Chapter 2. Biogenesis and nuclear export of cellular mRNA transcripts

2.1) Biogenesis of eukaryotic mRNA transcripts

The transcription of protein-encoding genes gives rise to pre-messenger RNAs (pre-mRNAs). In eukaryotes, transcription of mRNAs is carried out by the RNA Polymerase II (Pol II), which is also responsible for the production of UsnRNPs and 5S rRNAs (reviewed by Maniatis and Reed 2002; Orphanides and Reinberg 2002). Pre-mRNAs are bound to multiple proteins, as pre-mRNA ribonucleoprotein complexes (pre-mRNPs) (Dreyfuss *et al.*, 2002). The recruitment of proteins to the nascent transcript starts co-transcriptionally and continues during all the steps of pre-mRNA maturation (reviewed by Maniatis and Reed 2002; Neugebauer 2002; Orphanides and Reinberg 2002). These proteins are diverse and have RNA-binding properties, RNA chaperone and/or RNA helicase activity (Dreyfuss *et al.*, 2002; Linder and Stutz 2001). The assemblage of pre-mRNAs into pre-mRNPs confers to the transcript stability towards enzymatic RNA degradation, and information about its maturation status. The successful maturation of a pre-mRNA gives rise to a competent transcript (mRNA), which is assembled into a final mRNA ribonucleoprotein complex (mRNP). The mRNP is like the pre-mRNP, protected from enzymatic degradation, but also endowed with nuclear export and cytoplasmic translation capabilities (Dreyfuss *et al.*, 2002).

Typically, the maturation process of a pre-mRNA transcript includes several inter-related steps, which are: 5’-capping; splicing of intron-containing mRNAs; 3’-end polyadenylation and nucleocytoplasmic export of the mature mRNP. In the cytoplasm, the mRNP is recruited to the polysomes, where ribosome-mediated protein synthesis takes place (figure 7) (reviewed by Maniatis and Reed 2002; Orphanides and Reinberg 2002; Erkmann and Kutay 2004). Also, several quality control mechanisms exist, both in the nucleus and the cytoplasm, for mRNA turnover and the degradation of aberrant mRNAs, generally referred to as mRNA surveillance mechanisms, such as nonsense-mediated decay (NMD) and the exosome machinery (reviewed by Mitchell and Tollervey 2001; Baker and Parker 2004).

