, GSDS-like, gene structure display where exons and introns are scaled relatively to their length and cintrons are visible with gene structures aligned on the first common intron and (ii) a tab-delimited table containing introns' IDs for each gene as well as lengths and phases.

2.3 Web-server and standalone package

The idea of developing GECA came from our specific interest in a set of large multigenic families, the peroxidases. The evolution of these families is not yet elucidated and new tools to understand their evolutionary history was needed. Thus, GECA is available as a tool from our family database, the PeroxiBase (Koua *et al*., 2009). The PERL CGI::LITE server is accessible in the PeroxiBase environment (http://peroxibase.toulouse.inra.fr/) as a further analysis after a BLAST or a multicriteria search.

An independent demo version, accessible from the PeroxiBase website, provides scientists with a user-friendly interface to GECA (https://peroxibase.toulouse.inra.fr/geca_input_demo.php). From this page, a package providing all PERL scripts for local installation and execution is available under open source license. GECA is controlled by a single program, that can be customized using a configuration file and requires pre-installation of PERL (minimum version tested 5.8.8), MAFFT and CIWOG.

Aligning exon/intron structures accompanied with the similarities between sequences and common introns information is very helpful while manually checking structural annotation. Moreover, it will provide major information about gene evolution, such as gain and loss of introns, and will bring new clues for the intron origin debate.

ACKNOWLEDGEMENTS

The Demo version of GECA is hosted by the Toulouse Midi-Pyrénées bioinformatics platform. The authors thank Alan Walton for the critical reading and English editing.

Conflict of Interest: none declared.

REFERENCES

Chuan,Y. *et al.* (2007) GSDS: a gene structure display server. *Bioinformatics*, **29**, 1023–1026.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Koua,D. *et al.* (2009) PeroxiBase: a database with new tools for peroxidase family classification. *Nucleic Acids Res.*, **37**, D261–D266.

Rambaldi,D. and Ciccarelli,F. (2009) FancyGene: dynamic visualization of gene structures and proteindomain architectures on genomic loci. *Bioinformatics*, **25**, 2281–2282.

Wilkerson,M.D. *et al*. (2009) Common introns within orthologous genes: software and application to plants. *Brief. Bioinform.*, **10**, 631–644.