INTRODUCTION

1.1 Motivation

The last decade has seen a huge spurt of activity in genome sequencing. With improved technologies and reducing cost, more than 1000 viruses, 100 microbes and 11 eukaryotic whole genomes have been sequenced so far. Such a massive amount of data available in the public domain opens a whole array of possibilities to understand the mechanism of living organisms in detail. This revolution is likely to boost both the basic and applied science of various fields with opportunities for better food, health, and environment.

The highlight of all sequencing efforts is undoubtedly the announcement of the finished human genome sequence in summer 2003 by the International Human Genome Sequencing Consortium (IHGSC, 2001). This landmark achievement of a species reading its own genomic content is just the beginning rather than the end. Already progress is underway to tap this potential and understand the making and working of this complex organism. However, our current understanding is more limited and even defining complete functions of a single celled microorganism remains an uphill task. Nevertheless, recent high-throughput techniques, with supporting bioinformatics tools, have thrown out exciting results. Even complex human behaviours, like homosexuality and handedness are now linked to genes (Gibson and Dormor, 2003; Van Agtmael *et al.*, 2003). These are great surprises as scientists traditionally correlated these characters to environmental, social and cultural factors than genes. Such results emphasise the old genetic understanding that phenotype is the result of both genotype and environment even in complex human behaviours. Genotyping the expression of genes and their functions at molecular, cellular and physiological levels will answer such enigmatic questions in biology. This was emphasised again with the availability of two complete genomes – *Drosophila melanogaster* (Celniker *et al.,* 2002) and *Caenorhabditis elegans* (The *C. elegans* Sequencing consortium, 1998). Drosophila, having more complex developmental stage and nervous system, has fewer genes than the 1mm long soil nematode with only 959 cells in total.

To understand the functioning of organisms, it is necessary to know where and when a gene is expressed. The first step in this process is to identify the number of genes in the organism and map them in the genome. Unfortunately this has been a difficult task due to various issues such as intervening sequences (introns), pseudogenes and repetitive elements. In humans, we are still not clear about the exact number of genes. However research so far has helped to narrow down the number to around 30,000 (IHGSC, 2001). This is significantly lower than the 120,000 predicted sometime back (for discusssion, Ashurst and Collins, 2003; Ewing and Green, 2000; Liang *et al.*, 2000). Although the number might seem to be low for a complex organism, the number of transcripts produced from these genes is quite high as a result of alternative promoters, splicing and polyadenylation. In humans, it is estimated that an average of 2.5 alternative transcripts are produced per locus (Ashurst and Collins, 2003).

Until now, gene identification in the genomic sequence has been mainly focused on protein coding genes with less attention paid to pseudogenes, non-coding RNA genes and internal (embedded) genes. Non-coding RNA genes include an array of different types of regulatory RNA genes with newer types still appearing (Cawley *et al.*, 2004; Mattick, 2001).

Identifying, mapping and confirming the presence of these genes and different regulatory signals in the genomic sequence is referred to as *annotation*. This is done using an ensemble of different experimental and computational tools, with computational approaches usually facilitating the initial steps. Many gene prediction algorithms, such as Genewise (Birney and Durbin, 1997) rely on evidence from the alignment of EST, mRNA or protein sequences to the genome. Such algorithms generate accurate gene predictions, but only where expressed sequence data is available. Here, I am interested in *ab initio* methods that can predict from genome sequence alone. *Ab initio* gene prediction programs used in annotation can be broadly classified into comparative and non-comparative methods depending on whether they predict from an alignment of genome sequences or a single genome sequence. To date the majority of work was done using non-comparative *ab initio* algorithms and are based on different methods, namely neural networks (example programs include *GRAIL* (Uberbacher *et al.*, 1996), *GENEPARSER* (Snyder and Stormo, 1995)), discriminant analysis (*HEXON* (Solovyev *et al.*, 1995), *MZEF* (Zhang, 1997)) and hidden markov models (*GENSCAN* (Burge and Karlin, 1997)). Besides these, there are other old methods such as rule-based

methods (*GENEID* (Guigo *et al.,* 1992), *GENEFINDER* (Wilson *et al.,* 1990)), linguistic methods (*GENLANG* (Dong and Searls, 1994)) and decision trees (*MORGAN* (Salzberg *et al.,* 1998)). Some programs were developed by combining different methods and *GENIE*, an example, combines hidden markov models and neural networks (Reese *et al.,* 2000). Few other *ab initio* gene prediction programs, like *QRNA*, were developed to detect non-coding RNA genes (Rivas *et al.,* 2001). Reviewing all these methods and programs is beyond the scope of this chapter and hence I refer the reader to these reviews (Mathe *et al.,* 2002; Zhang, 2002).

In general, *ab initio* gene prediction programs use sequence signals and coding measures to predict gene structures. Coding measure (a feature measured computationally but not used by the biological system) is the important component as it is likely to differentiate exons (coding sequences) from introns (intervening sequences). However, this limits the identification of pseudogenes and non-coding RNA genes and the performance of the gene prediction programs are poor even in simple cases (Rogic *et al.,* 2001). So, a gene prediction program based purely on DNA regulatory signals is likely to overcome this problem. Towards this future objective, I attempt to develop prediction models that can efficiently detect signals from genomic sequence context.

Before describing my research objectives, I devote the rest of this chapter to introduce the basics of gene structure, different regulatory signals in the DNA sequence and the process of transcription and translation.

1.2 An overview of gene structure

A typical higher eukaryotic protein coding gene, as depicted in Figure 1, has a defined promoter region with exons and introns splitting the transcription unit. Transcription initiates from a transcription start site and terminates a few hundred bases downstream of the cleavage site. Exon and intron boundaries are marked by the donor and acceptor splice site regions and on pre-mRNA maturation, introns get spliced out by the spliceosome complex. The 5’ cap and 3’ poly(A) tail added to the matured transcript play major roles in mRNA stability, export and translation initiation (Manley, 2002; Proudfoot *et al.,* 2002). Processed and stable transcripts, exported to cytoplasm, are translated by the translation machinery in the cytoplasm with start and stop codon acting as its signals. Traditionally, as

noted in *in vitro* experiments, transcription, splicing, capping, polyadenylation, termination and export were considered to be independent of each other. However latest research suggests that all these processes occur co-transcriptionally with the carboxy-terminal domain (CTD) of RNA polymerase II playing a major role (for review see, Neugebauer, 2002; Proudfoot *et al.*, 2002).

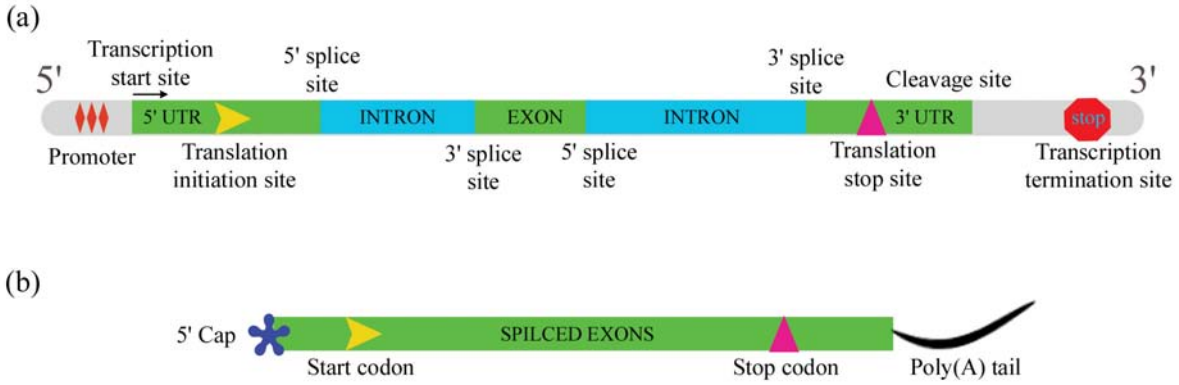

Figure 1. Schematic diagram showing (a) Typical gene structure of protein coding gene transcribed by RNA polymerase II (b) Matured RNA transcript with 5' cap and 3' poly(A) tail.

1.3 Defining transcription termination

Transcription termination has been defined to have two major steps: release of the transcript from the elongating polymerase and the dissociation of the polymerase complex from the DNA. An accurate and efficient system is required to pursue this function as the elongating polymerase would otherwise run-over into the adjacent transcription units. In yeast, many such cases have been reported in places where genes are closely spaced (Greger *et al.*, 1998). Also, terminating transcripts allow recycling of the polymerase and stops unnecessary transcription of intergenic regions. Various biological systems have been employed to understand this mechanism for many years now. All the results show termination can occur either depending upon bipartite or tripartite sequence components or on a stem-loop secondary structure basis. Here, I present a brief overview of the different termination systems identified so far.

1.4 Transcription termination in prokaryotes

Most prokaryotic genes do not have introns and the DNA is not isolated as prokaryotes do not have nucleus. Therefore, coupled transcription and translation is a common mechanism. Also, unlike eukaryotes, prokaryotic genes are transcribed by a single RNA polymerase.

