Bovine promoter annotation platform for the identification of transcription factor binding sites in genes involved in early pregnancy

Magalie Leveugle, Valentin Loux, Jean-François Gibrat

Bovine promoter annotation platform for the identification of transcription factor binding sites in genes involved in early pregnancy

Magalie Leveugle, Valentin Loux, Jean-François Gibrat
{magalie.leveugle, valentin.loux, jean-francois.gibrat} @jouy.inra.fr

Introduction

Deciphering mechanisms of early implantation of the embryo in cattle is of economical and fundamental interest, as these findings could help to reduce the implantation failure and also give clues regarding human sterility issues.
One possible approach is to identify precisely the transcription factors (TF) responsible for gene expression in the uterus-conceptus dialogue. TF are proteins that interact directly with the DNA molecule by recognizing small and slightly degenerate nucleotide motifs (TFBS) in the promoter region of the genes

In the present work we propose the combination of several bioinformatic approaches to identify regulatory motifs in cattle promoters for coregulated genes. We integrate results from phylogenetic footprinting (Footprinter) with orthologs from human, mouse and rat genomes and genetic algorithm (GALF-G) and statistical-based motif search (Weeder).

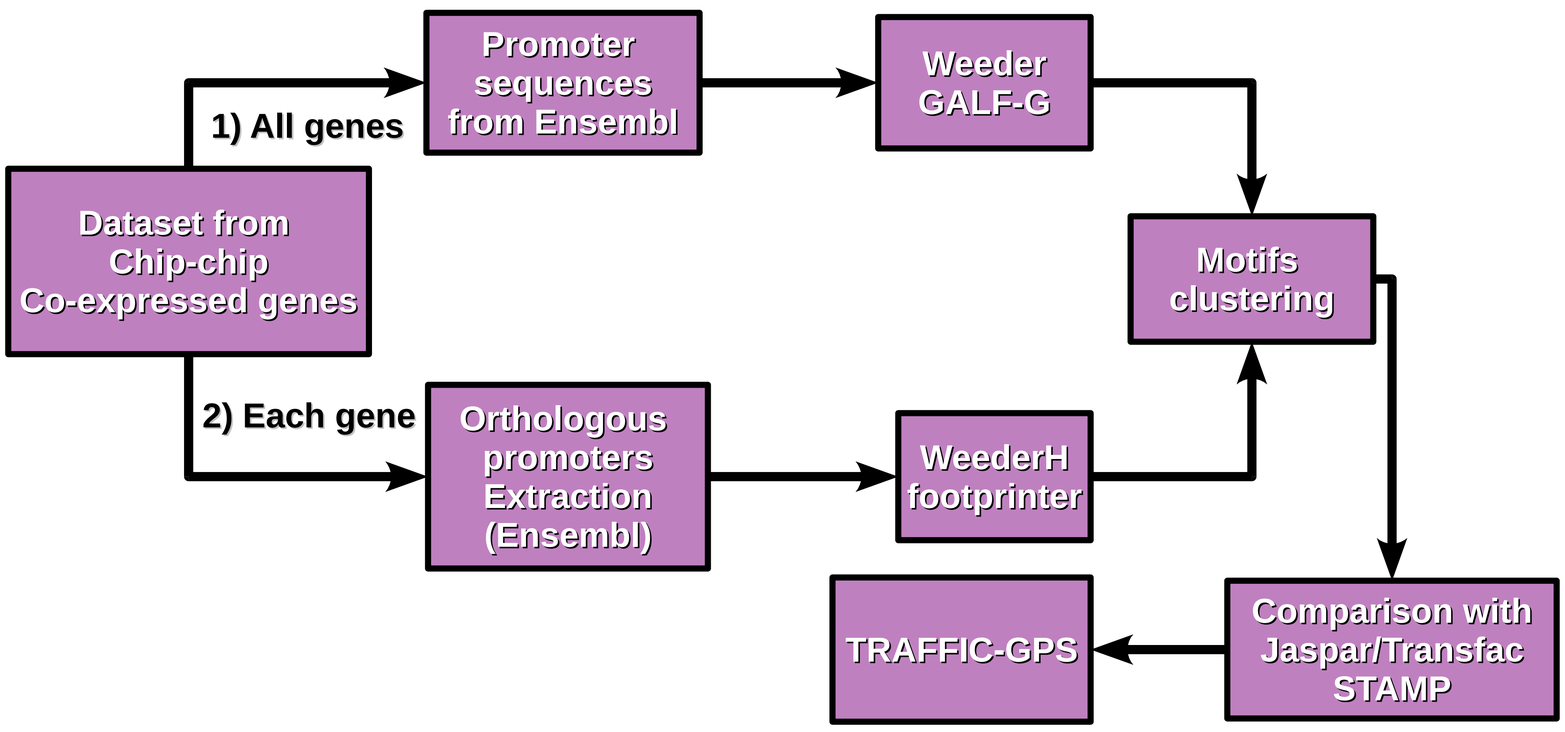
Methods

- **Program selection**

We chose programs which are available as command-line version on Unix platform and with good performance in TFBS detection according to bibliographic review and local tests.