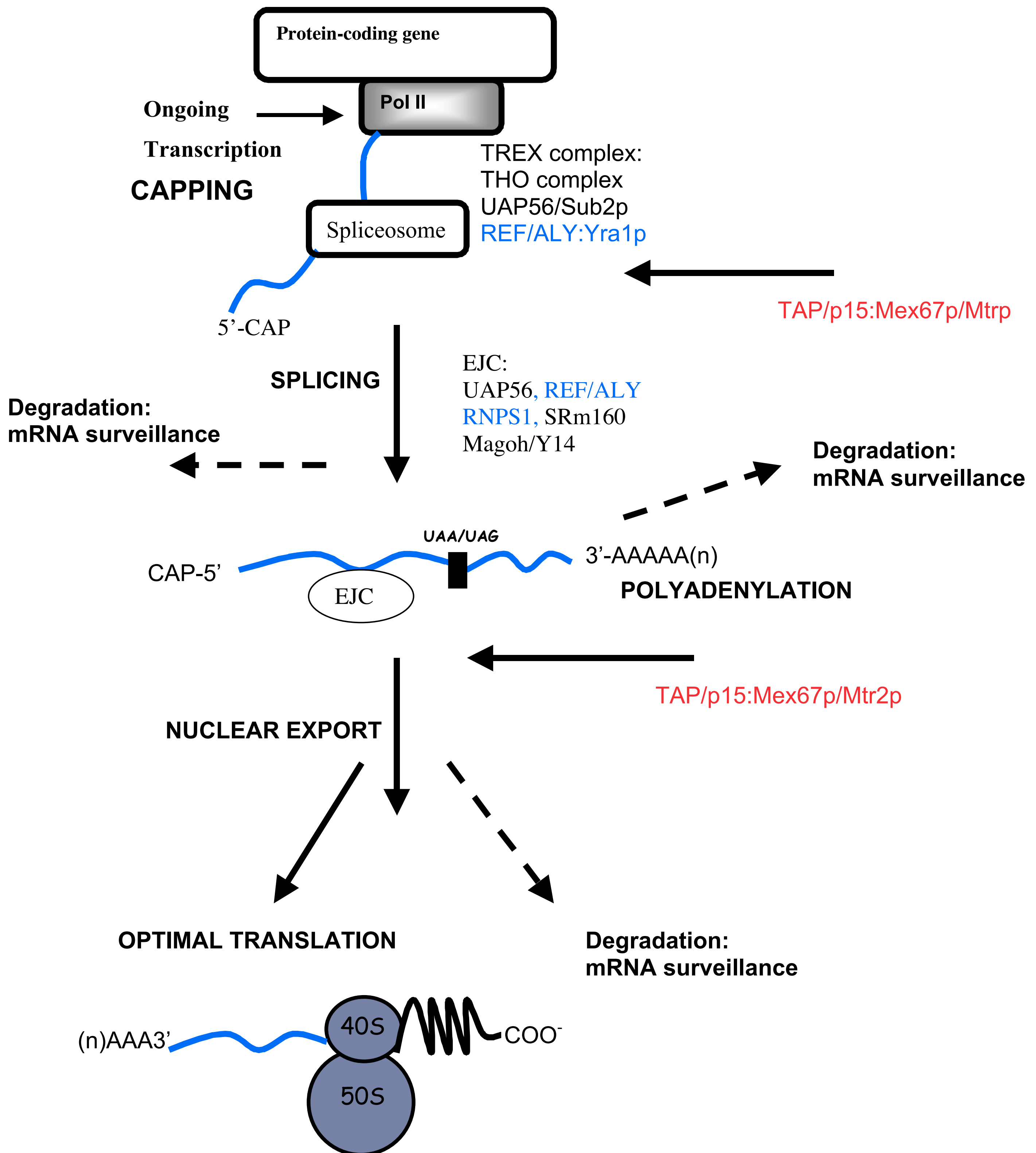


Figure7. The biogenesis and nuclear export of cellular mRNA transcripts. Highlighted in red, are the sites where TAP/Mex67p is apparently recruited to the mRNP. In blue, are the nuclear proteins that are thought to recruit TAP/Mex67p to the exportable mRNPs. Dashed arrows indicate the optional degradation of mRNP that have not been properly packed or contain premature termination codons.

1) 5’-Capping

Soon after transcription initiation, the 5’-end of a nascent pre-mRNA is capped (the 5’ triphosphate is replaced by a methylated guanosine monophosphate, m7GpppN). The process of capping confers stability to the nascent mRNA towards 3’- to 5’-endonucleases. It also leads to the recruitment of the cap binding protein (CBP) in the nucleus. In the cytoplasm, the CPB is exchanged by the elongation factor eIF4E, which participates in mRNA translation (reviewed by Neugebauer 2002; Proudfoot and Furger 2002).

2) mRNA splicing

The process of splicing is carried out in the nucleus by the spliceosome, which is a complex protein/RNA assembly that promotes the enzymatic removal of non-coding mRNA sequences (introns) and the ligation of corresponding exons (reviewed by Neugebauer 2002; Proudfoot and Furger 2002). In humans, the splicing reaction leads to the deposition of a complex of proteins referred to as the exon-junction complex (EJC), onto the mRNP. The EJC is thought to enhance the cytoplasmic translation rates by enhancing mRNA stability and recruiting the mRNP to the polysomes in the cytoplasm (Wiegand *et al*., 2003; Nott *et al*., 2004).

