

# OMGProm: A Database of Orthologous Mammalian Gene Promoters

Saranyan K. Palaniswamy, Victor X. Jin, Hao Sun, Ramana V. Davuluri\*

Human Cancer Genetics Program, Comprehensive Cancer Center, Department of Molecular Virology, Immunology & Medical Genetics, The Ohio State University, 420 W.12<sup>th</sup> Avenue, TMRF 524, Columbus, OH 43210, USA.

## ABSTRACT

### Summary:

Sequence comparisons between human and rodents are increasingly being used for the identification of gene regulatory regions. The effectiveness of such an approach largely depends on the quality and availability of promoter sequences. We developed OMGProm by integrating three data sources: (1) experimentally supported full-length cDNA, promoter and first exon sequences (2) homology information from HomoloGene and (3) the human and mouse genomic sequences. The current version of OMGProm contains 8,550 promoter pairs of 6,373 orthologous human and mouse genes, where supporting experimental evidence for transcription start site annotation exists in at least one species.

### Availability:

OMGProm can be accessed from <http://bioinformatics.med.ohio-state.edu/OMGProm>.

\*Contact: [davuluri-1@medctr.osu.edu](mailto:davuluri-1@medctr.osu.edu).

### Supplementary information:

Additional information on methods and implementation is available at

<http://bioinformatics.med.ohio-state.edu/OMGProm/si.jsp>.

## INTRODUCTION

Comparisons between human and rodent DNA sequences are widely used for the discovery of genes (Ureta-Vidal *et al.*, 2003) and gene regulatory regions (Lenhard *et al.*, 2003; Chapman *et al.*, 2004). Although it is widely appreciated that the protein coding regions are significantly conserved between human and rodents, very little is known about the extent of conservation in gene regulatory regions and the capability of existing algorithms to predict regulatory elements. Even with the advent of whole-genome sequencing, there is a relative dearth of well-curated data sets that can be used to train and test these algorithms. Here, we present a database of orthologous mammalian gene promoters (OMGProm) as a platform for comparative genomics of transcriptional regulation, in order to facilitate the identification of gene regulatory elements, such as core promoters and transcription factor binding sites that are conserved in the upstream regions of orthologous genes (Blanchette and Tompa, 2003; Loots *et al.*, 2002).

Extensive molecular research in the field of transcription regulation has produced invaluable promoter sequence data that are being deposited into GenBank (Wheeler *et al.*, 2004). In parallel, recent advances in sequencing technologies (Imanishi *et al.*, 2004) have generated full-length cDNAs of mammalian genes. We systematically integrated the cDNA and genome sequence data and curated a set of 8,550 promoters of orthologous mammalian genes. We believe that this data repository will be a valuable control set for designing novel promoter prediction tools and for testing the sensitivity of existing

programs, such as FirstEF (Davuluri *et al.*, 2001). Our database serves to complement similar databases, such as DBTSS (Suzuki *et al.*, 2002) and PromoSer (Halees and Weng, 2004).

## DATA ACQUISITION

We developed a data-mining pipeline ([Web Figure 1](#)) that integrated data from different resources. We collected 4,802 full-length cDNAs from DBTSS (Suzuki *et al.*, 2002), 2,159 full-length 5'UTR sequences from the 5'UTR database (Davuluri *et al.*, 2000), and human and mouse promoters from the EPD database (Schmid *et al.*, 2004). We searched GenBank, using complex queries {e.g., “*homo sapiens*” [ORGN] AND (“5'UTR” [FKEY] OR *promoter* [FKEY] OR *exon* [FKEY] OR *mRNA* [FKEY] OR *prim\_transcript* [FKEY])}, for retrieving experimentally supported first exons, promoters and full-length 5'UTR sequences in humans. Then, we used a Perl script (*available upon request*) to parse these GenBank records into first exons, full-length mRNAs, full-length 5'UTRs and promoter sequences that are supported by experimental evidence. The Perl script scans each GenBank nucleotide record for *mRNA*, *exon*, *5'UTR*, *prim\_transcript*, *promoter* and *CDS* feature key annotations. If a feature was annotated as incomplete at the 5' end, e.g., “mRNA: <1..250”, “putative” or “evidence=not experimental”, the record was ignored. The script also ignored those records that had identical start coordinates for both the mRNA (or first exon) and CDS. We mapped all those first exons, full-length 5'UTR/mRNA and promoter sequences to the human (NCBI Build 34) and mouse (NCBI Build 32) genomes using BLAT (Kent and Brumbaugh, 2002). After eliminating the overlapping first exons that are mapped to the same genomic template, we had a set of 11,165 human and 7,297 mouse real first exons.

