The Human - Mouse Promoter Machine at IFOM: a tool for retrieval of orthologous promoter sequences from genome sequence data

A.Guffanti1, L.Lassandro1, G.Finocchiaro1,2 & H.Muller1,2

1: IFOM – FIRC Institute of Molecular Oncology. Via Adamello, 16 – 20139 Milano, Italy
2: IEO – European Institute of Oncology – Via Ripamonti, 435 – 20141 Milano, Italy

Gene expression in eukaryotes is a highly coordinated process involving regulation at many different levels. The regulation of transcription initiation is an important, and often rate-limiting, step in this process. Although several types of cis-acting DNA sequence elements contribute to this regulation, the simplest element to locate may be promoters, as they are located just upstream of transcription start sites. Until recently, most functional studies of promoters were conducted on a gene-by-gene basis, but there also have been recent attempts to identify promoters on a large-scale with strictly computational methods (Davuluri et al. Computational identification of promoters and first exons in the human genome. Nat.Genet. 29:412-417).

The DNA sequences of entire genomes are being determined at a rapid rate. The extensively annotated human and mouse genome assemblies are available from the joint Sanger Institute – EBI project "Ensembl" (http://www.ensembl.org) at different levels of access, including a dedicated API for direct access of the remote relational database layer.

Starting from these considerations, we have established a project for automated retrieval of human and mouse orthologous genomic regions upstream of the first exon of annotated genes, starting from generic gene identifiers.
In order to compile the list of human and mouse genes that are linked by a relation of homology and possible orthology, at least at a sequence similarity level, we compiled a list from two NCBI resources: HomoloGene (http://www.ncbi.nlm.nih.gov/HomoloGene/) and LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/). We have written a set of scripts, using the BioPerl EnsEMBL API, in order to convert these genes to EnsEMBL identifiers and retrieve from the human and mouse assemblies the DNA sequence corresponding to the first exon and 1000 bp upstream of the first exon start site.
After assigning a unique identifier to these sequences, we have built a query interface that performs an intermediate conversion through the LocusLink database, mirrored and indexed under the SRS system at IFOM. It is possible to query this database using a range of generic gene identifiers (LocusLink identifiers, Gene Names or Gene Aliases, Accession Numbers, Pfam domains, Gene Ontology terms, RefSeq Accession Numbers, EnsEMBL identifiers).
It is possible to retrieve human genomic sequences, mouse genomic sequences, mouse orthologous sequences starting from human identifiers and viceversa. The output consists of DNA sequences in FastA format, where the region corresponding to the first exon is uppercase and the region corresponding to 1000 bp of genomic DNA is lowercase.

We aimed to provide an updated data set of putative promoter regions from the human and mouse assemblies that can be easily queried with generic gene identifiers. This dataset should be useful for researchers interested in in silico promoter work. Future development will include the addition of a set of annotations such as known regulatory elements and cross-species conserved regions; the use of orthology information directly from EnsEMBL; the addition of a graphical output for the annotated regions; the addition of other organisms to the computational pipeline.

The Human<>Mouse promoter machine at IFOM is freely available at web address
http://bio.ifom-firc.it/PROM_MACHINE/index.html