BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 8 2011, pages 1150–1151
doi:10.1093/bioinformatics/btr075

Genome analysis

Advance Access publication February 10, 2011

eGOB: eukaryotic Gene Order Browser

Marcela Dávila López and Tore Samuelsson*

Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, SE-405 30 Göteborg, Sweden

Associate Editor: John Quackenbush

ABSTRACT

Summary: A large number of genomes have been sequenced, allowing a range of comparative studies. Here, we present the eukaryotic Gene Order Browser with information on the order of protein and non-coding RNA (ncRNA) genes of 74 different eukaryotic species. The browser is able to display a gene of interest together with its genomic context in all species where that gene is present. Thereby, questions related to the evolution of gene organization and non-random gene order may be examined. The browser also provides access to data collected on pairs of adjacent genes that are evolutionarily conserved.

Availability: eGOB as well as underlying data are freely available at http://egob.biomedicine.gu.se

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

Contact: tore.samuelsson@medkem.gu.se

Received on October 4, 2010; revised on January 3, 2011; accepted on February 5, 2011

1 INTRODUCTION

The large number of eukaryotic genome sequences currently available has made extensive comparative analysis possible. For instance, studies of the conservation of gene order will help to understand genome evolution in general and what factors restrict the shuffling of genes. Databases and genome browsers have therefore emerged to compare genomes and gene order in different species. Some focus on specific organelles such as the OGRe (Jameson *et al*., 2003) and Plastid Gene Order Database (Kuhrihara and Kunisawa, 2004), while others cover a specific phylogenetic group like YGOB (Byrne and Wolfe, 2006) and Genomicus (Muffato *et al*., 2010). Typically gene comparisons between species are possible. We here present the eukaryotic Gene Order Browser (eGOB) which is useful for comparing and displaying genes with respect to their genomic environment. There are features that distinguishes eGOB from previously available 'gene order browsers'. First, non-coding RNA (ncRNA) genes are considered in addition to protein coding genes. Secondly, gene orthologs/homologs are identified using both OrthoMCL and Pfam. Thirdly, we consider a wide range of eukaryotic species representing all important eukaryotic phylogenetic groups; 19 metazoans, the choanoflagellate *Monosiga brevicollis*, 27 fungal species, 7 viridiplantae, 6 alveolata, 5 heterokonts, 2 amoebozoans, 4 euglenozoans and the three deep branching organisms *Giardia lamblia*, *Naegleria gruberi* and *Trichomonas vaginalis*.

*To whom correspondence should be addressed.

2 METHODS

eGOB stores information about the location of protein and ncRNA genes. The dataset corresponds to 1 122 102 protein genes and 395 149 ncRNA genes.

Protein ortholog and homolog relationships were identified with OrthoMCL (Chen *et al*., 2007) and Pfam classification (Finn *et al*., 2010), respectively (Davila Lopez *et al*., 2010). In addition, protein genes were functionally annotated with respect to gene ontology (GO) and a measure of functional similarity was calculated using the GS^2 method (Ruths *et al*., 2009).

As most genomes are missing adequate ncRNA annotation, we carried out such annotation for all of the organisms considered. First, all ncRNA genes from Rfam 9.1 (Gardner *et al*., 2009), including 975 ncRNA gene families, were used as queries in BLAST searches against genomic sequences. The resulting hits were scored with the INFERNAL 0.81 software (Eddy and Durbin, 1994) using the gathering cutoff as threshold (see also Supplementary Material). The locations of all predicted ncRNAs were finally recorded and added to the previously obtained information regarding the location of protein genes.

3 THE GRAPHICAL VIEW

The eGOB allows a user to display any eukaryotic gene and its environment in different species. The graphical view (Fig. 1) shows the reference gene or gene pair in the center with seven neighboring genes on both sides. Genes are represented by arrows which denote the relative direction of transcription. Thick and thin arrows denote protein and ncRNA genes, respectively. Each gene is color-coded according to its cluster ID and the clustering method used. Identically colored genes indicate an orthology relationship. It is possible to toggle between coloring schemes to emphasize either the clustering based on orthology (OrthoMCL) or the grouping based on domain architecture (Pfam). The user can navigate through the maps by scrolling to the left or to the right. The scrolling may be performed on each individual genome or on all of them simultaneously.

4 BIOLOGICAL APPLICATIONS

There are different problems that may be addressed by using eGOB. Two examples are shown here. One of them illustrates a situation where a user is able to monitor the evolution of a locus or chromosomal region. In the other example the evolution of pairs of adjacent genes is studied, with the aim of identifying genes that could be transcriptionally linked.

Example 1: visualizing the genomic context of the ParaHox cluster genes: the ParaHox genes (*Gsx, Pdx and Cdx*) code for homeodomain transcription factors that regulate the patterning of the anterior–posterior axis of animals (Gellon and McGinnis, 1998). These three genes are co-localized in the genomes of mammals,