Termination of transcription in prokaryotes is widely found to occur in two ways depending on the requirement of the protein factor, rho (reviewed in Henkin, 1996). In the 'intrinsic' or 'rho-independent termination', a G+C rich stem-loop structure followed by a series of U residues at the end of the transcript, hinders the proceeding polymerase and thus pauses, destabilizes and releases from the DNA (Figure 2). In 'rho-dependent termination', the protein factor, rho hexamer binds to a *rut* (rho utilization) site on the 3' end of the transcript. This RNA:protein interaction brings a change in the elongating polymerase resulting in the release of transcript and dissociation of polymerase by hydrolysing ATP as the energy source (Figure 3).

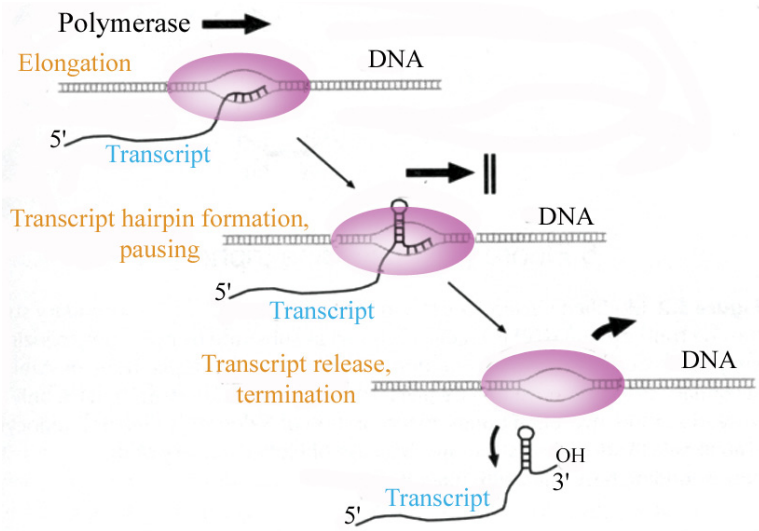

Figure 2. Rho-factor independent transcription termination in prokaryotes.

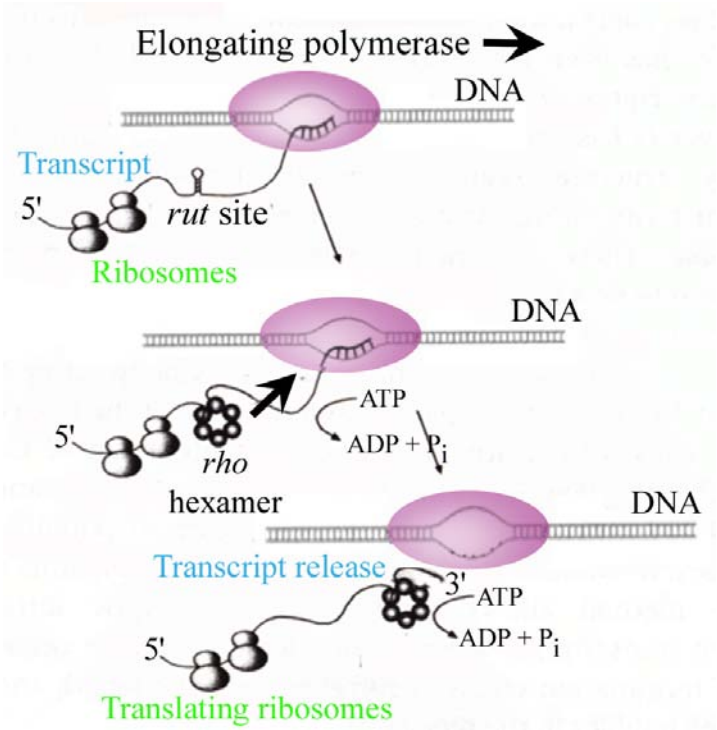

Figure 3. Rho-factor dependent transcription termination in prokaryotes.

However, in both mechanisms, termination is due to pausing of RNA polymerase at a specific site followed by destabilization of the complex due to the formation of a RNA:DNA hybrid in the transcription bubble and changes in the processivity of the polymerase (Henkin, 1996).

1.5 Transcription termination in eukaryotes

Unlike prokaryotes, eukaryotic transcription termination is complicated as there are three different types of polymerases responsible for transcribing various types of RNA molecules.

1.5.1 Polymerase I transcription termination

Transcription termination of Polymerase I, that syntheses rRNA, is mediated by protein factors reb1p in yeast (Lang *et al.,* 1994; Lang and Reeder, 1993) and TTF-I in mouse (Evers *et al.*, 1995). Polymerase I terminator sequence has two components: a binding site for the protein factor and an upstream element that codes for the last 10-12 nucleotides of the terminated transcript (Figure 4). The reb1p/TTF-I factor binds the DNA sequence element in the correct orientation and pauses the elongating polymerase. This halt stimulates the release of the transcript and dissociation of the complex. TTF-I is also found to recruit

additional releasing factors for this process. However reb1p does not require any additional factors and the dissociation of the transcript depends only upon the instability of RNA:DNA hybrid in the active site of the polymerase due to stretches of A:U base pairing (Reeder and Lang, 1997).

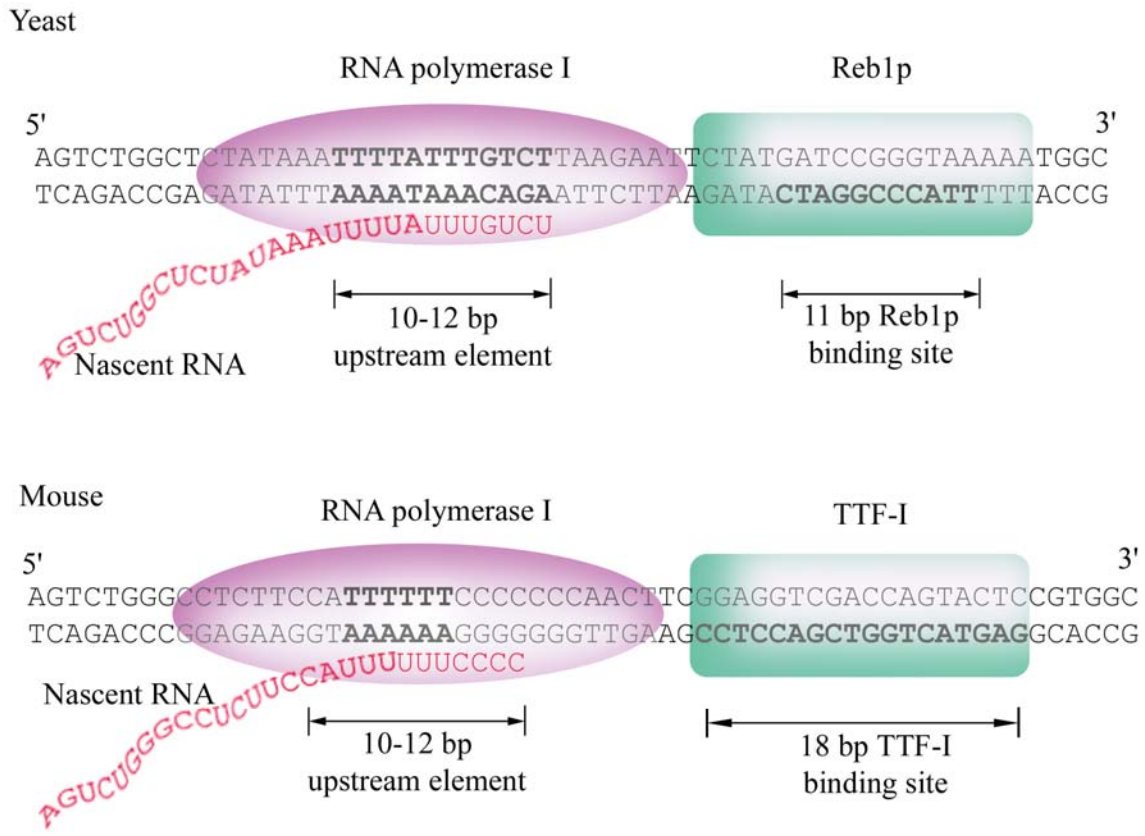

Figure 4. Structure of RNA polymerase I terminators from yeast and mouse.

Reb1p binding site was also found to have partial pausing activity for Polymerase II in the forward orientation and no activity in the reverse orientation (Lang *et al.*, 1994).

Polymerase I gene terminators are found to behave as DNA replication terminators as well. Bi-directional replication forks proceeding from the nearby *ori* site are stopped by the barrier created with TTF-I:DNA interaction. This barrier function is orientation dependent but has opposite polarity to transcription termination (Gerber *et al.*, 1997). However such a function is yet to be proved for yeast reb1p protein.

Polymerase I terminators are different from prokaryotic terminators as there are no inverted repeats and thus there is no hairpin structure formation and requirement of orientation-specific DNA binding proteins.