3) 3’-end polyadenylation

Upon transcription temination, the nascent transcript is cleaved at its 3’-end, released from the site of transcription and stabilized by the addition of a poly-adenine (A) tail. Importantly, mRNA transcription termination occurs only if a functional polyadenylation signal is present. Likewise, 3’-end cleavage and polyadenylation are triggered by transcription termination (reviewed by Maniatis and Reed 2002; Neugebauer 2002; Proudfoot and Furger 2002).

4) nuclear 3’- to 5’- degradation of aberrant transcripts: the exosome

The exosome is an assembly of 3’- to 5’-exonucleases. Its function is to degrade mRNAs that could not reach the cytoplasm, apparently due to the presence of aberrant

3’-ends and/or unproper assemblage of the mRNP. It is also involved in the maturation process of structured RNAs (e.g. rRNA, snRNA, snoRNA) (reviewed by Mitchell and Tollervey 2001).

5) nuclear export of mature mRNPs

The bulk of mature mRNPs are exported to the cytoplasm, by the heterodimer Mex67p/Mtr2p in yeast and its metazoan ortologue TAP/p15 in higher eukaryotes. TAP/p15:Mex67p/Mtr2p is recruited to intron-containing and intron-less mRNPs by various, apparently redundant protein adaptors. The best characterized of these protein adaptors is REF/Aly:Yra1p, which is part of the transcription-export (TREX) complex in yeast and the EJC in metazoa (reviewed by Stutz and Izaurralde 2003; Vinciguerra and Stutz 2004).

2.2) TAP/NXF1: the export factor receptor of most cellular mRNAs

The export of mature mRNA transcripts is primarily driven by transport factors different from karyopherins, which belong to a family of proteins referred to as nuclear export factors (NXF). NXFs are evolutionary conserved in eukaryotes, and have more than one member per species in metazoans, but only one in yeast (appendix A) (Herold *et al.*, 2000). These proteins share a similar domain organization, but only Mex67p in yeast and NXF1 (also called TAP) in metazoa have been shown to participate in the export of cellular messenger RNA (reviewed by Stutz and Izaurralde 2003; Vinciguerra and Stutz 2004).

The human protein TAP (Tip-associated protein) was initially identified in a yeast two hybrid screen, as a cellular factor that interacts with the viral protein Tip (Tyrosine kinase-interacting protein) from herpes virus saimiri (Yoon *et al.*, 1997). Later, Izaurralde and co-workers provided evidence that hTAP is a nuclear export factor, recruited by the genomic RNA of certain simple retroviruses through a cis-acting RNA structural element, the constitutive transport element CTE (Grüter *et al.*, 1998). In agreement with this, Hurt and co-workers had previously shown that in yeast, the protein Mex67p (mRNA export factor) is essential for nuclear poly(A)-RNA export (Segref *et al.*, 1997). As expected for a nuclear export factor, human

TAP accumulates at the nuclear rim in the nucleoplasm, and is capable of bidirectional shuttling across the nuclear envelope (Bear *et al*., 1999). An essential role for TAP/NXF1 in mRNA nuclear export has also been demonstrated in *C. elegans* (Tan *et al*., 2000) and *D. melanogaster* (Herold *et al*., 2001).

