Britta Mersch
Dr. Sc. hum.

**Predictive statistical methods for genomic sequence elements affecting splicing and translation**

Geboren am 08.01.1981 in Münster
Diplom der Fachrichtung Biomathematik am 17.06.2005 an der Universität Greifswald

Promotionsfach: DKFZ (Deutsches Krebsforschungszentrum)
Doktorvater: Prof. Dr. rer. nat. S. Suhai

At present, it seems that even the detailed knowledge about the complete human genome sequence has not been sufficient to unravel mechanisms underlying the diversity of gene products and their regulation. Especially in eukaryotes, alternative splicing is responsible for the production of a large number of proteins in spite of a relatively small number of genes. It is intuitive that this complex mechanism requires efficient regulation strategies in order to guarantee that the correct proteins are produced. This fact was the motivation of this thesis in which bioinformatics methods were used to shed light on different aspects of the regulation of the splicing mechanism.

A recently proposed kernel – the oligo kernel – for support vector machines (SVMs) was refined to recognize small regulatory sequence elements. Its suitability was demonstrated by using it for the detection of translation initiation sites (TISs) from the *E. coli* genome. This particular organism was chosen because a corresponding reliable set of biochemically verified sequences is available. In addition to the good performance of the SVM using this kernel, biologically relevant information about the nature of TISs could be extracted. For eukaryotic genes, the prediction of the correct TIS is not sufficient for gene finding as the genes are separated into exons and introns. Therefore, the correct detection of splice sites is necessary. In the described work, SVMs using refined oligo kernels were shown to produce very satisfying classification rates. In addition, they provided insight into features that are relevant for classification of splice sites from the eukaryotic *C. elegans* genome.

Especially for non-consensus splice sites, there are other sequence elements which help to define a splice site, such as, exonic or intronic splicing enhancers (ESEs, ISEs). Point mutations can lead to the inactivation of ESEs, which promote the inclusion of exons in which they reside. This can change the splicing patterns and result in erroneous proteins which might be involved in disease development. In order to analyze the impact of ESEs an accurate detection method is required. In this thesis, an SVM using a refined oligo kernel was employed for this purpose. Since supervised machine learning methods such as SVMs require appropriate training samples, a new method for obtaining these samples from genomic sequences was developed. The resulting trained SVM classifier showed excellent performance and can now be used to verify potential ESE occurrences in order to reduce incorrect predictions.

As a special case of alternative splicing, transposed elements (TEs) can be exonized into the mRNA since TEs provide alternative splice sites and potentially ESEs. This event was observed to be more frequent in the human genome than in the mouse genome, especially due to the primate-specific Alu elements. Exonizations of TEs can cause genetic disorders as well as tissue- or tumor-specific splice forms. In this thesis, several potential tissue- or tumor-specific isoforms were identified using a newly developed analysis pipeline. This pipeline relies on expressed sequence tags (ESTs), but it is not possible to infer the level of expression in a certain tissue directly from the ESTs due to inherent biases in the available annotated databases. Methods of Bayesian statistics were used to cope with

this drawback and filter for statistically significant signals. The previously unknown or unstudied transcript variants for genes that were finally obtained could potentially provide new knowledge about gene function or provide new targets for drug development. Alternative splicing is often ignored in the pharmaceutical industry, but could be highly significant for drug discovery programs.

Exonization of TEs is possible because of sequence motifs inside of the TEs which differ from regular splice sites in one nucleotide only. An active splice site can be induced by single nucleotide polymorphisms (SNPs) providing exactly the mutation that induces an active splice site. In human TEs, a higher SNP density was observed than in mouse TEs which is consistent with the fact that a larger amount of exonizations was found in the human genome. Additionally, a higher density of SNPs in the primate-specific Alu element was obtained, supporting the high exonization capability of Alu elements. As a SNP always has two alleles from which only one induces the splice site, there may be individuals in the population which do not possess a certain exonization. 14 SNPs in the human genome and 3 SNPs in the mouse genome were identified which occurred in splice sites of exonized TEs. An inquiry of population frequency data for these SNPs suggested that such exonizations indeed create divergence in the proteome within the individuals of a species. These individuals may then differ in their regulation of pathways, their response to certain drugs or in the course of certain diseases.

In later biomedical research, the obtained results provide the opportunity to determine the effect of regulatory elements related to alternative splicing on diseases, especially cancer since it is known that cancer can arise as a genetically induced disorder.

In future work it would be interesting to investigate the interaction of exonic splicing enhancers and transposed elements. Using the methods and tools developed in this thesis, it could be possible to shed light on the role of ESEs during exonization.