1.5.2 Polymerase III transcription termination

RNA Polymerase III responsible for transcription of tRNA, 5S rRNA and U6 snRNA can recognize termination sites accurately and efficiently without any requirement for protein factors (Cozzarelli *et al.,* 1983) and bring about termination with a simple cluster of four or more T residues (Bogenhagen and Brown, 1981). However, efficiency of release of paused polymerase was shown to improve with the recruitment of PTRF factor. Attempts to prove the requirement of La auto-antigen in Polymerase III transcription termination remains inconclusive (Lin-Marq and Clarkson, 1998; Maraia *et al.,* 1994; Yoo and Wolin, 1997).

1.5.3 Polymerase II transcription termination

RNA Polymerase II responsible for transcription of the remainder and vast majority of genes and is the subject of the work described in this thesis.

Polymerase II transcription termination occurs at least in three different ways depending on the gene it is transcribing, namely, snRNA and snoRNA genes, histone genes and protein coding genes. Before embarking into the details of these mechanisms, it is necessary to understand the 3'-end processing signals of protein coding genes.

The 3'-end processing involves an endonucleolytic cleavage of the nascent transcript and subsequent addition of poly(A) tail to the newly formed 3'-end. This process thought to occur for all transcribed genes along with capping and splicing of introns makes a nascent RNA matured. The 5'-cap and 3'-poly(A) tail have been found to have major roles in mRNA stability, export, translation initiation and other events. Endonucleolytic cleavage at the 3'-end of the transcript occurs at the cleavage site that has a consensus sequence of CA dinucleotide (Sheets *et al.,* 1990), flanked by a highly conserved poly(A) signal at the 12-30 bases at the upstream region and U-rich and or GU-rich motif immediately at the downstream site (Zarudnaya *et al.,* 2003) (Figure 5). In the majority of mammalian pre-mRNAs, the poly(A) signal is found to be composed of AAUAAA or AUUAAA

(MacDonald and Redondo, 2002) and has been suggested to be required for effective splicing (Cooke *et al.,* 1999) and transcription termination (Edwalds-Gilbert *et al.,* 1993; Yeung *et al.,* 1998) as well as for polyadenylation. The 160 kDa subunit of the cleavage and polyadenylation specificity factor (CPSF) binds to this hexamer element while the 64 kDa cleavage stimulation factor (CstF) binds to the U-rich sequence immediately downstream of the cleavage site. The binding of these factors is co-operative and each factor enhances the affinity of other factors towards its binding site. Once the processing site is recognized, two cleavage factors (CF I and CF II complex) get recruited and cleave the nascent transcript at the cleavage site. To the newly formed 3'-end, poly(A) polymerase (PAP) adds at least 250 nucleotides of adenine. Poly(A) binding proteins (PABP II) bind to this stretch of adenine nucleotides which enhances the stability of the tail. Although release of the transcript occurs after the cleavage at the cleavage site, RNA polymerase does not get released from the DNA at this site, but several hundred bases downstream.

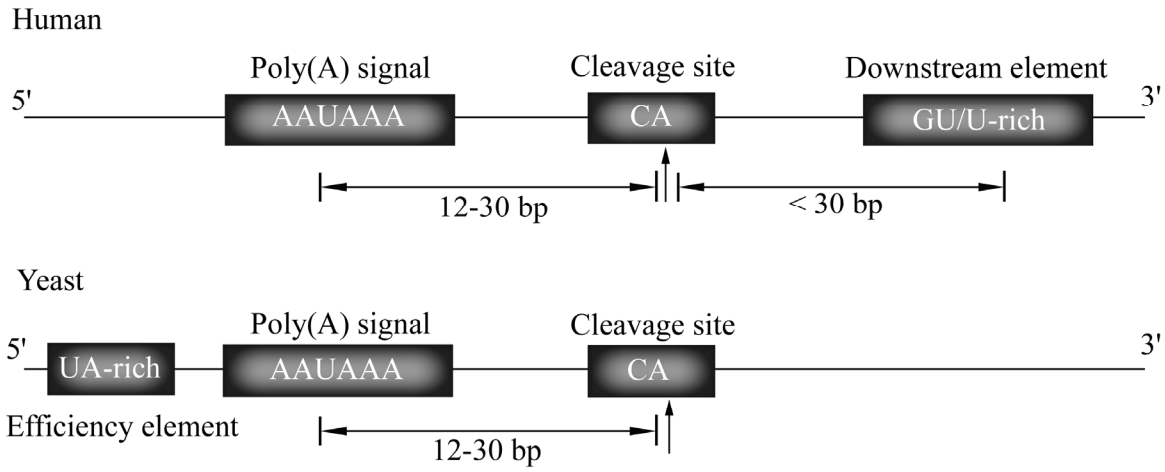

Figure 5. Schematic representation of 3'-end processing signals in human and yeast.

Determining the exact position of the polymerase release has been a challenge to study as the 3'-end product of the cleavage has very short half-life and the maturation of the 3'-end of the transcript (cleavage and polyadenylation) occurs co-transcriptionally. Fortunately, nuclear run-on assay can trap such nascent transcripts and help in analyzing transcription termination.

In the nuclear run-on technique, the nuclei transcribing a specific gene is isolated and allowed to incubate with radiolabelled ribonucleotide triphosphates for incorporation in the

newly synthesized RNA molecules. This labeled nuclear RNA is then purified and hybridized to Southern blots of DNA probes carrying the gene sequence. The hybridization techniques equate directly to the polymerase density at the position of the probe and hence the point at which signal is no longer detectable corresponds to the site of termination. A gradual decrease in polymerase density always occurs downstream of the cleavage site. However in many instances before this decrease, a short higher polymerase density site is noticed. This is referred to as the *pause site*.

Now it is understood that transcription termination requires the 3'-end processing signals and a pause site. However 3'-end maturation does not require termination of the transcribing polymerase. In fact, both transcription termination and 3'-end processing processes are found to be coupled *in vivo* (Birse *et al.,* 1998; Dichtl *et al.,* 2002b) and are largely facilitated by the carboxy-terminal domain (CTD) of rpb1, the largest sub unit of Polymerase II.

Existence of pause site for termination has not been thoroughly accepted and there have been studies showing termination occurring without any requirement of pause site and with sole perturbation by poly(A) signal (Orozco *et al.,* 2002). However, several attempts have been made to identify consensus pause elements that are responsible to create a transient pause and thus enhance poly(A) signal recognition and termination.

Earlier studies identified an orientation-specific CCAAT element in the adenovirus late promoter that recruits CP1 protein and effectively terminates transcription from upstream genes (Connelly and Manley, 1989a, b). In yeast, Yhh1p, a subunit of CPF complex was identified to play this role (Dichtl *et al.,* 2002b).

In *Saccharomyces pombe*, both ura4 and nmt2 are found to possess downstream sequence elements that induce termination. These sequence elements are orientation specific and are composed of multiple and redundant signals. One of the sequence elements found in ura4 gene having pause activity, has two copies of pentanucleotide ATGTA with the last GTA playing an important role for binding an unknown factor responsible for pausing. However in the nmt2 gene, the pause elements are less compact and there is no homology with ura4

gene elements (Aranda and Proudfoot, 1999; Birse *et al.,* 1997). Similar pause sites were also found in α-globin genes and C2 and factor B genes (Yonaha and Proudfoot, 2000).

A detailed run-on assay in the mouse β-major globin gene identified a 69 bp AT-rich sequence that is active based on its position from the cleavage site (Tantravahi *et al.,* 1993). A similar experiment in human β-globin gene showed that a region 900 to 1600 bp downstream of the transcript cleavage site is essential for termination. Interestingly, it was also found that more cleavage of the nascent transcript occurs at this downstream termination region apart from the original cleavage site. These cleavages are termed as *co-transcriptional cleavage* and found to be necessary in addition to the 3' end processing signals for polymerase pause and release. However co-transcriptional cleavage was found to occur independent of 3' processing signals and thus deleting termination region does not affect 3' processing and vice versa. Nuclear run-on assay repeated on ε-globin genes found that the termination region is more diffuse than for the β-globin gene. Nevertheless the region is found to be as AT rich as the mouse globin gene, although the human region is longer (Dye and Proudfoot, 2001). Likewise, an A-rich 92 bp sequence at the 3' flanking region of human α2 globin gene is found to improve efficiency of upstream signals and thus processing events (Enriquez-Harris *et al.,* 1991).

Transcriptional studies in the intergenic region between human complement C2 and B genes showed the sequence element, GGGGGAGGGGG and the zinc-finger regulatory protein, MAZ that binds the sequence, can effectively stop transcription run-over from upstream genes and bring termination (Ashfield *et al.,* 1991). An upstream sequence element, mainly U-rich, was also found in human complement factor C2 and Lamin B2 gene (Moreira *et al.,* 1998).

Thus these experiments define various signals (CCAAT, ATGTA, AT-rich sequence, A-rich sequence and G-rich sequence) and factors (CP1, SP1 and MAZ) responsible for polymerase II transcription termination.