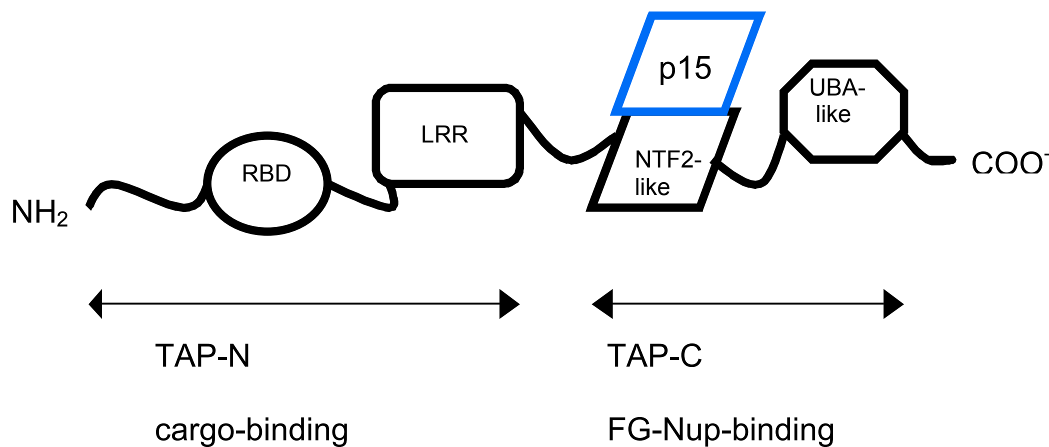
The protein TAP/NXF1 is only functional when in complex with p15 (Guzik *et al*., 2001; Levesque *et al*., 2001; Wiegand *et al*., 2002). In yeast, the protein Mtr2p is a functional analogue of p15. Although Mtr2p shares no sequence similarity with p15, it displays significant structural similarity (Fribourg and Conti 2003) and has a similar functional role (Katahira *et al*., 1999). Even if both hTAP and Mex67p have general RNA binding activity (Braun *et al*., 1999; Kang and Cullen 1999; Liker *et al*., 2000), NXF1/p15:Mex67p/Mtr2p heterodimers contact the mature mRNA through protein adaptors, such as the protein REF/Aly:Yra1p (Rodrigues *et al*., 2001). The directionality of TAP-mediated nuclear transport is not yet understood, as TAP/Mex67p does not interact with Ran, or any other Ras-like GTPase (Clouse *et al*., 2001).

2.2.1) Structure of hTAP/NXF1

TAP/NXF1 is a modular protein. In humans, the polypeptide is composed of four structured domains that are interconnected by flexible peptide linkers (Suyama *et al*., 2000). According to its binding properties, hTAP can be dissected into two functional parts (figure 8). The N-terminal part of TAP (TAP-N) is involved in general RNA-binding (Braun *et al*., 1999; Kang and Cullen 1999; Liker *et al*., 2000), protein-adaptor recognition (Stutz *et al*., 2000) and specific interaction with the CTE viral RNA (Braun *et al*., 1999; Kang *et al*., 1999). The C-terminal part (TAP-C), on the other hand, heterodimerizes with p15 (Guzik *et al*., 2001; Levesque *et al*., 2001; Wiegand *et al*., 2002) and features two domains, that are responsible for the interaction with nucleoporin FG-repeats (Bachi *et al*., 2000; Braun *et al*., 2002). Starting from the N-terminus, TAP-N consists of an unstructured N-terminus followed by a non-canonical RNP-type RNA-binding domain (RBD), and a leucine rich repeat (LRR) domain (Liker *et al*., 2000). The remaining C-terminal part, TAP-C, comprises an NTF2-like domain that heterodimerizes with p15 (Fribourg *et al*., 2001), followed by an UBA-like domain (Grant *et al*., 2002). Both the NTF2-like and the

UBA-like domains contain an FG-repeat binding-pocket (Fribourg *et al.*, 2001; Grant *et al.*, 2003). These two C-terminal domains are involved in NPC translocation and are relatively interchangeable, as a TAP construct lacking the NTF2-like domain, but having two UBA-like domains, is functionally competent, albeit not to the extent of the wild-type protein (Braun *et al.*, 2002). p15 is also an NTF2-like domain, whose function seems to be structural, maintaining the NTF2-like domain of TAP in a FG-repeat binding conformation (Fribourg *et al.*, 2001, Fribourg and Conti 2003). Independent structures of TAP-N (Liker *et al.*, 2000), the NFT2-like heterodimer (Fribourg *et al.*, 2001, Fribourg and Conti 2003) and the UBA-like domain (Grant *et al.*, 2002; Grant *et al.*, 2003) are now available (figure 9).

