# Gene structure prediction and validation based on homologous genomic sequences

T. Wiehe, M. Burset[+], J. Abril[+], S. Gebauer-Jung and R. Guigò[+]

Max Planck Institute for Chemical Ecology, Department of Molecular Genetics and Evolution,
Tatzendpromenade 1a, 07745 Jena, Germany,
Phone: +49-3641-64 3631, Fax: +49-3641-64 3668, e-mail: twiehe@ice.mpg.de

[+] Institut Municipal d' Investigacio Medica, Medical Bioinformatics,
C/ Dr. Aiguader, 80, 08003 Barcelona, Spain.

Pairwise gapped alignments of homologous genomic sequences of two species usually show a patchwork pattern of conserved and less conserved fragments, the so called phylogenetic footprint. Furthermore, if species are evolutionarily closely related, such as mouse and man, we have observed that homologous splice site signals are generally highly conserved. We have developed a method that uses information on sequence and splice site conservation to predict simultaneously the gene structure in syntenic sequences from two different species. To test this algorithm we have generated from EMBL (release 56) a test set of human/rodent homologous genomic sequences for which both CDS and mRNA are annotated. We found that this algorithm is highly specific in its predictions of coding sequences and exons: we evaluated the performance in terms of standard measures for the gene prediction quality and found a sensitivity of 0.93 (0.87) and a specificity of 1.00 (0.93) on nucleotide and exon levels, respectively. Syntenic gene prediction may be particularly useful to identify lowly expressed transcripts or transcripts of uncommon structure, where other gene prediction tools are prone to fail. Another use of the above algorithm is the validation of gene annotations: in our test set we found several instances where the annotated gene may contain errors such as wrongly annotated splice sites, start or stop codons and sequencing errors. With the help of our algorithm some of these potential errors can be readily corrected or a putative structure can be validated *in silico* if a homologous reference sequence, which may itself even be unannotated, is available.

Our algorithm will be freely distributed for academical use from a publicly accessible ftp-server (apolo.imim.es).