1.5.4 Computational detection of transcription termination signals

Apart from the experimental evidences mentioned above, a few related computational studies were also conducted around 500 bp upstream and downstream of cleavage and transcription start sites. Analysis on the cleavage site regions showed the common signals AATAAA and GT-rich sequence elements along with other signals. Prominent among them is CCCC, CCCTC and CCTCCC motifs. These motifs were also found peaking at -75 base pair and -200 to -100 bp upstream of transcription start site. Similarly the frequency of A_4 and G_4 motifs is higher before transcription start sites and after cleavage sites. Thus homo-oligomers A_{4-5}, G_{4-5}, T_{4-5} and C_{4-5}, C_{3-4} interspersed with T (CCCTC and TTCTT) and alternations of T and G (TGTGT) and GGAGG are found peaked around the 5' and 3' ends of genes (Nussinov, 1986a, b). Among all these signals, GTG/CAC and CTC/GAG DNA sequences are more interesting as they are frequently encountered in the regulatory DNA sequences and are likely target sites for several regulatory protein factors (Nussinov, 1986a). Another interesting result showed complementary signals on the same DNA strand have asymmetry behavior, i.e. the TGTGT peak patterns do not need to be the same for its complementary sequence, ACACA. This is more pronounced for complementary homo-polymers around transcription start site and cleavage site. This suggests some directionality in DNA bending and orientation-specific recognition by protein factors (Nussinov, 1986a). Similar signals were found in non-mammalian vertebrate DNA sequences as well (Nussinov, 1986b).

In another study (Nussinov, 1987) it was found that the distribution of the nucleotides showed opposite trends around the mammalian gene 5' and 3'-ends i.e., R_6 motifs (stretch of 6 purine residues) are found more frequently before transcription start sites, whereas Y_6 motifs (stretch of 6 pyrimidine residues) occur less frequently. In the 3' termini, Y_6 are less just before the end and R_6 motifs are more following it. In the non-mammalian vertebrate genes, these conditions are more pronounced. Two Y_6 peaks found at the 3' termini might be due to poly(C) and poly(T) residues. The R_6 peaks in the gene upstream might be due to high concentration of AGGG and GGGC and to a lesser extent of A_4. This G runs might contribute to the bendability feature of the DNA molecule (Figure 6, reproduced from Nussinov, 1987).

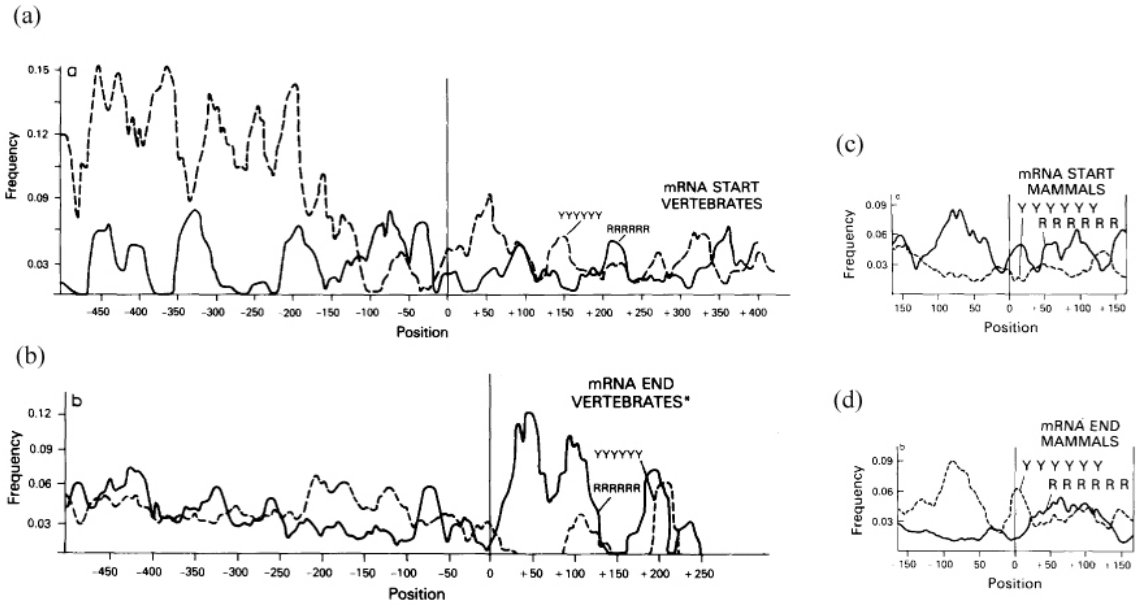

Figure 6. The nucleotide distribution of Y6 and R6 runs around transcription initiation and cleavage site. (a), (b) shows distribution in vertebrate mRNAs while (c) and (d) in mammals.

Although all these studies show that RNA polymerase II transcription termination signals are quite complicated, in general the system appears to work mainly based on two sequence components: 3'-end processing signal and pause sites. However these are not universal for all Polymerase II transcribed genes, as alternatives are found in histone and snRNA genes.

Histone genes are not spliced and the majority are not polyadenylated. The mature 3'-end of the transcript is formed by the endonucleolytic cleavage of the primary transcript and polymerase terminating in the A-rich sequence flanking the 3'-end (Briggs *et al.*, 1989). This cleavage is enacted by the stem-loop structure formed upstream of the cleavage site (roughly, 600 bp in case of H2A gene). The sequence at the stem-loops are well conserved with GGYYYU in the stem followed by a four-base loop, UYUN and the complementary sequence ARRRCC (Lanzotti *et al.*, 2002). Specialized protein factors called SLBP bind to this structure and stabilize the transcript mimicking the role of a poly(A) tail (Johnson *et al.*, 1986; Lanzotti *et al.*, 2002; Zanier *et al.*, 2002). The cleavage site efficiency is improved by a downstream element interacting with the U7 snRNP. Thus, both SLBP and U7 snRNP together recruit a complex capable of performing pre-mRNA processing reactions.

Polymerase II termination in snRNA genes require a 3' box element located 9-19 nt downstream of the end of the nascent transcript. U1 transcripts terminate just after the 3'

box whereas U2 snRNA nascent transcripts are found up to 250 nucleotides downstream. These results show a 3' box, a RNA processing signal and a downstream signal where interaction between protein-DNA leads to termination (reviewed in, Hernandez, 1992). HeLa cells with transfected constructs of 3' box and downstream sequence elements confirmed the termination activity of these signals (Cuello *et al.*, 1999). This mechanism sounds similar to the mRNA bipartite termination process, requiring RNA processing signals and the termination elements.

However, poly(A) signal or 3' processing signals are not essential for all cases of termination and recently it was reported in yeast that there is poly(A)-independent Polymerase II termination mechanism for snRNA and snoRNA genes with protein factors Nrd1 and Nab3 complex, Sen1 helicase and the CTD domain of Polymerase II (Steinmetz *et al.*, 2001).

1.5.5 Transcription termination models

Based on the available experimental evidence, three different models of Polymerase II transcription termination have been proposed.

(i) In the 'RNA cleavage' or 'torpedo' model, cleavage occurs firstly at the cleavage site, leaving two products: the upstream RNA later forming a matured transcript and a 3' product still attached to the elongation complex. Rapid degradation of this 3' product by the 5'→3' exonuclease aided by helicase, 'catches up' the elongating polymerase and triggers termination (Proudfoot, 1989). However, recent evidences suggests cleavage is not necessarily required for termination to occur (Osheim *et al.*, 1999).

(ii) In the 'polymerase change' or 'anti-terminator' model, Polymerase II complex upon passage and recognition of poly(A) signal, undergoes conformational changes in the complex making it termination competent; this results in pause and release from the DNA (Logan *et al.*, 1987).

(iii) Recent experiments show that both these models are not mutually exclusive and a combination of both might exist (Proudfoot *et al.*, 2002). In the combined model, a co-transcriptional cleavage occurs at the downstream termination site first, with still

interaction between the CTD of the polymerase and 3' end processing signals remaining active. Subsequently, 'polymerase change' occurs in this interaction leading to cleavage at cleavage site and polymerase release.

1.6 Splicing and transcription

Recent experiments have clearly indicated that transcription and mRNA processing occur together and all the steps in the mechanisms are linked with each other, with the CTD of the RNA Polymerase II itself playing a major role. Therefore it is important to know about the splicing process, where intervening sequences were removed from the pre-mRNA to form mature mRNA, ready for translation.

1.6.1 Splicing mechanism

Exons and introns are determined by their boundary sequences with definite consensus patterns. Introns predominantly start with GT and end in AG dinucleotide. Figure 7 (reproduced from www.sanger.ac.uk/HGP/Chr22/cwa_archive/splice_site_analysis.shtml) shows the nucleotide distribution calculated from 3,673 introns from human chromosome 22. These predominant splice signals are called canonical splice sites and they form the basis for the GT-AG splicing rule (Mount, 1982). However, apart from the GT-AG rule, other intron boundaries, GC-AG and AT-AC, were also reported (Burset *et al.*, 2001). Also along with the splicing boundary signal, another consensus pattern called Branch Point Sequence (BPS) was found to be present upstream of AG dinucleotide and shown to be required for the splicing process.