a) human TAP/NXF1:p15/NXT1 heterodimer



b) yeast Mex67p/Mtr2p heterodimer

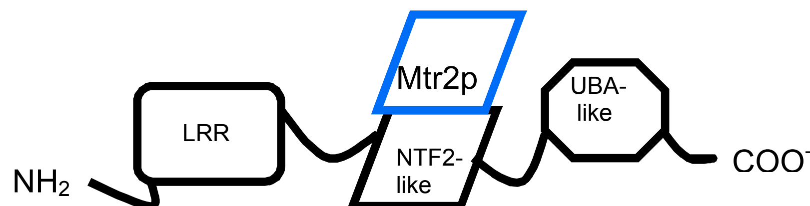


Figure 8. Domain organization of TAP/NXF1 and Mex67p/Mtr2p. Mex67p is, according to sequence analysis, predicted not have an RBD domain.

2.3) The REF family of protein adaptors

REF proteins were first described in yeast in 1997 (Portman *et al.*, 1997) and since then have been attributed several functional roles. A Yeast homologue of metazoan REF proteins was initially described as a heteronuclear ribonucleoprotein (hnRNP)-like protein with RNA-annealing activity and named Yeast RNA annealing protein 1 (Yra1p) (Portman *et al.*, 1997). A mammalian homologue was cloned later and termed Aly, for Alloy of LEF-1 and AML-1 (Bruhn *et al.*, 1997). Aly was described as a nuclear protein that enhanced the DNA-binding activity of the transcriptional activators LEF-1 and AML-1 (Bruhn *et al.*, 1997). Two years later, Green and colleagues found that the same protein enhanced the DNA-binding properties of transcription factors that contain a basic region leucine zipper (bZIP). The acronym BEF (bZIP enhancing factor) was used this time (Vibrasius *et al.*, 1999). Finally, Yra1p was found to interact *in vitro* with Mex67p, which is the major mRNA nuclear export factor in Yeast (Strässer and Hurt 2000; Stutz *et al.*, 2000; Zenklusen *et al.*, 2001). Further, experiments in oocytes showed that Aly interacts with the human protein TAP/NXF1. This led to the proposal that Yra1p in yeast and its metazoan homologue Aly, are members of an evolutionary conserved family of proteins. Given their role in the nuclear export of cellular mRNA, these proteins are referred to as the REF family of proteins (appendix B). Where the acronym REF stands for RNA and export factor binding protein (Stutz *et al.*, 2000).

REF proteins have a conserved central RNP-type RBD domain, flanked by arginine-glycine-glycine (RGG)-rich regions of variable length and short highly conserved N- and C- termini, referred to as the N- and C-terminal boxes respectively (figure 10) (Rodrigues *et al.*, 2001; Stutz *et al.*, 2000). Biochemical studies have shown that REF proteins interact with TAP/Mex67p and RNA through their N- and C-terminal boxes and RGG variable regions, but apparently not through the RBD domain (figure 10) (Rodrigues *et al.*, 2001). The role of REF proteins in mRNA export is based on the assumption that REF/Yra1p recruits the nuclear transport factor TAP/Mex67p to the mature mRNP (reviewed by Stutz and Izaurralde 2003 and Vinciguerra and Stutz 2004).

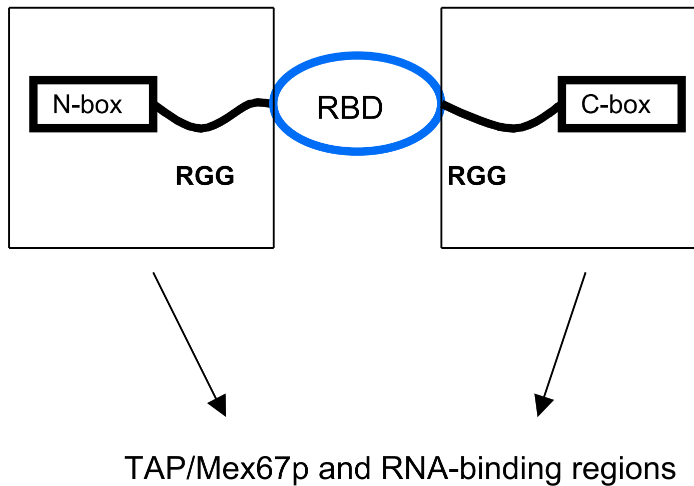


Figure 10. Schematic cartoon of the structural motifs endowed by the REF proteins. blue: the RNP-type RBD, black: variable RGG-motifs and the conserved N- and C- terminal boxes.

