Bioinformatics, 31(8), 2015, 1302–1304
doi: 10.1093/bioinformatics/btu798
Advance Access Publication Date: 29 November 2014
Applications Note

OXFORD

Sequence analysis

GenePainter v. 2.0 resolves the taxonomic distribution of intron positions

Stefanie Mühlhausen, Marcel Hellkamp and Martin Kollmar*

Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on August 20, 2014; revised on November 12, 2014; accepted on November 25, 2014

Abstract

Summary: Conserved intron positions in eukaryotic genes can be used to reconstruct phylogenetic trees, to resolve ambiguous subfamily relationships in protein families and to infer the history of gene families. This version of GenePainter facilitates working with large datasets through options to select specific subsets for analysis and visualization, and through providing exhaustive statistics. GenePainter's application in phylogenetic analyses is considerably extended by the newly implemented integration of the exon–intron pattern conservation with phylogenetic trees.
Availability and implementation: The software along with detailed documentation is available at http://www.motorprotein.de/genepainter and as Supplementary Material.
Contact: mako@nmr.mpibpc.mpg.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The rapid growth of sequenced eukaryotes increasingly allows incorporating exon–intron pattern information into phylogenetic analyses. Conserved intron positions have helped resolving the relationship of taxonomic lineages (Grewe *et al.*, 2013; Lehmann *et al.*, 2013) and have also extensively been used to reveal subfamily relationships within larger gene families (Du *et al.*, 2013; Yan *et al.*, 2014) and to reconstruct ancient genes (Odronitz and Kollmar, 2008). Because the early eukaryotes were rather intron rich and the gene structure evolution towards the extant species was dominated by intron-loss events (Koonin *et al.*, 2013), considerable taxonomic and sequence sampling is necessary to reconstruct the gene structure history of a gene or gene family across the eukaryotic tree of life. For example, mammals and plants are known to have intron-rich genomes, while most unicellular organisms are intron poor. However, although the yeast *Saccharomyces cerevisiae* is known to contain only ~275 introns in total (Lopez and Séraphin, 2000), the yeast's dynein light chain LC8 coding region is split by two introns (other ascomycetous fungi contain even up to seven introns) while the human LC8 orthologs only contain a single intron. In general, intron gain events are rare and are most often recent events (Li *et al.*, 2009). Existing tools for the comparison of gene structures like Exalign (Pavesi *et al.*, 2008), CIDA/CIWOG (Wilkerson *et al.*, 2009), GECA (Fawal *et al.*, 2012) and GenePainter (Hammesfahr *et al.*, 2013) compare exon lengths or map intron positions to positions in multiple amino acid sequence alignments (MSA). However, none of the tools integrates gene structure conservation with taxonomic and phylogenetic tree information. Here, we present a new version of our GenePainter software, v. 2.0, which maps intron positions onto the multiple sequence alignment and assigns intron gain and loss events to taxonomic branches and extant species.

2 Features

2.1 Input files

The most accurate way to compare gene structures is to align the intron positions based on corresponding protein sequence alignments (Hammesfahr *et al.*, 2013). GenePainter therefore requires a protein MSA (fasta format) and corresponding gene structures (YAML or GFF v. 3 format). For phylogenetic analyses, GenePainter by default uses the NCBI taxonomy tree as reference tree (Federhen, 2012) but user-provided trees can also be employed.

