

# 1 Motivation

In the post-genomic era, deciphering the relationship between genomic sequence and complex proteomic world has become a major challenge. The intermediate link between the genome and the proteome is the transcriptome. An enormous amount of diversity is introduced during the process of transcription with a single gene giving rise to multiple transcripts which eventually form different proteins. The fact that this process is carefully manipulated by nature to produce completely different tissue types is even more intriguing. During the last decade, several approaches have succeeded to produce evidences of transcript expression in different tissues/ conditions. The transcript-related data is mainly derived from one of the three resources, viz. microarrays, expressed sequence tags (ESTs) and serial analysis of gene expression (SAGE). The amount of data produced by these approaches is enormous. Bioinformatics supplies a means to analyze this increasing amount of data. This thesis focuses on the bioinformatics based analysis of EST data. Due to low sequence quality of ESTs as well as frequent genomic contamination, it is imminent that any analysis of EST data should be preceded by a cleaning step. Such a step would improve reliability of the predicted transcripts. Subsequently, owing to the tissue-related annotations present in the EST data, the putatively reliable transcripts can then be used to derive (specific) expression patterns. An understanding of the possible limitations of the data can be achieved via independent validations like comparison of the initial computational analysis with laboratory experiments.

It is the scope of this thesis to develop EST-based tools for transcriptome analysis. This includes studying the variation in expression patterns of transcripts across different tissues. By the comparison of computational predictions with laboratory experiments, the strengths as well as limitations of the EST data are identified. The thesis concludes with the description of a comprehensive world-wide-web (www) portal to enhance the applicability of the EST-based predictions. The benefits of having such a resource are pointed out using biological applications.