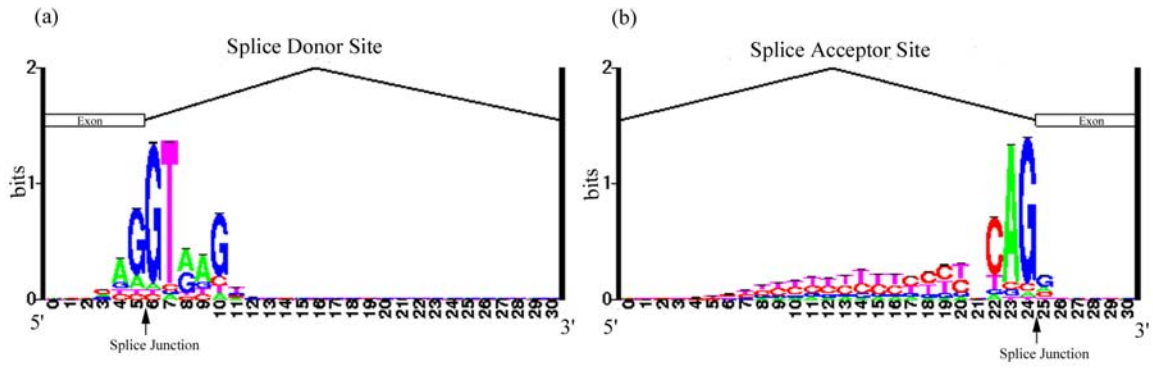

Figure 7. Nucleotide Distribution at Donor and Acceptor site analysed from 3,673 introns from human chromosome 22

Splicing of introns is mediated by a mega-Dalton RNA-protein complex formed with snRNA (small nuclear RNA) and around 50 to 100 protein molecules. The details of this complex mechanism is beyond the scope of this chapter, however, I will briefly cover some important aspects of the splicing process (Figure 8).

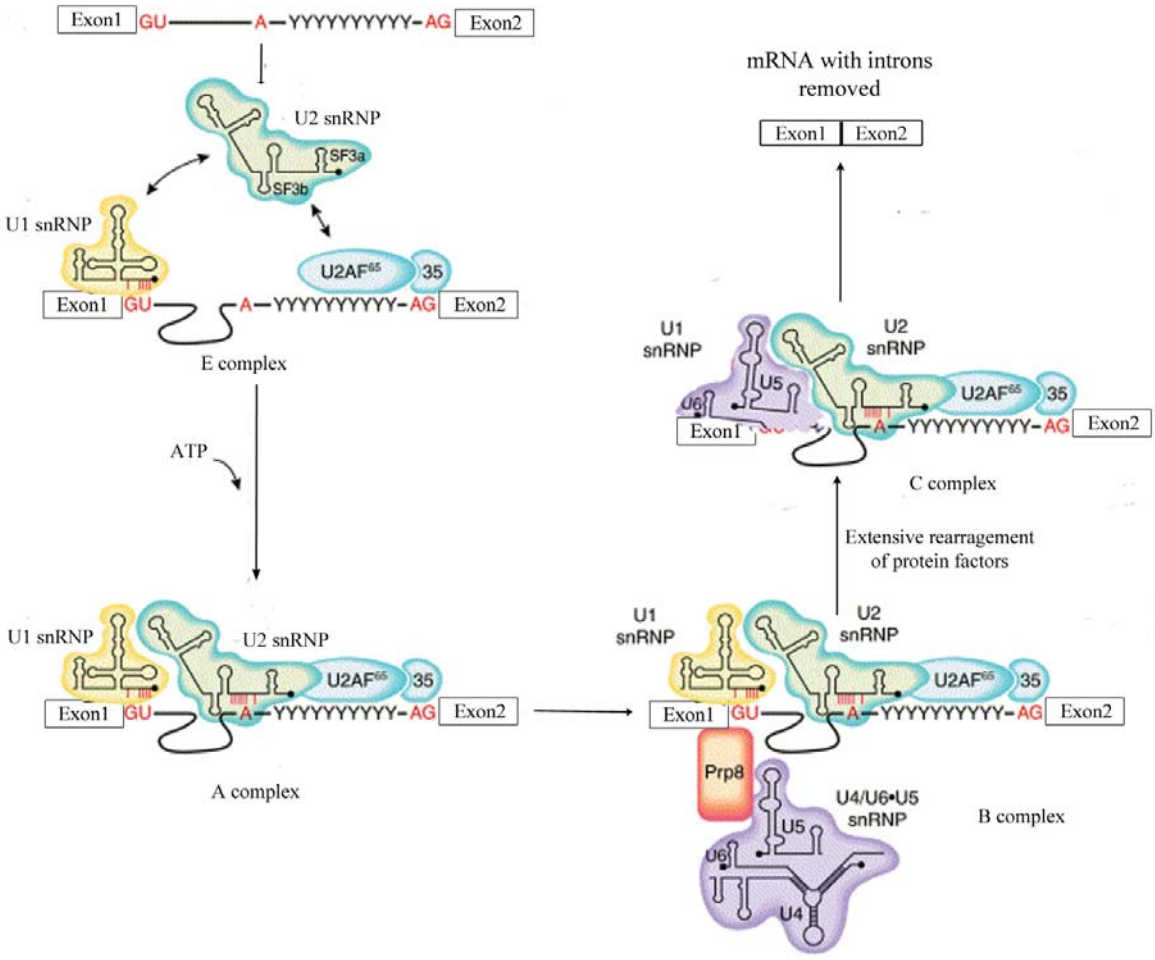

Figure 8. Splicing mechanism where introns are spliced and exons are linked.

The splicing of an intron from a nascent RNA is a two step process requiring two distinct trans-esterification reactions. Initially, cleavage occurs at the donor splice site (the site where introns start) facilitating the first base of the intron to form a lariat structure with the BPS signal present upstream of the acceptor site (the site where introns end). This step is referred to as *branching*. Next, a new phosphodiester bond is formed between the last base of the upstream exon and the first base of the downstream exon. The intron is then released (Jurica and Moore, 2003). These reactions occur within the spliceosome complex,

responsible for recognizing splice sites and catalyzing the reactions. The spliceosome is largely made up of five RNA-protein complexes known as small nuclear ribonucleoproteins (snRNPs).

Before cleavage at the donor site, the signals at this site are recognized by U1 snRNP with the formation of commitment complex (E complex). This process does not require any energy component like ATP and it was noted recently that the step is not a strict requirement, as introns were found spliced efficiently *in vitro* even in the absence of U1 snRNP (Crispino *et al.*, 1996). A key role of U1 snRNP complex is to promote the association of U2 snRNP complex with the BPS signal. This interaction is dependent on two other interactions – $U2AF^{65}$ with the polypyrimidine tract of the BPS and $U2AF^{35}$ with the intron terminal AG dinucleotide (reviewed in Reed, 2000). This step is an ATP dependent process where six proteins, including DEAD box protein UAP56 and components of essential splicing factors, SF3a and SF3b, bind either upstream or downstream of the BPS. The association of U1 and U2 snRNPs defines complex A.

Association of the tri-snRNP complex containing U4, U5 and U6 snRNPs with the complex A is required to form complex B. This interaction was recently found to be promoted by the splicing factor SPF30 although this transition remains poorly defined (Rappsilber *et al.*, 2001). This tri-snRNP complex interacts with the donor and acceptor splice sites, recruits other factors including the highly conserved Prp8 protein, and forms the catalytic core of the spliceosome (reviewed in Jurica and Moore, 2003). Although the complete role of this catalytic core is under investigation, it is understood that the tri-snRNP brings about a series of RNA-RNA rearrangements, with the displacement of U1 snRNP from the donor splice site by U6 snRNA, creating the catalytically competent C complex. These rearrangements have been found to be directed by an RNA helicase of the DExD/D box protein family (Schwer, 2001).

The catalytically competent C complex facilitates the second trans-esterification reaction between the upstream and downstream exon with the excision of the spliced intron and mature mRNA.

In yeast, branching occurs with an almost invariant BPS signal UACUA**A**C (with branch A in bold letter), 20-30 nucleotide upstream of the acceptor splice site. The mammalian BPS is less well conserved but generally conforms to the consensus YNYUR**A**Y signal. Recognition of BPS is mediated by base pairing of an invariable sequence in U2 snRNA. It has been suggested that the Branch Point (BP) nucleotide is bulged out from this RNA duplex and this may activate the 2' hydroxyl group for nucleophilic attack. The natural BP nucleotide is adenosine; however, exceptions have been reported. For example, branching of the first intron of the human growth hormone gene and the third intron of the human calcitonin/CGRP gene occur mainly at a cytosine and uridine residue respectively (Adema *et al.*, 1988; Hartmuth and Barta, 1988).

Branching in higher eukaryotes requires other elements found near BP nucleotide. In human, the branch sites map 18-37 nucleotide upstream from the highly conserved AG dinucleotide separated by a polypyrimidine tract of variable length. The length and uridine content of the tract are important factors for branching. At a very early step in spliceosome assembly (complex E formation) the $U2AF^{65}$ and SF1 bind the polypyrimidine tract and the BPS signal. SF1 recognizes primarily the two most conserved nucleotides in the BP sequence YNY**U**R**A**Y. In addition, a direct interaction between SF1 and $U2AF^{65}$ has been demonstrated that may account for the coupled recognition of the BP sequence and polypyrimidine tract.