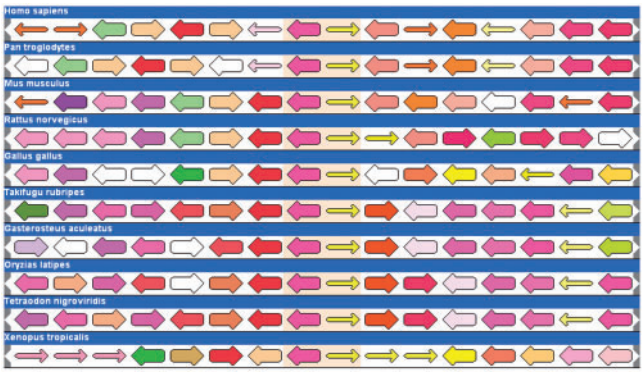

Fig. 1. Genomic context and evolutionary conservation of the *RNU12*/*POLDIP3* gene pair. Genes are represented by arrows, which denote the relative direction of transcription. Thick and thin arrows denote protein and ncRNA genes, respectively. Each gene is color-coded according to the OrthoMCL/Rfam cluster to which it belongs.

birds, frogs as well as Branchiostoma. It is believed that this arrangement was the ancestral gene organization. However, in the teleost fishes rearrangements gave rise to an organization where the three genes were separated (Mulley *et al*., 2006). Using eGOB, the ParaHox genes may be displayed in their genomic context. As described in detail in Supplementary Material, a gene may be identified on the basis of identifier or protein description information, or by a BLAST search. All orthologs may then be displayed in the gene order browser window. For instance, using the Swissprot identifier PDX1_HUMAN as a starting point, orthologs to the Pdx1 protein are found and gene order information may be displayed as shown in the Supplementary Figure S1. This figure shows that whereas the three ParaHox genes are tightly clustered in species such as human and mouse, they are separated in all the fishes available in eGOB, *Danio rerio*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Gasterosteus aculeatus* and *Oryzias latipes*. Also in *Ciona intestinalis*, the three ParaHox genes are separated (Ferrier and Holland, 2002). The *Gsx1* gene seems to be missing in chicken, but this may be a result of a gap in the current genome assembly (Prohaska and Stadler,2006).

Example 2: identifying genes that might be transcriptionally linked: pairs of genes that are divergently transcribed and that have a relatively short intergenic distance are expected to be related in terms of transcriptional control. Such pairs have been identified involving protein genes (Adachi and Lieber, 2002; Davila Lopez *et al*., 2010; Koyanagi *et al*., 2005; Piontkivska *et al*., 2009; Trinklein *et al*., 2004). Browsing through eGOB, we may also identify pairs of this nature that include a ncRNA. For instance, the human U12 small nuclear RNA gene (*RNU12*) is adjacent to the polymerase delta interacting protein 3 (*POLDIP3*) (Fig. 1). This gene pair is shown to be present in 10 other metazoans and these pairs are all associated with a short intergenic distance. This information indicates that the *RNU12* and *POLDIP3* genes are transcriptionally linked.

5 IMPLEMENTATION AND AVAILABILITY

MySQL (version 5.0.26) was used to store and manage the data in eGOB. Scripts for data querying and retrieving were written in PHP. The web interface was designed using HTML and JavaScript with all major browsers supported. All information on gene maps and gene pairs can be downloaded without any restrictions as tab separated value (TSV) files. Resulting data at the different steps in the query process can be exported as TSV files. Graphical representations of gene order maps may also be exported as HTML files as well as in an XML format.

ACKNOWLEDGEMENTS

We thank Moisés Salvador Meza Moreno for excellent technical support during the development of the web interface and valuable comments on the manuscript.

Funding: The Erik Philip-Sörensen Foundation.

Conflict of Interest: none declared.

REFERENCES

Adachi,N. and Lieber,M.R. (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell*, **109**, 807–809.

Byrne,K.P. and Wolfe,K.H. (2006) Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic Acids Res*., **34**(Database issue), D452–D455.

Chen,F. *et al*. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, **2**, e383.

Davila Lopez,M. *et al*. (2010) Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS ONE*, **5**, e10654.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.

Ferrier,D.E. and Holland,P.W. (2002) Ciona intestinalis ParaHox genes: evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. *Mol. Phylogenet. Evol*., **24**, 412–417.

Finn,R.D. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res*., **38**(Database issue), D211–D222.

Gardner,P.P. *et al*. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res*., **37**(Database issue), D136–D140.

Gellon,G. and McGinnis,W. (1998) Shaping animal body plans in development and evolution by modulation of Hox expression patterns. *Bioessays*, **20**, 116–125.

Jameson,D. *et al*. (2003) OGRe: a relational database for comparative analysis of mitochondrial genomes. *Nucleic Acids Res*., **31**, 202–206.

Koyanagi,K.O. *et al*. (2005) Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene*, **353**, 169–176.

Kuhrihara,K. and Kunisawa,T. (2004) A gene order database of plastid genomes. *Data Sci. J*., **3**, 60–79.

Muffato,M. *et al*. (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, **26**, 1119–1121.

Mulley,J.F. *et al*. (2006) Breakup of a homeobox cluster after genome duplication in teleosts. *Proc. Natl Acad. Sci. USA*, **103**, 10369–10372.

Piontkivska,H. *et al*. (2009) Cross-species mapping of bidirectional promoters enables prediction of unannotated 5′ UTRs and identification of species-specific transcripts. *BMC Genomics*, **10**, 189.

Prohaska,S.J. and Stadler,P.F. (2006) Evolution of the vertebrate ParaHox clusters. *J. Exp. Zool. B Mol. Dev. Evol.*, **306**, 481–487.

Ruths,T. *et al*. (2009) GS2: an efficiently computable measure of GO-based similarity of gene sets. *Bioinformatics*, **25**, 1178–1184.

Trinklein,N.D. *et al*. (2004) An abundance of bidirectional promoters in the human genome. *Genome Res*., **14**, 62–66.