7 Summary

Alternative splicing generates multiple products from a single gene and partly explains the diversity in eukaryotic transcriptomes. Expressed Sequence Tag (EST) data is a major resource that enables identification of exon-intron structure of transcripts, including the alternative ones. However, technical artifacts like contamination of the EST data with unspliced mRNA, gaps in the alignment, etc. lead to incorrect predictions of exon-intron boundaries. Therefore, this thesis aims to separate potentially reliable splice sites from the frequent data-/method-related artifacts. Once the gene structure has been efficiently delineated, the more complicated derivation of tissue-/tumor-specific transcripts could be addressed with increased reliability.

As a first step however, the predicted splice sites need to be separated from the multitude of inter-mingled data artifacts. In the absence of a distinct set of rules that govern the splicing process, definition of a mathematical score function is a difficult proposition. As an alternative approach, I have used fuzzy logic that offers a robust approximation for the combination of various EST-based parameters into a single score. Using fuzzy logic modeling, a *quality* value for all predicted splice sites is computed. This splice-site quality is subsequently used to estimate *confidence* in putative alternative splice events. Applying the method on a set of known alternative exons (AEDB database: Stamm et al. (2000)), almost all listed exons were assigned high quality values (>70%). Additional validation was provided by performing RT-PCR experiments performed for a limited number of alternative splice events. In most cases (17 out of 19), the method correctly classified the true positives and true negatives.

Another feature of the Expressed Sequence Tag data is the annotation of tissue and/or tumor source of the cDNA libraries that were used to generate the EST sequences. This annotation was exploited for assigning tissue-/tumor-specificity to the predicted (alternative) splice events. Upon validation of the expression pattern of these transcripts using RT-PCR experiments over a large set of tissue types, the expression of transcripts in the respective tissues was always confirmed. However, the experiments often revealed expression in additional tissues that are not represented in the EST data. This could partly be explained by the variation in protocols of EST generation as well as lack of ESTs for some tissues. In principle, such irregularities could be compensated by a (statistically) significant number of ESTs confirming tissue-specific events. Since that is usually not the case for individual splice events, there is a need for large scale validation experiments and/or comparison to independent datasets.

In order to facilitate integration of our EST based tools with external datasets, our

predictions of (alternative) transcripts and their expression patterns have been implemented into a relational database schema (T-STAG: Tissue-Specific Transcripts And Genes). In addition, this database includes estimates of tissue-/tumor-related gene expression levels as well as man-mouse orthology relationships. Therefore, apart from being a portal for these individual datasets, the T-STAG web-interface is designed to integrate underlying resources, thereby enabling applications like the detection of differentially expressed genes in tumors, the retrieval of orthologs with significant expression in the same tissue and genes specific to groups of tissues. Furthermore, the refined categorization of ESTs according to the normalization of cDNA libraries allows to search for putative low abundant transcripts.
To conclude, in this thesis the EST data has been employed to reliably identify (alternative) transcripts. Subsequent evaluation of the expression patterns using RT-PCR experiments confirmed the tissues in which the transcripts were predicted to be expressed, but often revealed expression in additional tissues. Therefore, such predictions of expression patterns of transcripts need to be supported by large scale validation experiments, for example using cutting edge microarray technology (Johnson et al. (2003)). Such an integration is facilitated by our comprehensive database, T-STAG. Advanced features of the T-STAG web-interface enable several biological applications like the detection of tumor markers and the evolution of tissue-specific expression of transcripts.