For the second trans-esterification reaction, the conserved AG dinucleotide, at the acceptor splice site, plays a highly important role and usually the first AG dinucleotide downstream of the BPS is generally used. This is probably selected by a scanning mechanism.

The limiting step in the whole splicing process lies in the recognition of the intron itself. In yeast, where introns are short, the spliceosome is thought to form directly on the intron through the process of *intron recognition* (Talerico and Berget, 1994). However, in human, where short exons are interrupted by long introns, recognition is thought to be based on an alternative model called *exon recognition* (Berget, 1995). Recognition in both models involves splicing associated SR proteins, which play a major role in bringing spliceosome components together (Graveley, 2000).

The splicing process explained so far obeys the normal GT-AG rule and the spliceosome is referred to as the U2-type. However, there is another set of donor and acceptor sites, which displays the AT-AC rule. These splice sites are utilized by a distinct spliceosome called U12 spliceosome that contains U11, U12, U4atac, U6atac and U5 snRNPs (Tarn and Steitz, 1996, 1997). Interestingly, U12-dependent system is lacking in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*.

Distinct differences have been observed between U2- and U12-dependent types of introns. U12-dependent signals exhibit strongly conserved and informative donor and branch signals whereas U2-dependent ones exhibit only moderately informative signals at the donor and acceptor sites and a highly degenerate BPS. Additionally, the polypyrimidine tract found in U2-dependent introns is either not present or weaker in U12-dependent introns (Will *et al.,* 1999).

However, both the systems are not entirely independent of each other and are often found to evolve together. Recent results have found a strikingly high degree of similarity of overlap between the proteins and non-coding RNAs of both systems. These include U5, Prp8, 8 snRNP Sm proteins, SF3b components and SR proteins. Moreover similarity in secondary structures and interactions between the set of non-coding RNAs U11, U12, U4atac and U6atac and the set of U1, U2, U4 and U6 in U2-dependent systems argue that both the systems are homologous to each other (Hastings and Krainer, 2001; Schneider *et al.,* 2002; Will *et al.,* 2001; Will *et al.,* 1999).

1.6.2 Roles of splicing

In transcription: The role of introns in the genome and their probable function has been a fascinating area of study for quite sometime. Along with other functions reported, introns are considered to be a rich source of regulatory elements with the first introns having most elements (for details see, Le Hir *et al.,* 2003; Mattick, 1994; Salamov *et al.,* 1998a). For example, the 280 nucleotide regulatory elements in the first intron of the *c-myc* gene blocks transcription elongation (Pan and Simpson, 1999). In mice, intronless transgenes are transcribed 10-100 times less efficiently than their intron-containing counterparts (Brinster *et al.,* 1988; Le Hir *et al.,* 2003). In yeast, promoter proximal introns enhance transcription

initiation with the association of U1 snRNA and initiating factor, TFIIH (Kwek *et al.*, 2002).

The role of CTD of RNA polymerase on the recruitment of processing factors and mRNA maturation has been well established. However, with the latest studies it was shown the communication actually goes both ways, with the assembling spliceosome providing positive feedback to the polymerase. The tat-specific factor (TAT-SF1) recruited on newly transcribed introns interacts with the kinase, pTEFb, capable of phosphorylating the C-terminal domain. This increased CTD phosphorylation is necessary for both promoter clearance and efficient transcription elongation.

Similarly, the interaction of spliceosome component with cap binding complex and poly(A) processing factors enhances the recognition of the 5' most and 3' most introns respectively. *In vitro* studies show an upstream 3'-splice site can significantly enhance use of a downstream polyadenylation site, and a downstream polyadenylation site can, likewise, increase excision of the 3'-most intron (Proudfoot *et al.*, 2002). Protein-protein interaction experiments confirm that both the snRNP protein U1A and SRm160 (SR-related matrix protein of 160 kDa), a splicing co-activator, interacts with the cleavage-polyadenylation specificity factor, CPSF 160. Furthermore, interactions between the C terminus of poly(A) polymerase and the splicing factor U2AF65 and U1A can enhance upstream 3'-splice site recognition (Vagner *et al.*, 2000).

These interactions between splicing and transcription components are not only related physically but temporally too. In α-TM, constitutive splicing factors bind to the splice site signals of exon 3, committing it to the normal splicing pathway. In regulated splicing, an alternative set of factors are thought to bind to the URE and DRE in the flanking introns, forming an inhibitory complex for constitutive splicing. Delaying the transcription of the DRE element through the introduction of some spacer sequences and hindering the regulated splicing complex formation, and removed the inhibitory effect. This indicates that on transcription, splicing factors along with its regulators are available and the decision for constitutive or regulated transcription can occur due to the lag between the transcribing polymerase and splice site and the relative distance between competing elements. Thus, the

rate of transcription and the pausing of the polymerase while transcribing might decide the processing pathways (Roberts *et al.,* 1998).

In translation: In Xenopus, splicing was reported to influence translational efficiency as well without significantly altering the steady-state cytoplasmic mRNA levels. When a mature mRNA is injected directly into oocyte nuclei, it is translationally repressed after export to the cytoplasm. This repression can be overcome with a spliceable intron in the 3' UTR. Splicing can apparently enable an mRNA to escape masking of mRNPs and to actively engage ribosomes (Braddock *et al.,* 1994). In another experiment, Matsumoto *et al* found that an intron placed in the 5' UTR was highly stimulatory, whereas the same intron placed in the 3' UTR repressed translation to below the level of the corresponding intronless mRNA (Matsumoto *et al.,* 1998).

In pre-mRNA processing: Apart from influencing transcription and translation processes, adjacent introns in a pre-mRNA affect one another's splicing efficiency too. Results from related experiments form the basis for an *exon recognition* model, which depicts that the acceptor splice site of an upstream intron helps to increase the efficiency of recognition of the donor splice site of a downstream intron through components of the splicing machinery and vice versa. The interactions, which link the upstream acceptor splice site and the downstream donor splice sites, involve U1 snRNP and U2AF65 and these are thought to be mediated by SR proteins. SR proteins generally possess one or two RNA-binding domains (recognition motifs, RRM) and an arginine-and-serine rich region (the RS domain). RRMs often target SR proteins to exonic splice enhancers. The RS domain then appears to provide a molecular 'glue' allowing RS-RS interactions between interacting factors and thus facilitating the recognition of intron-exon boundary by the splicing apparatus (Graveley, 2000).

These SR proteins are also found associated with the CTD and are either referred to as CTD-associated SR-like protein or SR-like CTD-associated factor. The heptad-repeat sequence of CTD is micro-heterogeneous and that might result in different levels of phosphorylation and affect significant levels of SR-protein interaction with CTD (Graveley, 2000).

Transcription and Splicing Rate: Using a well-documented alternatively spliced intron from the highly intronic gene for fibronectin, it was shown that different types of promoters initiate various splicing pattern of transcripts (Cramer *et al.*, 1997). Over expressing various SR proteins are also found to affect the splicing patterns, sometimes antagonizing the promoter effects (Cramer *et al.*, 1999). These results are consistent with a model in which SR-protein interactions with the CTD are set up early in the transcriptional initiation process. Also, the correlation between transcriptional rate and splicing was also shown previously (Roberts *et al.*, 1998). When transcription slows down its rate on specific parts of the gene, it might influence the splicing patterns of nearby exon sequences. Thus these results emphasize, mRNA processing and transcription are interlinked.

With this understanding, it is clear that analyzing the splicing mechanism is imperative while discussing RNA polymerase II transcription and translation.

1.6.3 Computational detection of splicing signals

Consensus signals for splice sites were quickly recognized and were used to determine the gene structure. However, it was recognized later that many functional splice sites shared only a few bases of similarity and more sophisticated models were required.

Simple independent weight matrices or frequency tables that yield a probabilistic log-odds score for each base at each position in a sequence were initially developed and they are still used extensively (Staden, 1984). Weight matrices were derived from a training set of true sites to generate the frequency table and then score potential sites by summing the scores of individual bases in a pre-defined window. This was improved with the incorporation of first-order dependencies into the weight matrix framework (Zhang and Marr, 1993).

The next set of improvements came with the components of a successful gene prediction system *GENSCAN* (Burge and Karlin, 1997). It uses a maximal dependence decomposition approach, where the donor sites are broken into a set of classes based on dependencies between bases in the splice site signal and then uses a simple weight matrix to model each class individually (Burge, 1998). For acceptor sites, it uses a windowed weight array method, which models BPS region using a modification of first-order dependencies

approaches that groups sets of neighboring bases together in order to avoid problems caused by limited data.

Later, multiple signals were used to identify splice site regions. *GeneSplicer* (Pertea *et al.*, 2001) combines a traditional log-odds score based on a slight variant of maximal dependence decomposition, a measure of local coding potential and a local optimality requirement. But this approach did not yield improved results.

Another approach was used to identify precise splice sites from among a number of nearby or proximal false positives. This approach used a decision tree to discriminate true and false sites and may prove useful for annotation purposes (Thanaraj, 2000). However, these models produce too many false positives per kb. Typically, if thresholds are set to detect 99% TP, then 12 FP per kb and for thresholds to include 95% TP, 6 FP per kb were reported (Levine, 2001a).