REF/Aly:Yra1p is believed to be recruited to mRNA transcripts early during transcription, presumably by the splicing factor/nuclear export factor UAP56 in metazoa (Luo *et al.*, 2001) and Sub2p in yeast (Strässer and Hurt 2000). In metazoa, REF and at least 5 other proteins (e.g. Magoh, RNPS1, SRm160, UAP56 and Y14) are deposited onto spliced mRNAs, constituting the EJC, which has been defined above. This multiprotein complex is deposited 20-24 nucleotides upstream of exon-exon junctions, in a sequence- independent manner (Le Hir *et al.*, 2000a; Le Hir *et al.*, 2000b). The EJC is thought to confer a post-splicing molecular tag to the cellular mRNA transcripts, giving (as mentioned before) stability, translational competence to the mature mRNP, and at least in some cases (e.g. *oscar* mRNA), provides information about the mRNP's cytoplasmic localization (reviewed by Stutz and Izaurralde 2003; Tange *et al.*, 2004).

Although the presence of REF in the EJC and its ability to interact with the export factor TAP initially suggested a role for the EJC in facilitating the export of spliced mRNAs, more recently it has been shown that the EJC does not play a significant role in nuclear export (Wiegand *et al.*, 2003; Nott *et al.*, 2004). Also, the recruitment of REF proteins to the mRNP is not necessarily splicing-dependent, since REF is found both on spliced and on unspliced mRNAs (Hieronymus *et al.*, 2003). Moreover, both Sub2p and Yra1p are essential for the nuclear export of mRNA transcripts in yeast (Stutz *et al.*, 2000), whereas in *D.melanogaster*, REF/Aly is dispensable for mRNA

export, but UAP56 is essential (Gatfield and Izaurralde 2002). It therefore appears that in metazoa, the essential role of UAP56 is not restricted to REF/Aly recruitment. Further, the non-essential role of REF in metazoan mRNA export, is possibly due to the presence of other protein adapters, which might also mediate the recruitment of TAP/NXF1 to the mature mRNP complex (reviewed by Stutz and Izaurralde 2003; Vinciguerra and Stutz 2004).

Recently, the SR proteins RNPS1 (Huang *et al.*, 2003), 9G8 (Huang *et al.*, 2004) and ASF/SF2 (Huang *et al.*, 2004) have been shown to promote the recruitment of TAP/NXF1 to metazoan mRNPs. Interestingly, 9G8 and ASF/SF2 do so, only when hypophosphorylated (Huang *et al.*, 2004; Lai and Tarn 2004). Furthermore, the yeast protein Np13p is thought to regulate the recruitment of Mex67p to the mature mRNP. The recognition of Mex67p by Nl3p is as well phosphorylation-dependent (Gilbert and Güthrie 2004).

2.3.1) The structure of REF/Aly

The solution structure of the RBD domain of REF/Aly has recently been determined by NMR spectroscopy (figure 11) (Perez-Alvarado *et al.*, 2003). The function of the REF RBD is however, still unclear, since it is conserved among the REF family of proteins, but does not participate in RNA or TAP-binding (Stutz *et al.*, 2000).

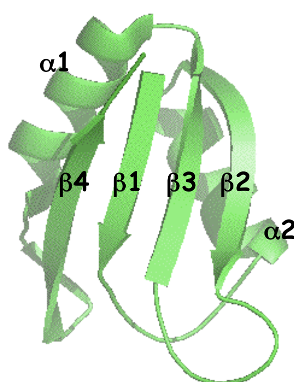


Figure 11. Ribbon cartoon of the NMR structure of REF/Aly, Perez-Alvarado *et al.*, 2003.