- **Dataset : Bovine EST from ChIP experiments.**

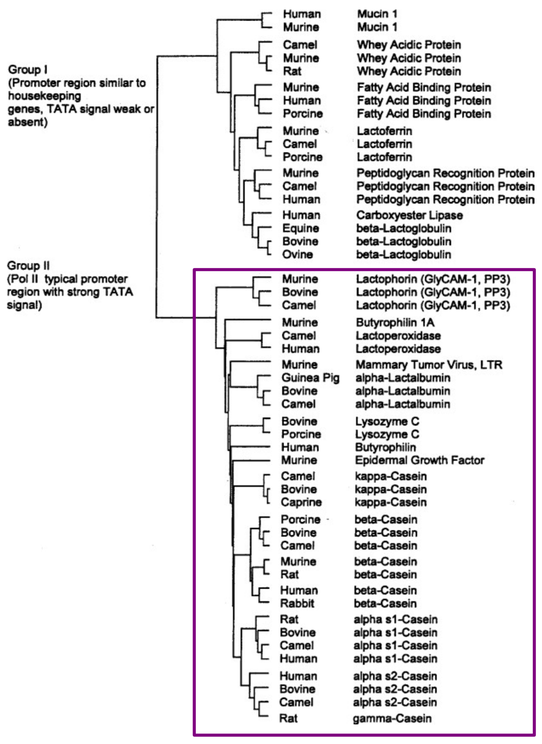
Each EST is mapped onto the corresponding gene in Ensembl.
Promoter sequences and their orthologs in mouse, human and rat are extracted from the Ensembl database through the Compara and Core API.
The promoter sets are analyzed with Weeder and GALF-G for co-expressed genes (1) and with WeederH and Footprinter for each group of orthologs and co-orthologs (2).
Resulting motifs are clustered and compared to motifs databases with Stamp.



Test Dataset

Camel, cow and human milk proteins.

As described in Kappeler et al. we used the promoters from genes expressed in milk. These genes possess strong TATA signal and typical PolII promoters (Group II in the tree). We extracted the 2000 bp sequences before the annotated TSS for human and cow in Ensembl, and the 5' sequences available in GenBank for the Camel.

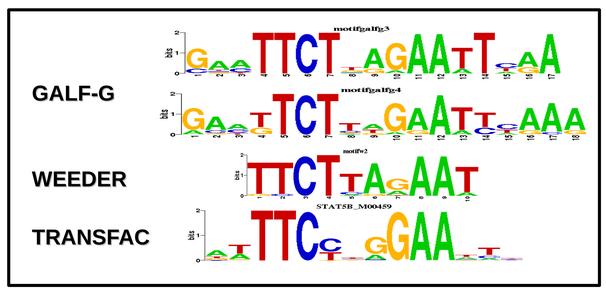


TFBS were searched with our pipeline and the best results for each method were clustered and compared with the TRANSFAC database.

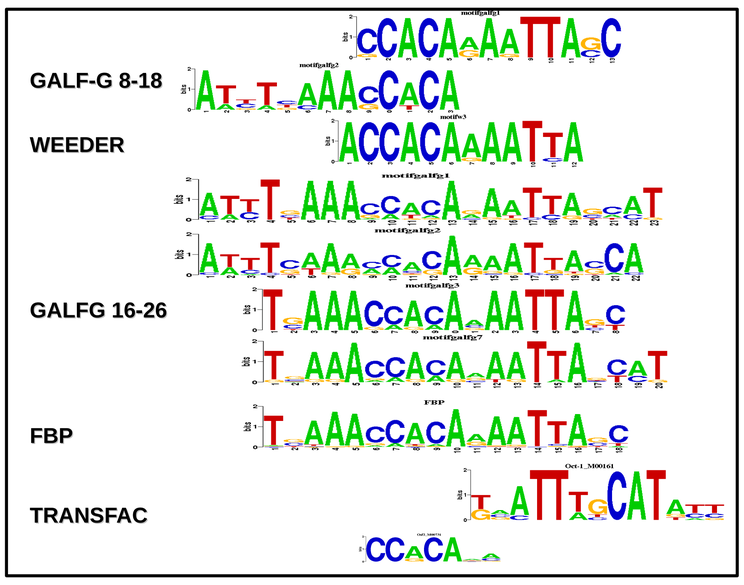
Only the motifs found by at least two methods were kept.

STAT and Oct1 motifs are the most represented in the results and are found by all the programs except GALF-G with 8-18 bp motif width parameter. This results confirmed the usefulness of several methods in combination with different ranges of parameter to reduce the false positive results and find all the relevant motifs.

STAT Motif



Oct1 Motif



Future Developments

- **Improvement of the motif selection process, since this version of the pipeline still requires a step of human curation.**
- **Integration of the pipeline in an interface dedicated to the analysis and prediction of TFBS for the cattle biologists.**

References

- **Blanchette M, and Tompa M, *Nucl. Acids Res. 31: 3840-3842.***
- **Chan TK, Li G, Leung KS,Lee KH, *BMC Bioinformatics, 2009, 10:321***
- **Kappeler SR, Farah Z, Puhan Z. *J Dairy Sci. 2003 Feb;86(2):498-508.***
- **Mahony S, Auron PE, Benos PV, *PLoS Computational Biology (2007) 3(3):e61***
- **Matys V, Kel-Margoulis OV, Fricke E, et al. *Nucl. Acids Res. 2006 Jan 1;34:D108-10.***
- **Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, and Pesole G. *Nucl. Acids Res. 34: W566-W570.***

MIG – Mathématique, Informatique et Génome
Domaine de Vilvert,
78352 Jouy-en-Josas
This work is funded as part of the ANR-08-GENM-037 project MEETAC