EST sequences have also been used to confirm the site signals on a large scale basis and in analysis of canonical and non-canonical introns (Burset *et al.*, 2000), though their use means the algorithm is no longer truly *ab initio*.

As an attempt to improve previous methods, another program, *Stratasplice* (Levine, 2001a) was developed in which true and false positives were differentiated using the base composition near the splice signals. The local GC content with a first-order dependence weight matrix combination model is used by the predictor to predict the human splice sites. This resulted in better prediction of splice sites of genes in GC-rich sequences.

However, all the programs developed so far, are limited in that they produce excessive false positives when applied on a genome scale. Hence, I attempt to develop a few splice site models that will do fairly on the genomic sequence and will complement the transcription termination predictor in identifying real transcription terminators.

1.7 Transcription and translation

The protein coding mRNA, transcribed by RNA polymerase, is later used for coding for protein synthesis by a process called *translation*. Transcription and translation are coupled

in prokaryotes where there is no defined nucleus or nuclear membranes to separate genetic material from the cytoplasm. However, in eukaryotes it is traditionally believed that these processes occur separately with the transcribed RNA product processed and exported to the cytoplasm where the translation occurs. Also, it is often suggested that the membrane evolved to segregate splicing and translation so that they do not interfere with each other. This understanding was recently challenged with the recent finding of nuclear translation in mammalian cells (Hentze, 2001; Iborra *et al.,* 2001). Three types of evidences supported the possibility of coupled transcription and translation in the eukaryotic cell just like in bacteria. (i) Nuclei contain all the components required for protein synthesis, (ii) Isolated nuclei can incorporate radiolabelled amino acids to make new protein molecule and (iii) Nonsense-mediated Decay (NMD), which is responsible for degradation of transcripts with termination codon near to the 5'-end support the transcription and translation coupling in eukaryotic nucleus (for details see, Hillman *et al.,* 2004; Iborra *et al.,* 2004). NMD, which mostly occurs in the cytoplasm, is also found in the nucleus and this poses a challenge to the current consensus. However, this phenomenon can be explained if some translation occurred within nuclei by the protein machinery present within the nuclei. So, the present model is that ribosomes are assembled within nucleoli and are exported to both nucleoplasm and cytoplasm, where they associate with transcripts and become active. Some nuclear ribosomes are incorporated into the transcription factories and proof-read the newly made transcripts as they emerge from polymerases. Any pre-mature codon in the transcript would trigger the NMD pathway and degrade the transcripts with nearby proteasomes. If no premature stop codons are found, the transcript would be exported to the cytoplasm where it could support multiple translation initiations. Thus, there is evidence that transcription and translation mechanisms are interlinked, so understanding translation signals and modeling them may complement the transcription start site and termination models in predicting the gene structure.

One of the mechanisms by which pre-termination codons are incorporated in the transcript is a frame shift splicing mechanism, and this triggers the NMD pathway (Lewis *et al.,* 2003). This leads to the understanding that identifying translation termination codons will help to legitimate the correct splice sites and screen out the numerous splice site-like signals from the genomic DNA. Thus, translation models may supplement other models in predicting genes in the genomic DNA.

1.7.1 Translation mechanism

Explaining the translation mechanism in detail is beyond the scope of this thesis. So I will give a brief overview of the mechanism in prokaryotes and eukaryotes instead.

1.7.1.1 Translation initiation

The translation initiation mechanism in prokaryotes differs from that in eukaryotes and the process in both is more than a mere assembly of protein components. The initiation phase sets the reading frame which is normally maintained throughout all subsequent steps in the translation process. Moreover, protein synthesis is regulated at the level of initiation, which adds to its importance.

Initiation in prokaryotic polycistronic mRNA is usually selected *via* base pairing with ribosomal RNA. This initiation is regulated by *cis*- and *trans*-acting signals. In eukaryotes, translation initiation sites are reached *via* a scanning mechanism from the AUG codon near to the 5' end of mRNA. However there are also other mechanisms through which initiation can occur. These are context dependent leaky scanning, reinitiation and internal initiation where translation initiation is directed from an AUG that is not the nearest to the 5' end (for details refer, Gray and Wickens, 1998; Kozak, 1999, 2001; Kozak, 2002; Pain, 1996; Sonenberg and Dever, 2003).

At the start codon, the 30S ribosomal subunit forms an initiation complex with a special form of tRNA (fMet-tRNA) and a GTP-binding protein IF2. IF1 and IF3 stabilize the binding of fMet-tRNA·IF2·30S complex and thus initiate polypeptide chain formation with addition of methionine. AUG is the common initiator codon because it forms a stable interaction with CAU anticodon in fMet-tRNA. GUG and UUG are also used as start codons in >10% of bacterial genes. AUU codon is used in a single *Escherichia coli* gene. The initiation phase is completed with the 50S ribosomal subunit forming a 70S unit with fMet-tRNA occupying the P-site of the ribosome.

Start codons in prokaryotic mRNA are distinguished by an upstream purine-rich sequence that pairs with a complementary sequence in the 16S rRNA component of the small

ribosomal subunit. This sequence, called the *Shine-Dalgarno* (SD) sequence, consists of three to nine contiguous bases in the mRNA that form standard base pairs (not including G·U) with bases from 1534 to 1542 (ACCUCCUUA) at the 3'-end of 16S rRNA. This SD interaction augments initiation by anchoring the 30S subunit in the vicinity of the start codon. Apart from the SD signal present nearby the start codon, several trans-acting signals and factors have been reported. However, the SD sequence is not essential in all initiations as some AUG codons are found to initiate without SD augmentation. Similar cases were reported for chloroplast mRNAs as well. In these cases, the SD sequence is generally considered to be substituted by a low GC content (hence minimal secondary structure) in the 5' UTR region (for review see, Kozak, 1999).

Efficient formation of initiation complexes requires the sequence immediately preceding the SD element to be devoid of any secondary structure. Some additional sequence elements present downstream of the AUG codon might substitute for the main SD element. These elements have patchy complementarity to 16S rRNA and include weak G·U pairings and so their significance remains inconclusive. Many prokaryotic mRNAs are polycistronic and ribosomes translating the first open reading frame will often, upon termination, slide a few bases upstream or downstream to reinitiate at the next start codon.

The eukaryotic mechanism differs with the 40S ribosomal subunit entering near the 5' end and sliding its way to identify the first AUG codon, which is recognized by base pairing with the anti-codon in Met-$tRNA_i$. AUG is the most common initiating codon; however, ACG and CUG codons are also used. Methionine is the first amino acid even when the first codon is other than AUG. Eukaryotic initiation depends on the m7G cap added to the 5' end of mRNA molecule. In vertebrate mRNAs, the initiation sites has a consensus sequence of GCC**R**CC**AUGG** with R (purine, mainly A) and G at -3 and +4 positions showing more active role (Iida and Kanagu, 2000; Kozak, 1987). Poly(A) tail and 3' UTR might also influence translation initiation (Figure 9).

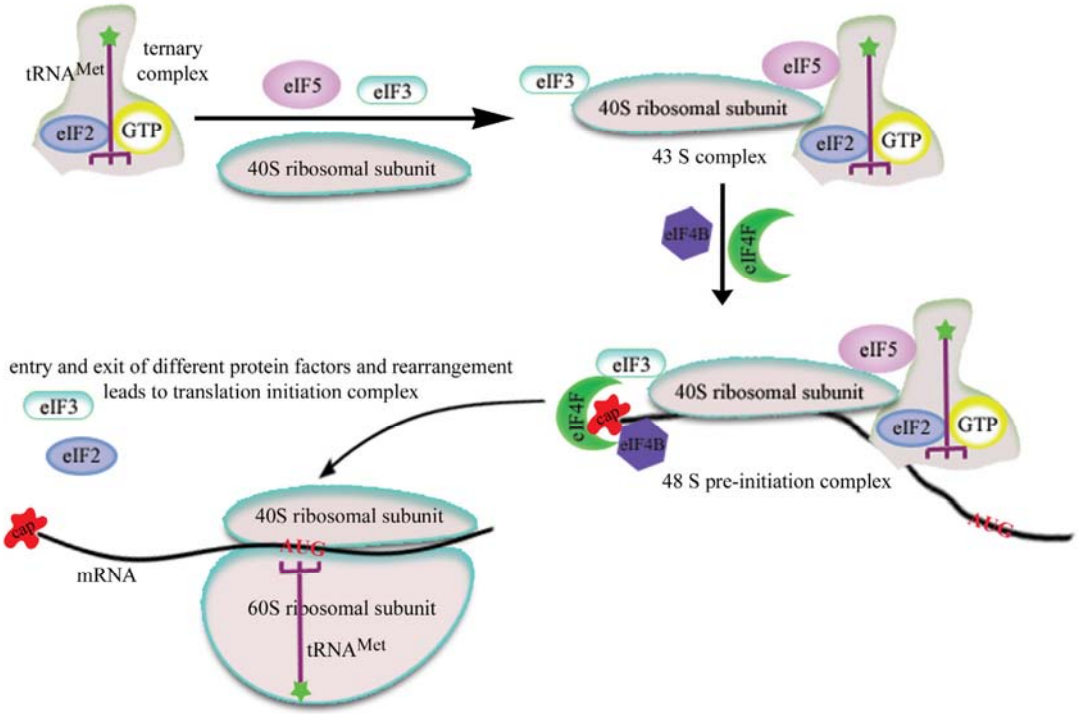

Figure 9. Translation initiation in eukaryotes

Leaky scanning allows 40S ribosomal subunits to by-pass the first AUG codon and initiate instead at the second or rarely at the third AUG codon. This is mainly due to sub-optimal context near to the first AUG codon. There is some evidence that initiation can occur with non-AUG codons as well. Re-initiation in eukaryotes occurs if the initiation complex gets terminated at some distance near to the 5' end. Scanning then continues until the next authentic AUG is reached. IRES (Internal Ribosome Entry Site) is another mechanism, wherein translation of mRNA occurs from an internal initiation site (Houdebine and Attal, 1999).