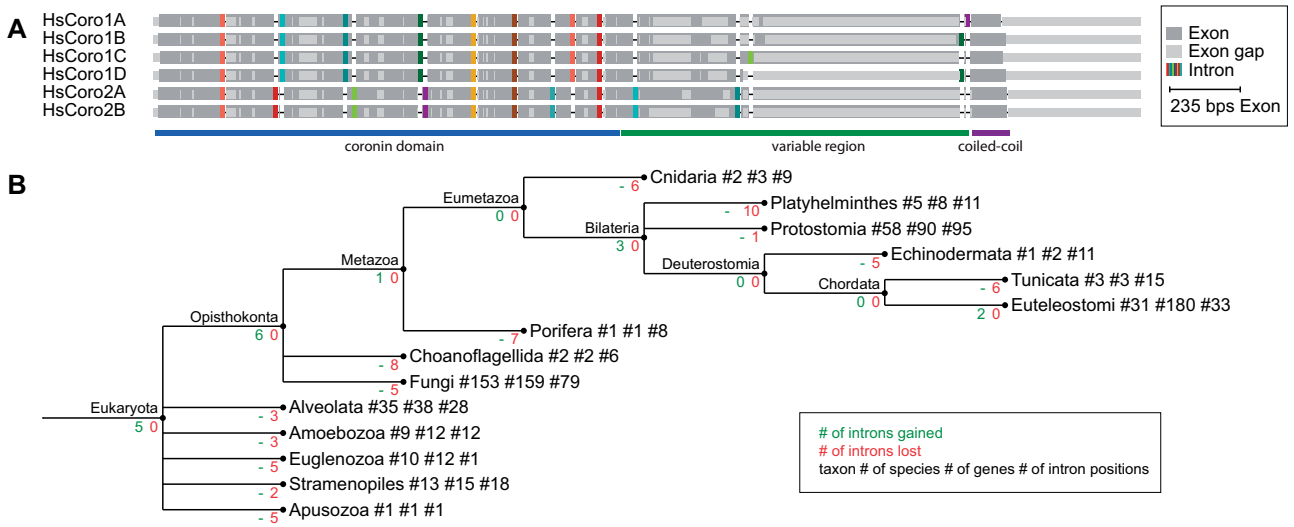

Fig. 1. (**A**) Representation of the human coronin gene structures from an alignment of 526 genes with introns highlighted. Exons are drawn to scale, while introns are drawn length independent in colour. Introns at the same nucleotide position get the same colour. To keep the alignment, exons in those genes not containing an intron are split at the respective position. Alignment gaps are indicated as light grey boxes. For better orientation, the conserved coronin- and coiled-coil domains are indicated, which are separated by a region highly divergent in length and sequence. (**B**) The conservation of the introns in the human coronin genes was analysed against 312 species from most major eukaryotic branches based on the NCBI taxonomy tree. The numbers of introns gained (first number) and lost (second number plotted on branch) in human coronin genes are plotted on each branch. Statistics plotted on taxa refer to full dataset of 312 species (Color version of this figure is available at *Bioinformatics* online.)

2.2 Mapping introns onto taxonomy

For placing intron gain and loss events, GenePainter currently uses a simple maximum parsimony model requiring the least intron-loss events. The model is based on the assumption that it is highly unlikely that introns would have independently been introduced at the exact same sequence position with exact same reading frame in species whose last common ancestor had lived hundreds of million years ago. Thus, introns shared by the genes of two species are regarded as already been present in their last common ancestor and are marked as intron gain events there. If species of sub-branches miss the respective intron, an intron-loss event is marked for this branch.

2.3 Data selection and visualization

Visualization of hundreds of sequences with dozens to hundreds of shared and unique intron positions becomes challenging. Therefore, we implemented several different options for users to reduce the visualized data, while statistics and intron gain and loss event mappings are still computed and provided for the whole dataset.

2.4 Output files

GenePainter v. 2.0 provides a number of output files such as an extended multiple sequence alignment, gene structure alignments including a binary representation for evolutionary analyses, graphical outputs on base-pair and other scales (Fig. 1A), scripts for mapping intron positions onto protein structures at user-provided conservation levels, extensive statistics and an extended phylogenetic tree with intron gain and loss events plotted onto the respective branches as svg (Fig. 1B) and in Newick format.

