Aequatus: An open-source homology browser

Anil S. Thanki[1,*], Sarah Ayling[1], Javier Herrero[2], Robert P. Davey[1,*]

1. Digital Biology, The Genome Analysis Centre, Norwich, NR4 7UH, UK
2. Research Department of Cancer Biology, UCL Cancer Institute, London WC1E 6DD, UK

*To whom correspondence should be addressed.

Abstract

Background: The phylogenetic information inferred from the study of homologous genes helps us to understand the evolution of gene families, the study of homology plays a vital role in finding ancestral gene duplication events as well as identifying regions those are under positive selection within species. Conservation of homologous loci results in syntenic blocks, and there are various tools available to visualise syntenic information between species. These tools provide an overview of syntenic regions as a whole, reaching down to the gene level, but none provide any information about structural changes within genes such as the conservation of ancestral exon boundaries amongst multiple genomes.

Findings: We present Aequatus, a standalone web-based tool that provides an in-depth view of gene structure across gene families, with various options to render and filter visualisations. It relies on pre-calculated alignment and gene feature information held in an Ensembl database, typically generated through, but not limited to, the Ensembl Compara workflow. We also offer Aequatus.js, a reusable JavaScript module that fulfils the visualisation aspects of Aequatus.

Availability: Aequatus is an open-source tool freely available to download under GPLv3 license at https://github.com/TGAC/Aequatus and a demo is available at http://aequatus.tgac.ac.uk

Contact: : Anil.Thanki@tgac.ac.uk and Robert.Davey@tgac.ac.uk

Introduction

Inferring the homology of genes across or within species is a commonplace technique to investigate synteny [1]. The inference process involves carrying out multiple sequence alignments comprising multiple steps and these can be computationally intensive even for small numbers of data points [2].

There are many methods available for findings of genome-wide orthology descriptions, for example MSOAR [3], OrthoMCL [4], HomoloGene [5], TreeFam [6], TreeBeST [7]. TreeBeST gives combined results based on species trees and dN/dS nucleotide and protein measures, unlike others which typically provide clustering without considering a given species tree topology. PhyOP [8] uses a tree-based method but it is useful only for closely related species. For these reasons, TreeBeST is used in the Ensembl Compara pipeline [9] a computational workflow developed by the Ensembl Compara team to infer familial relationships that includes clustering, multiple alignment, and tree generation. The Ensembl Compara schema is able to store comparative data such as gene families, syntenic regions, and protein families, and Ensembl Core database stores gene feature informations and other genomic annotations at the species level. The Ensembl project (release 84) at EMBL-EBI houses 87 species [10] on both production and early access websites, among them precomputed multiple alignments and gene family information for 70 vertebrate species.

There are many ways to represent and view comparative datasets, with the traditional method being phylogenetic trees, but also using tools such as Ensembl Browser [11], Genomicus [12], SyMap [13], and MizBee [14]. These tools are able to provide an overview of syntenic regions as a whole, with some reaching down to the gene order and orientation level. However, whilst retaining ancestral information, phylogenetic trees do not represent the underlying

information regarding structural changes within genes, such as the conservation of ancestral exon boundaries between multiple genomes or variants within genes that can be correlated to phenotypic changes. To begin to build these gene-level visualisations, basic genomic feature information is required. Therefore, we have developed Aequatus to bridge the gap between phylogenetic changes and gene feature information. Here we show that Aequatus allows identification of exon/intron boundary changes and mutations, which might be related to mis-annotations, pseudogenes [15] or polyploidisation in animal and plant genomes.

Aequatus

Aequatus is built using open-source technologies (Java on the server-side; HTML, JavaScript, jQuery and SVG on the client-side) to provide a fast and intuitive web-based browsing experience over complex data. It retrieves and process comparative genomics information directly from Ensembl Compara and the Ensembl Core database. Precalculated genomic alignments, in the form of CIGAR strings, are held in Ensembl Compara and Aequatus cross-references these sequences to Ensembl Core databases for each species to gather genomic feature information. Aequatus then processes both the comparative and feature data to provide a visual representation of the phylogenetic and structural relationships among the set of chosen species, using a shared colour scheme for coding regions to represent homology in internal gene structure alongside their corresponding gene trees. Aequatus is also able to indicate insertions and deletions in homologous genes with respect to shared ancestors.

As a part of the Aequatus project, we have developed Aequatus.js [16], an open-source JavaScript library. We have extracted the visualisation modules from the standalone Aequatus browser in order to make the tool more accessible and reusable. Aequatus.js comprises the latest web technologies such as SVG, jQuery, JavaScript and D3.js [17]. As a JavaScript library, it preserves the full interactive functionality of the full Aequatus browser tool, but it is able to be integrated with other third-party web applications.

Result

The main view of Aequatus is shown in figure 1, depicting homologous genes relating to a given selected reference gene. The selected gene acts as a 'master' (depicted as a black leaf node with a red label) and other genes are mapped to it based on alignment. Exons sharing the same colour represent those that are shared among genes, and black bars within exons represent insertions specific to a given gene compared with the 'master'. There is also a gene tree on the left indicating responsible events (e.g duplication, speciation, and gene split) for the proposed evolution of the gene family. You can also see the Aequatus control panel on the far left with various options for search, settings, filter, export and help. These options are explained below.