We used the [HomoloGene](#) database (Wheeler *et al.*, 2004) to obtain orthologous gene counterparts for all the genes in our database that have real first exons. We found 9,330 orthologous mouse promoters for 6,688 human genes, and 5,618 orthologous human promoters for 4,253 mouse genes. We then aligned the real first exon and promoter sequence of a gene in one species with the 5' upstream region of its orthologous counterpart as follows. We retrieved two sequences, Seq-1 (-5 kb to +2 kb relative to the TSS of the real first exon of human/mouse) and Seq-2 (-20 kb to +2 kb relative to the 5' end of the annotated first exon of the mouse/human orthologous counterpart). We aligned Seq-1 and Seq-2 using BLASTZ (Schwartz *et al.*, 2003). As Seq-1 has real first exon coordinates, this sequence is analogous to a template that is used to annotate the first exon of the orthologous gene in Seq-2. The degree of conservation of the core promoter (-35 bp to +35 bp) and donor site in the BLASTZ alignment determined the readjusted position of the TSS and donor site in Seq-2. The alignment was further refined by an additional alignment of Seq-1 and Seq-2 by ClustalW (Thompson *et al.*, 1994). We found 6,251 (67%) human promoters conserved in mouse and 4,288 (76%) mouse promoters conserved in human. In the combined set of 10,539 first exons, if any two first exons of a gene overlapped and the distance between transcription start sites was less than 50 bp, we eliminated the shorter first exon. This resulted in 8,550 promoter pairs, which include alternative promoters, for 6,373 unique gene pairs ([Web Table 1](#)).

## DATA PRESENTATION

The OMGProm database is presented via an interactive and easy-to-use web interface. The data are displayed with links to the raw data and related references. A user may query the database with a Unigene ID, Gene Symbol or GenBank accession number of interest. Each record displays information about a pair of promoters of orthologous genes. We provide the BLASTZ and ClustalW alignments of orthologous sequences, in which the first exons are displayed in different color. Additionally, CpG island and human-mouse sequence conservation tracks are presented as annotated in the UCSC genome database (Karolchik *et al.*, 2003). We implemented the database in MYSQL, and the web-interface and promoter visualization using GDVTK (Sun and Davuluri, 2004). For the convenience of our users, we have also run the promoter sequences through the RepeatMasker program (Smit, A.F.A. & Green, P; <http://repeatmasker.org>), and these masked sequences are also available for downloading.

We will periodically update OMGProm by including additional organisms, such as rat and chimpanzee, and by continuously updating the data utilizing our data-mining pipeline. The results of different phylogenetic footprinting and sequence alignment programs will be included in the future updates of OMGProm.

#### ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their constructive suggestions, Twyla Pohar for editing the manuscript and Sang-Gook Han for assistance with web implementation.

#### REFERENCES

- Blanchette, M., Tompa, M. (2003) FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840-3842.
- Chapman, M.A., Donaldson, I.J., Gilbert, J., Grafham, D., Rogers, J., Green, A.R. and Gottgens, B. (2004) Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. *Genome Res.*, **14**, 313-318.
- Davuluri, R.V., Suzuki, Y., Sugano, S., Zhang, M.Q. (2000) CART classification of human 5' UTR sequences. *Genome Res.*, **10**, 1807-1816.
- Davuluri R.V., Grosse, I., Zhang, M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat Genet.*, **29**, 412-417.
- Halees, A.S., Weng, Z. (2004) PromoSer: improvements to the algorithm, visualization and accessibility. *Nucleic Acids Res.*, **32**, W191-194.
- Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PloS Biology.*, **2**, 3-20.
- Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51-54.
- Kent, W.J. and Brumbaugh, H. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.*, **12**, 656-664.

- Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., Wasserman, W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J Biol.*, **2**, 13.
- Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., Rubin, E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832-839.
- Schmid C.D., Praz, V., Delorenzi, M., Perier, R., Bucher, P. (2004) The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res.*, **32**, D82-85.
- Schwartz, S., Kent, W.J., Smith, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-Mouse Alignments with BLASTZ *Genome Res.*, **13**, 103-7.
- Sun, H. and Davuluri, R.V. (2004) Java-based application framework for visualization of gene regulatory region annotations. *Bioinformatics*, **20**, 727-734.
- Suzuki, Y., Yamashita, R., Nakai, N. and Sugano, S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328-331.
- Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-80.
- Ureta-Vidal, A., Ettwiller, L., Birney, E. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet.*, **4**, 251-262.
- Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Suzek, T.O., Tatusova, T.A., Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35-40.