Data Structures for Genome Annotation, Alternative Splicing, and Validation

Sven Mielordt[1,*], Ivo Grosse[1], and Jürgen Kleffe[2]

[1] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK),
06466 Gatersleben, Germany
[2] Charité-Universitätsmedizin Berlin - Campus Benjamin Franklin
Institut für Molekularbiologie UND Bioinformatik,
Arnimallee 22, 14195 Berlin, Germany
mielordt@ipk-gatersleben.de, grosse@ipk-gatersleben.de,
juergen.kleffe@charite.de

Abstract. To establish a clean basis for studying alternative splicing and gene regulation in life science projects, a powerful data modeling and also a strict validation procedure for assigning levels of reliability to given gene models is essential. One common problem of public genome databases are insufficiently organized and linked description data, which make it difficult to study relations of the alternative isoforms of a gene that are relevant for medicine and plant genome research. This is a severe obstacle for the integration of biological data and motivated us to establish a new modeling instance and that we call splice template or sTMP. Every sTMP has a unique splicing pattern, but the length of the first and the last exon remains undefined. This allows to model different gene isoforms with the same splicing pattern. By utilizing this more fine-grained data structure, many cases of plurivalent mRNA-CDS relations are uncovered. There are more than 3,000 extra CDSs in the human genome compatible with the categories sTMP, mRNA and CDS, which exceed the classical one-to-one relations of mRNAs and CDSs. In one case, 11 extra CDSs are compatible with one mRNA. Crosslinks between mRNAs derived from different sTMPs leading to the same CDS are now accessible as well as disease-related ruptures in UTR regions. This allows discovering and validating disease and tissue specific differences in alternative splicing, gene expression and regulation. Another problem in public databases is a too much relaxed standard for labeling genes "confirmed by ESTs and full-length-cDNAs." We provide a pipeline that handles gene annotations from different sources, integrates them into complex gene models and assigns strict validation tags, constrained by a local low-error model for the alignments of genome annotation and transcripts. The data structures are being implemented and made publicly available at the Plant Data Warehouse of the Bioinformatics Center Gatersleben-Halle (http://portal.bic-gh.de/sTMP).

Keywords: Gene and genome annotation, alternative splicing, data integration, splice template, validation and confirmation, quality control, Fasta-XML format.

* Corresponding author.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 114 – 123, 2006.

1 Introduction

Gene annotation and prediction is still a challenging task. On average less than 40% of the *ab-initio* gene predictions for Arabidopsis are error-free and hence genes with complete full-length cDNA support are important for testing and training gene prediction software [1]. Studying alternative splicing and gene regulation in animal and plant genome projects not only demands a strict validation procedure, but also a powerful meta-data modeling. Insufficiently organized and linked description data makes it difficult to express and uncover relations between the alternative isoforms of a gene. This is a severe obstacle for the integration of genome data.

The EnsEMBL and TIGR-XML Annotation. EnsEMBL [2] and TIGR-XML [3] model protein-coding genes on the genomic DNA as fixed hierarchical tree structures as shown in Fig. 1. A gene locus may have one or more splice isoforms (mRNAs). Each mRNA splits into the protein coding CDS region and the two untranslated regions 5'UTR (upstream) and 3'UTR (downstream). This topology cannot deal with alternative start codons for the same mRNA or other alternative mRNAs with the same splicing pattern, such as mRNAs with alternative transcription start sites or alternative polyadenylation sites. Moreover, there are no crosslinks given between the different CDSs of a gene, although these would be instructive, since often alternatively spliced mRNAs differ only in the UTR regions but lead to the same CDS and therefore code for the same protein.

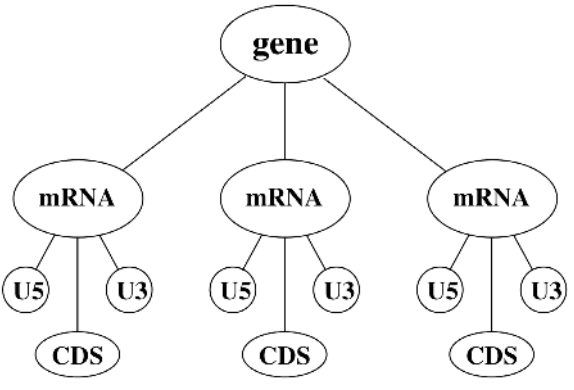

Fig. 1. The EnsEMBL and TIGR-XML entities model protein-coding genes on the genomic DNA as fixed hierarchical tree structures

The GenBank (RefSeq) Annotation. In contrast, NCBI [4] GenBank (RefSeq) annotations do not care for the relationship between mRNAs and CDSs and are therefore more general. However, they provide only unrelated lists of mRNAs and CDSs for each gene, as can be seen in **Fig. 2**. The users are left alone to build and run their own programs for matching mRNAs and CDSs in order to find the relations between mRNAs and CDSs.