eGOB: Eukaryotic Gene Order Browser

Marcela Dávila López¹, Tore Samuelsson₁, * ¹Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Box 440, SE-405 30 Göteborg, Sweden.

Associate Editor: Prof. 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 organisation 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. Contact: tore.samuelsson@medkem.gu.se

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 OGRA (Jameson et al. 2003) and Plastid Gene Order Database (Kurihara and Kunisawa 2004), while others cover a specific phylogenetic group like YGOB (Byrne and Wolfe 2006) and Genomicus (Muffato, Louis et al. 2010). Typically gene comparisons between species is 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 genes are considered in addition to protein coding genes. Secondly, gene orthologues/homologues 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.

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 non-coding RNA genes.

Protein orthologue and homologue relationships were identified with OrthoMCL (Chen, Mackey et al. 2007) and Pfam classification (Finn, Mistry et al. 2010), respectively (Davila Lopez, Martinez Guerra 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 method (Ruths, 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, Daub et al. 2009), including 975 non-coding RNA 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 cut off as threshold (see also Supplementary document). 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 (Figure 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, birds, frogs as well as Branchiostoma. It is believed that this arrangement was the ancestral gene organisation. However, in the teleost fishes rearrangements gave rise to an organisation where the three genes were separated (Mulley, Chiu et al. 2006). Using eGOB the ParaHox genes may be displayed in their genomic context. As described in detail in Supplemental document 1, a gene may be identified on the basis of identifier or protein description information, or by a BLAST search. All orthologues may then be displayed in the gene order browser window. For instance, using the Swissprot identifier PDX1_HUMAN as a starting point, orthologues to the Pdx1 protein are found and gene order information may be displayed as shown in the Supplementary figure 1. 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; Trinklein, Aldred et al. 2004; Koyanagi, Hagiwara et al. 2005; Piontkivska, Yang et al. 2009; Davila Lopez, Martinez Guerra et al. 2010). Browsing through eGOB, we may also identify pairs of this nature that include a non-coding RNA. 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: This work was supported by the Erik Philip-Sörensen Foundation

Conflict of Interest: None declared.

REFERENCES


