Chapter 1

This chapter gives a brief description about rationale of the work, aims and objectives and hypothesis drawn from conclusions. Finally it sheds light on structure of the thesis.

*It is one of the more striking generalizations of biochemistry - which surprisingly is hardly ever mentioned in the biochemical textbooks - that the twenty amino acids and the four bases, are, with minor reservations, the same throughout Nature.*

Francis Crick
1.1. Background

The flow of genetic information from DNA to protein via replication, transcription and translation was referred as “Central Dogma of Molecular Biology” by Francis Crick in 1956. The information for the synthesis of proteins is encoded by RNA. A newly synthesized RNA molecule is known as primary transcript or pre-mRNA. This transcript governs the type of proteins to be produced by a cell under a given environmental condition that essentially involves the mechanism of post-transcriptional processing of mRNA known as splicing. The most extensive processing of primary transcripts occurs in eukaryotic mRNAs.

In eukaryotes, single pre-mRNA transcribed from single gene can give rise to multiple mRNA transcripts which code for multiple protein isoforms by a complex splicing process called alternative splicing. Transcripts from approximately 95% of multi-exon human genes are spliced in more than one way, and in most cases the resulting transcripts are variably expressed in different cell and tissue types (Wang et al., 2008). Alternative pre-mRNA splicing is emerging as one of the most important mechanisms to control eukaryotic gene expression. It regulates numerous aspects of protein functions, such as binding properties, intracellular localization, enzymatic activity, stability and post-translational modifications. The sequences of metazoan genomes with their relatively low gene numbers have highlighted the question of how protein number can be expanded beyond the gene number for a complex organism. This discrepancy in the gene number and the gene products can be partly answered through alternative splicing as it produces multiple proteins from a single gene generating biological complexity.

Misregulation of alternative splicing is often involved in various diseases. Alternative RNA splicing is reported in many diseases including neurological and muscle diseases (Yap and Makeyev, 2013). In cancer-associated genes, alternative splicing has important roles in oncogenesis, tumor suppression, and metastasis (Hagen and Ladomery, 2012). Alterations in alternative splicing are commonly reported in various types of cancers (Miura et al., 2011). Spliced variants are also potential biomarkers (Yi and Tang, 2011) for cancer diagnosis/prognosis and may be the targets for cancer therapy based on specific splicing correction treatments. There are several methods of therapeutic intervention in cancer that either alter or exploit the alternative splicing event itself. Also, several direct antisense-based methods are being developed to
Chapter 1: Introduction

remedy aberrations of splicing in genetic disease including RNA interference, stable antisense, and hybrid protein nucleic acids (Garcia-Blanco et al., 2004). Many examples of alternative splicing changes are involved in apoptosis (Wu et al., 2003), so altering AS can be employed to selectively kill cancer cells. Numerous genes associated to cell death pathway are highly dependent on alternative splicing to produce protein isoforms of a particular apoptotic gene having either pro- or anti-apoptotic function.

The rapid expansion of nucleotide sequence data available in public databases is revolutionizing biomedical research. These databases have a variety of uses, including the discovery of novel genes, identification of homologous genes, analysis of alternative splicing, chromosomal localization of genes and detection of polymorphism. In the post-genomic era, deciphering the relationship between genomic sequence and complex proteomic world has become a major challenge. Careful *in silico* analysis can significantly reduce the amount of lab work required. In the present work, we have used a combination of computational and molecular biology techniques for the prediction and subsequently confirmed the alternatively spliced transcript variants of neurotransmitter receptor genes in mouse. Various gene/exon finding tools were used for the prediction of exons. The dry laboratory results were further validated and confirmed in wet laboratory through molecular biology techniques.

1.2. Aims and objectives

The general aim of this study was to use a combination of bioinformatics tools and molecular biology techniques for identification of alternatively spliced isoforms of few cancer associated genes from mouse genome which has not been annotated or identified earlier. The specific objectives were:

1. Identification of genomic sequences, transcripts and protein isoforms of cancer associated genes through database and literature survey.
2. Computational analysis of selected genes for possible alternatively spliced transcript isoforms using online available bioinformatics tools.
3. Validation of computational results in wet lab using molecular biology approach.
4. To identify whether isoforms are expressed differentially in different tissue types and temporally during different post-natal stages of development.
5. To characterise the conceptually translated protein isoforms, generated by alternative splicing by \textit{in silico} analysis.

\section*{1.3. Research hypothesis}

We have devised a pipeline/scheme employing use of computational technologies that include existing web-based tools and database systems and services for gene analysis, gene modelling, sequence comparison etc. Using this combinatorial approach, we predicted and confirmed sequences of several novel exons of cancer associated genes, and their presence in transcripts that have not been annotated earlier. The post translational modification studies of these alternative isoforms also showed their importance with respect to diverse functionality. Thus, imparting diverse and distinct specificities together with their differential expression in different tissues and developmental stages.

We propose a hypothesis that genes having large 5' untranslated sequences and relatively large first introns could produce several alternatively spliced transcripts which may encode protein isoforms with different N-terminals. This may help in achieving subcellular, cellular, tissue and developmental specificities. Our study could be of help in elucidating similar splicing events in humans where in many cases we found sequence similarity. The approach used in this study may be a way out to identify rare and unknown isoforms for other genes as well. Also the novel isoforms sequenced in this study might be implicated in different disease conditions in which case they can be targeted with isoforms specific drugs.

\section*{1.4. Thesis structure}

The aim of our study was to identify alternatively spliced new variants of cancer associated genes in mouse. Cancer associated genes mentioned in the list provided by Sanger institute (http://www.sanger.ac.uk/research/projects/cancergenome/) were screened through literature survey for performing computational analysis. The \textit{in silico} analysis of around sixty genes were carried out for prediction of novel isoforms. Based on the analysis, eight genes with high prediction score conforming to our selection criteria were selected for validation of computational results in wet lab using molecular biology approach. After initial experimental results of the selected genes, we have chosen three genes for the thesis work.
The subject matter is arranged in chapters and sections. The problem under investigation, aim and objectives of the study and proposed hypothesis is presented in this chapter. Chapter 2 explains our current knowledge on splicing, alternative splicing, methodologies for studying alternative splicing, association of alternative splicing with cancer and its possible treatment, myriad functions of alternative splicing, importance and evaluation of bioinformatics tools for solving the isoform repertoire of genes and genomes. Chapter 3 describes the computational methods used in this study, materials used in wet lab and wet lab experimentation to confirm the predicted alternatively spliced transcript variants of genes. The protein variants were studied on the basis of availability of specific antibody. The specific procedures for each gene analysis, primers and PCR conditions are given in “Materials and Methods” section of each chapter. Chapters 4 describe the prediction and validation of three newly identified alternatively spliced transcript variants of Arnt gene and their differential expression across tissues. Promoter regions and diverse post translational modifications were characterized by in silico analysis. Chapters 5 describe the identification of new alternatively spliced variant of Stk11 gene employing 5' RACE methodology. The expression of new variant across different tissues was observed at postnatal-7 stage and adult stage. Chapter 6 again describes the two novel N-terminal non-coding exons of MyD88 gene identified by bioinformatics tools and subsequently by RT-PCR and DNA sequencing. In silico analysis of new promoter region and various post translational modifications were also carried out. The expression of new larger variant of the protein was also detected using anti-MYD88 antibody. Chapter 7 summarizes the research findings, their implications, significance and future prospective. In each chapter designations of exons and transcripts has been done differently to avoid overlap.