Chapter 3. Biogenesis and nuclear export of retroviral mRNA transcripts

3.1) Retroviruses

Retroviruses resemble other RNA viruses in several aspects, except that they replicate their genomes through a DNA intermediate. The extra-cellular virus particle is composed of a single-stranded RNA genome that is wrapped in a core of viral proteins, which in turn are surrounded by an envelope consisting of viral glycoproteins. Although viral multiplication occurs only within a host cell and depends on cellular functions, an infecting retrovirus also brings along an organized collection of viral enzymes and non-coding RNA elements designed to direct the synthesis of a double-stranded DNA copy from its RNA genome (reverse transcription) and the precise joining of that DNA to the host chromosome (integration) (reviewed by Varmus 1988; Wödrich and Krausslich, 2001).

Generally, retroviruses contain two copies of a single-stranded positive-sense RNA genome of approximately 10 Kilobases. The RNA genome is reverse-transcribed into a linear double-stranded DNA by the viral enzyme reverse transcriptase. Subsequently, the viral DNA is inserted into the host genome (by a viral integrase enzyme), where it is retained during the life cycle of the infected cell and all daughter cells. The integrated retroviral genome is termed a provirus and is generally transcribed into a single primary transcript by cellular RNA polymerase II complexes (Pol II). All retroviruses have a similar genomic organization, consisting of several genes flanked by transcriptional control elements. Simple retroviruses feature open reading frames encoding for a minimum of three proteins. Gag (structural), Pol (replication enzyme) and Env (glycoprotein). Complex retroviruses on the other hand, need to modulate their gene expression patterns and so, harbour additional open reading frames that code for regulatory proteins. Transcriptional control elements, flanking the structural and regulatory genes, are duplications of terminal sequences, referred to as long terminal repeats (LTR). The 5’ LTR is a Pol II promoter whereas the 3’ LTR is a poly-adenylation signal (reviewed by Varmus 1988; Wödrich and Krausslich, 2001).

The successful completion of retroviral infection relies on the production of different translation products at different and defined ratios, starting from only one or two-coding transcripts. As a consequence, both unspliced and spliced viral mRNAs need to access the cytoplasm. There are currently several examples of how retroviruses interact with the cellular machinery to pursue with their replication cycle. One refers to the HIV-1 retrovirus, which is a *bonafide* complex retrovirus. Another concerns certain types of simple retroviruses, such as type D retroviruses, (reviewed by Cullen 1998; Cullen 2000).

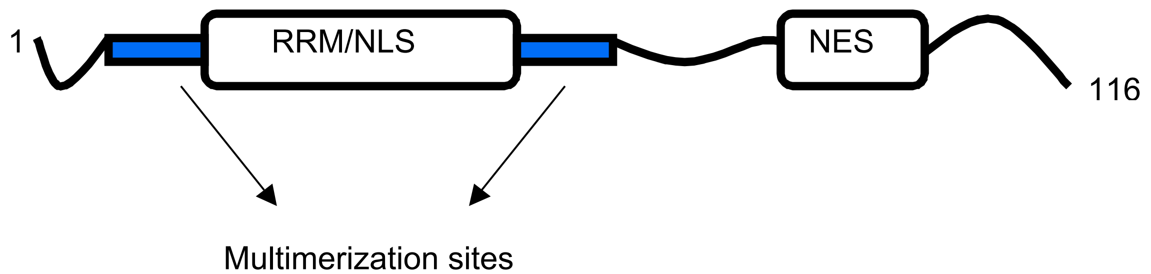
3.2) Complex retroviruses: HIV-1 and the karyopherin-mediated export