3 Example: coronin gene structure evolution

Analysis of 526 coronin class-I and class-II genes from 312 different species representing major eukaryotic branches reveals introns with relatively small taxonomic distribution as well as deeply conserved introns. The deep conservation of many of the intron positions becomes apparent when focusing on those positions present in at least one of the human coronin genes (Fig. 1A). The six human coronin genes contain together 17 different intron positions, of which none is restricted to the genus *Homo*. Instead, six of these positions are shared with fungal genes, indicating their presence in at least the last common ancestor of the metazoa and fungi, whose divergence is estimated to 700 up to 1500 million years ago (Chernikova *et al.*, 2011; Peterson *et al.*, 2008). Five intron positions are even deeper conserved, because they are shared between Opisthokonta and other basal eukaryotic branches such as Alveolata, Stramenopiles and Amoebozoa (Fig. 1B). Plotting all intron gain and loss events onto the NCBI taxonomy tree reveals an intron-rich coronin gene in the last common ancestor of the eukaryotes and the opisthokonts (Fig. 1B). The many intron-loss events in, for example, the Choanoflagellida and Platyhelminthes branches are most likely due to the very low sequence and species sampling in these branches, but may also be due to the low number of introns present in some of the respective genomes in general. The actual CPU time used for this computation was 38′27″, which reduced to 0′34″ when using a precompiled list of lineages instead of extracting the lineages of the 312 species from the complete NCBI taxonomic tree.

4 Implementation

GenePainter is a standalone tool written in Ruby and requires Ruby v. 2 or higher. Python is required to convert the Newick tree format into svg format and for invoking GenePainter output in PyMOL (The PyMOL Molecular Graphics System, http://sourceforge.net/projects/pymol). GenePainter has been tested under Linux, MacOS and Windows.

Acknowledgements

We thank Prof. Christian Griesinger for his continuous generous support, Fabian Meyer for helpful discussions and James Dong for extensive testing and bug reporting.

Funding: This project has been funded by the Deutsche Forschungsgemeinschaft [DFG Grant KO 2251/13-1 to M.K.] and a synaptic systems fellowship [to S.M.].

Conflict of interest: none declared.

References

Chernikova,D. *et al*. (2011) A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct*, **6**, 26.

Du,H. *et al*. (2013) Genome-wide identification and evolutionary and expression analyses of MYB-related genes in land plants. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes*, **20**, 437–448.

Fawal,N. *et al*. (2012) GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families. *Bioinformatics*, **28**, 1398–1399.

Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.

Grewe,F. *et al*. (2013) Complete plastid genomes from Ophioglossum californicum, Psilotum nudum, and Equisetum hyemale reveal an ancestral land plant genome structure and resolve the position of Equisetales among monilophytes. *BMC Evol. Biol.*, **13**, 8.

Hammesfahr,B. *et al*. (2013) GenePainter: a fast tool for aligning gene structures of eukaryotic protein families, visualizing the alignments and mapping gene structures onto protein structures. *BMC Bioinformatics*, **14**, 77.

Koonin,E.V. *et al*. (2013) Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdiscip. Rev. RNA*, **4**, 93–105.

Lehmann,J. *et al*. (2013) Near intron pairs and the metazoan tree. *Mol. Phylogenet. Evol.*, **66**, 811–823.

Li,W. *et al*. (2009) Extensive, recent intron gains in Daphnia populations. *Science*, **326**, 1260–1262.

Lopez,P.J., Séraphin,B. (2000) YIDB: the yeast intron database. *Nucleic Acids Res.*, **28**, 85–86.

Odronitz,F. and Kollmar,M. (2008) Comparative genomic analysis of the arthropod muscle myosin heavy chain genes allows ancestral gene reconstruction and reveals a new type of 'partially' processed pseudogene. *BMC Mol. Biol.*, **9**, 21.

Pavesi,G. *et al*. (2008) Exalign: a new method for comparative analysis of exon–intron gene structures. *Nucleic Acids Res.*, **36**, e47–e47.

Peterson,K.J. *et al*. (2008) The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philos. Trans. R. Soc. B Biol. Sci.*, **363**, 1435–1443.

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

Yan,J. *et al*. (2014) Evolution, functional divergence and conserved exon-intron structure of bHLH/PAS gene family. *Mol. Genet. Genomics MGG*, **289**, 25–36.