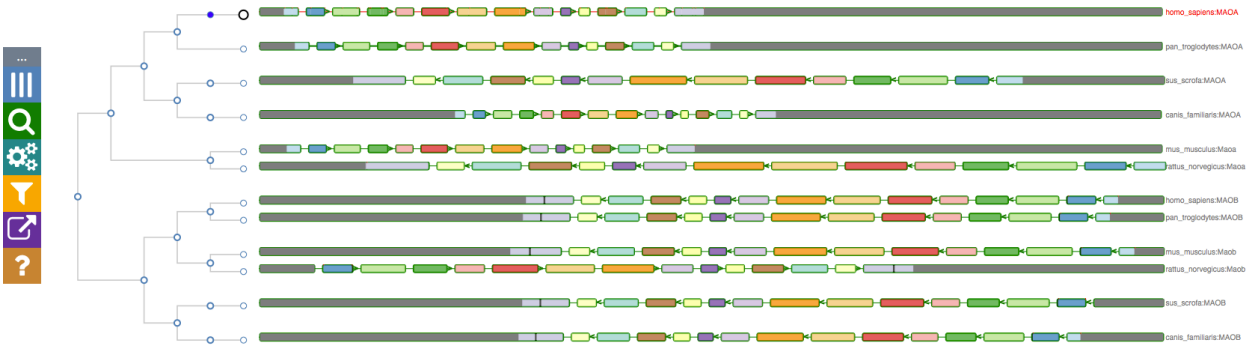

Figure 1 shows the homologous genes for monoamine oxidase (MAO) gene, with human gene as a master, alongside other homologous genes in the exon-focused view. Considering the gene tree on the left, it is clear that the MAO genes are separated into two clusters, one of which is MAO-A and the other is the MAO-B gene family. You can also see the Aequatus

Control Panel
Aequatus has an intuitive draggable control panel on left-hand side which contains options to modify gene views and labels, the search box, export options, and a help section. It also allows users to toggle showing genes from specified species in the main comparative view via a multi-select filter list of the currently available species.

Search
Search functionality in Aequatus is based on keyword lookups, in which the user can provide a given search term and a list is returned of all the gene members matching the query in gene name and stable_id (unique identifiers in the Ensembl project for each genomic annotation), and display_label (common name for a gene or protein), as well as matching any keyword in description.

Toggle Views
Aequatus provides two view types for homologous genes. The first is the default view in which all homologous genes are resized to the maximum available width in the web browser, showing introns and exons proportional to the real gene size. The second is the exon focused view (as seen in Figure 1) whereby all introns are set to a fixed width, as long introns can adversely affect the ability of a user to see detail in surrounding exons. This provides an easier browsing experience of the actual gene structure, especially when less screen real estate is available.

Export
Aequatus allows users to export genomic data for the homologous genes in various forms, such as simple lists of gene IDs, or sequence alignments in the form of protein sequences and CIGAR alignments alongside a list of corresponding gene/protein IDs or gene names. Users can also download gene trees in newick or JSON format.

Popup
Aequatus provides a contextual menu system via interactive popups which are displayed when a user clicks on a gene, supplying information such as the gene name and its position. The popup also has an option to prioritise the currently selected gene as the master in order to find insertions and deletions in homologous genes relative to that which is selected, and a link out to Ensembl for navigating to the genomic feature information held in the Ensembl Core service.

REST API
Aequatus provides a REST API to enable consistent access to genes of interest, making it easy to go back to the results of a previous search or to share information with collaborators via persistent URLs. The REST API is divided into three modes: query, search, and reference-based, where a user can share the link for a given homologous gene directly, the results of a search term, or a specific reference with a given species and chromosome, respectively.

Conclusions

The ultimate goal of Aequatus is to provide a unique and informative way to render and explore complex relationships between genes from various species at a level that has so far been unrealised. Whilst applicable to species with high-quality gold-standard reference genomes such as human or mouse, Aequatus has been designed with large fragmented genome references in mind, particularly hard-to-assemble polyploid plants. As Aequatus can visualise relationships using simple CIGAR strings, any tool that outputs this format can use Aequatus to view them. Finally, Aequatus.js represents an independent JavaScript version of the visualisation module, so that other web applications can take advantage of the comparative representation aspects without having to use the full Aequatus software.

Funding

This work was strategically funded by the BBSRC and through the EU TransPlant grant (BBS/E/T/000GP006).

Conflict of Interest: none declared.

References

1. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009;19:327–35.

2. Edgar RC, Batzoglou S. Multiple sequence alignment. Curr. Opin. Struct. Biol. 2006;16:368–73.

3. Fu Z, Chen X, Vacic V, Nan P, Zhong Y, Jiang T. MSOAR: a high-throughput ortholog assignment system based on genome rearrangement. J. Comput. Biol. 2007;14:1160–75.

4. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.

5. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2008;36:D13–21.

6. Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 2006;34:D572–80.

7. TreeSoft: TreeBeST [Internet]. [cited 2015 Dec 21]. Available from: http://treesoft.sourceforge.net/treebest.shtml

8. Goodstadt L, Ponting CP. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput. Biol. 2006;2:e133.

9. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, et al. Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res. 2003;31:38–42.

10. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.

11. Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz H-R, et al. The Ensembl Web site: mechanics of a genome browser. Genome Res. 2004;14:951–5.

12. Muffato M, Louis A, Poisnel CE, Crollius HR. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics. 2010;26:1119–21.

13. Soderlund C, Nelson W, Shoemaker A, Paterson A. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Res. 2006;16:1159–68.

14. Meyer M, Munzner T, Pfister H. MizBee: a multiscale synteny browser. IEEE Trans. Vis. Comput. Graph. 2009;15:897–904.

15. Vanin EF. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. 1985;19:253–72.

16. Thanki AS, Davey RP. TGAC/aequatus.js [Internet]. GitHub. [cited 2015 Dec 21]. Available from: https://github.com/TGAC/aequatus.js

17. Bostock M. D3.js - Data-Driven Documents [Internet]. [cited 2015 Dec 21]. Available from: http://d3js.org/