1.7.1.2 Translation termination

Translation termination is due to stop codons in the mRNA sequence. When a stop codon has been translocated into the ribosomal A-site by the action of elongation factor EF-G or eEF2, a cleavage of the ester bond between the peptide and tRNA moieties of the peptidyl-tRNA complex occurs at the peptidyl transferase centre of the ribosome. In prokaryotes, termination involves two different release factors recognizing UAA/UAG and UAA/UGA respectively, whereas in eukaryotes all the three stop codons are recognized by a single release factor. Eukaryotic release factor binding to the ribosomal A site is GTP dependent

and RF3·GTP binds at this site when it is occupied by a termination codon. Then, hydrolysis of the peptidyl-tRNA ester bond, hydrolysis of GTP, release of nascent polypeptide and deacylated tRNA and ribosome dissociation from mRNA ensue (Kisselev and Frolova, 1995) (Figure 10).

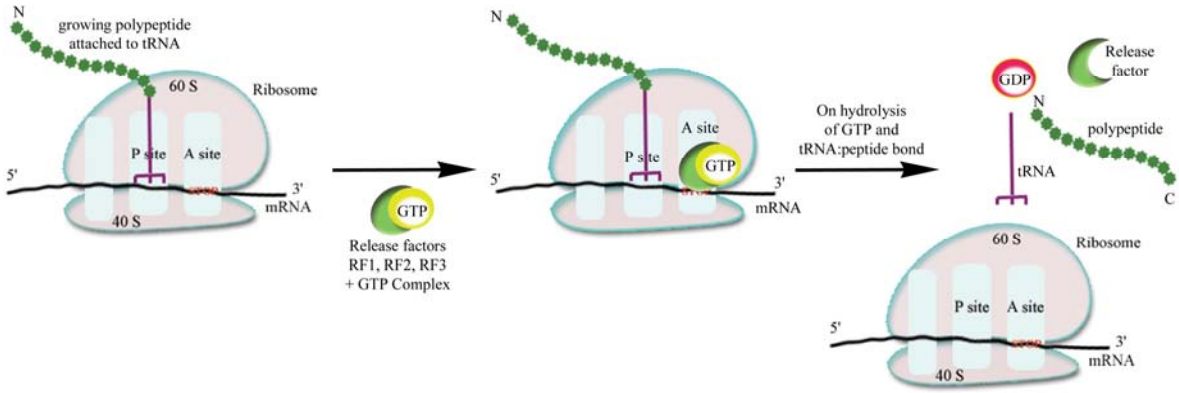

Figure 10. Translation termination mechanism mediated by release factors

Translation termination efficiency was found to be improved by the local context in yeast genes. The consensus sequence, CA(A/G)N(U/C/G)A, located downstream of the stop codon base pairs with the regions close to helix 18 and 44 of the 18S rRNA for augmenting translation termination efficiency (Namy *et al.*, 2001). In higher eukaryotes, the stop codons are biased towards purines (Cavener and Ray, 1991). Also, the CpG dinucleotide patterns present immediately downstream of the stop codons are significantly suppressed (Cavener and Ray, 1991).

The downstream context also plays a detrimental role for the UGA triplet in deciding whether it is used as a termination codon or selenocysteine codon.

Analysis of full length RIKEN mouse cDNA and eukaryotic UniGene clusters (Ozawa *et al.*, 2002) showed the following results –

(i) The occurrence of guanine at position +1 (immediately after the stop codon) was high in mammals. Adenine was high at this position in plants and Zebrafish.
(ii) The occurrence of cytosine at position +1 was low in plants.
(iii) The occurrence of cytosine at position +4 was high in mammals.

(iv) The occurrence of cytosine at position +2 was high in plants. In human positions +2, +3, +4, +7 and +13 after stop codons have some information content.

Apart from DNA signals, protein factors also influence translational efficiency. PABP1 interacts with initiation factors eIF4G and eIF4B and promotes the synergistic effect of having both a cap and poly(A) tail on translation efficiency. The translation termination factor eRF3 also interacts with PABP1 and so could relay information from the termination complex to both ends of mRNA and thus regulate subsequent translation initiation (Cosson *et al.*, 2002).

1.7.2 Computational detection of translation signals

Identifying translation start sites depends on the consensus signals identified near to the initiator codon. Several attempts have been made to correctly identify the translation start site and to screen true sites from the false sites in the genomic DNA.

In 1987, Kozak developed the first weight matrix from an extended collection of vertebrate mRNA data (Kozak, 1987). The consensus motif derived from the matrix is GGG**A**CC**ATGG**, where a single G nucleotide following the ATG codon and three A nucleotides upstream are two highly conserved positions.

Later prediction methods took the nucleotide context in the vicinity of the start site as well. These include the positional conditional probability matrix (Salzberg, 1997) and generalized second-order profile models (Agarwal and Bafna, 1998). In the Agarwal and Bafna model, an algorithmic idea of the ribosome scanning model was implemented. The search starts from the 5’ end of the mRNA and an AUG is defined as a putative start codon if followed by an ORF longer than 200 nucleotides. Likewise, in the Pederson and Nielson *NetStart* model, an Artificial Neural Network (ANN) was constructed with 100 bases upstream and downstream of AUG codon that recognizes the surrounding context (Pedersen and Nielsen, 1997b). These approaches are significantly better than weight matrix models but still generate high false positive rates.

So, to improve the prediction accuracy, Salamov *et al.,* developed a program called *ATGpr*, where the following six characteristics are applied to analyze the sequence around putative start sites:

(a) Positional weight matrix around an ATG.
(b) Hexanucleotide difference between upstream and downstream of ATG sequences.
(c) Preference for longer reading frames downstream of ATG.
(d) Signal peptide characteristic.
(e) Presence of another upstream in-frame ATG.
(f) Upstream cytosine nucleotide characteristic.

Linear discriminate analysis was used to finalize the score from these properties. The important components in the ATGpr model are the positional triplet weight matrix around AUG and the hexanucleotide difference between the upstream and downstream of the AUG in a 50 nucleotide long window (Salamov *et al.,* 1998a). Along with these properties, another program developed by Zhang *et al.* used 50 base pair downstream windows to screen for in-frame stop codons and local context to determine translation start site (Zhang *et al.,* 2000).

Recently, a method based on Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000) has been introduced by Zien *et al* (Zien *et al.,* 2000). To add to this, Liu *et al* used SVM as classifiers with possible amino acid patterns around start sites to differentiate true and false sites (Liu *et al.,* 2003). Similar to this, an ANN with the ability to determine coding/non-coding potential around the start codon and conversed motif was also developed (Hatzigeorgiou, 2002).

Contrary to these various translation start models, not much computational analysis has been carried out on translation stop prediction as identifying them becomes relatively easy if the correct translation start site and ORF can be determined.

1.8 Objectives of this project

With this understanding, it is clear that for a gene prediction program that works purely based on gene regulatory signals, it is necessary to have efficient methods to capture the complexity of regulatory signals linked to each process from the genomic sequences.

As a transcription start site predictor is available (Down and Hubbard, 2002), modeling transcription termination is the next important step as the task has proved challenging for nearly 25 years now. Extensive research on transcription termination through these years has still not cleared the enigma and a clear mechanism of the process is yet to be realized. So, the major aim of this project is to build a transcription termination model using the genomic sequences available with the different techniques explained in chapter 2. A successful predictor will be useful to identify the point where RNA polymerase II stops transcription and exits from the DNA sequence, and thus helping to sketch the gene structure. This is explained in chapter 3 along with some interesting results found by the model.

As explained previously, transcription is tightly linked with the splicing and translation process and thus identifying their regulatory signals may help to supplement the development of a transcription termination model. So I have set the objective of modeling splice site and translation start and stop signals as well. Chapters 4 and 5 detail the models trained to meet these objectives based on the learning techniques explained in chapter 2.

Finally, in chapter 6, I meet the objective of creating an *ab initio* gene prediction system based on DNA regulatory signals by linking the predictions of the models using GAZE (Howe *et al.,* 2002).

Apart from this goal; I worked on two other project areas as well. These are explained in the Appendices. Appendix A gives an overall view of the project with the aim of identifying domain insertions in known protein structures. Appendix B details the analysis of protein evolution based on sequence and structure conservation.