Complex retroviruses are endowed with a temporal regulation of their gene expression strategy. HIV-1, a prototypical complex retrovirus, regulates its gene expression pattern at the post-transcriptional level, mainly through the expression of the transacting viral protein Rev. Rev is required for the expression of the incompletely spliced mRNAs, which mainly encode regulatory proteins, but not fully-spliced mRNAs, which encode, on their majority for structural proteins. This means that the availability of regulatory proteins, during viral infection, relies over time, on the expression levels of Rev, which is expressed from a fully-spliced mRNA. All HIV-1-derived incompletely spliced mRNA transcripts, feature a cis-acting RNA motif referred to as the Rev-responsive element (RRE). Rev interacts specifically with and multimerizes onto the RRE RNA moiety of viral mRNA transcripts. In doing so, Rev recruits RRE-containing mRNA transcripts to Crm1, which is the protein responsible for the nuclear export of many classes of cellular proteins, U snRNAs, and 5S rRNAs (reviewed by Cullen 1998; Cullen 2000; Wödrich and Krauslich 2001).

3.3) The structure of Rev

The protein Rev is 116 amino acids long, shuttles between the nucleus and the cytoplasm by means of a NLS and NES, respectively, and features an arginine (R) rich motif (RRM) that recognizes the RRE specifically. Rev's domain organization comprises the RRM positioned at the N-terminal half of the polypeptide, which serves both as an RNA-binding platform and as a NLS (figure 12). This RRM is flanked on

both extremes by short multimerization sites. The NES is a leucine rich sequence of amino acids located to the C-terminal half of the protein (reviewed by Pollard 1998). An NMR-based atomic structure is available of the RRM of Rev in complex with the RRE (Battiste *et al*., 1996). Basically, the RRM of Rev penetrates the major groove of the RRE, on a region where the A-helical conformation is slightly distorted (figure 12).



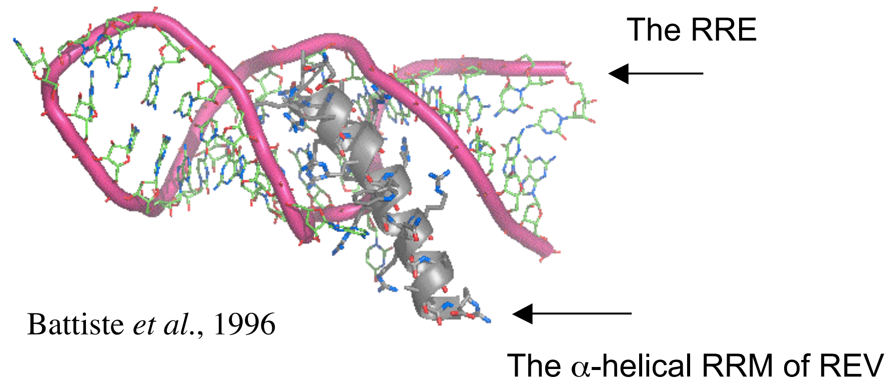


Figure 12. Domain organization of Rev and ribbon cartoon of the RRE (magenta), bound to the α-helix of Rev (grey).

3.4) Simple retroviruses: Simian type D retroviruses and TAP-mediated export

Type-D simian retroviruses, as all retroviruses, need to export unspliced genomic RNA for the package of their progeny virions and the synthesis of structural proteins (reviewed by Cullen 1998; Cullen 2000; Wödrich and Krausslich, 2001). The genomic RNA map of these retroviruses features a cis-acting RNA element, upstream of the 3' LTR, which represents its main device for proper replication and survival. This non-coding stem-loop RNA is known as the constitutive transport element (CTE), which interacts with cellular key factors of the cellular mRNA export pathway in metazoans (Ernst *et al.*, 1997a; Saavedra et al., 1997; Pasquinelli *et al.*, 1997), in particular with TAP/NXF1 (Tang *et al.,* 1997; Gruter *et al.*, 1998). The result of this interaction is the export of the viral genomic RNA (or any CTE-containing RNA), by-passing several steps of the canonical cellular mRNA metabolic pathway (reviewed by Cullen 1998; Cullen 2000; Wödrich and Krausslich, 2001). The CTE RNA is a 173 nucleotides long, extended RNA stem loop, with two conserved internal loops exhibiting 180° inverse symmetry (Ernst *et al.*, 1997b). See figure 13.

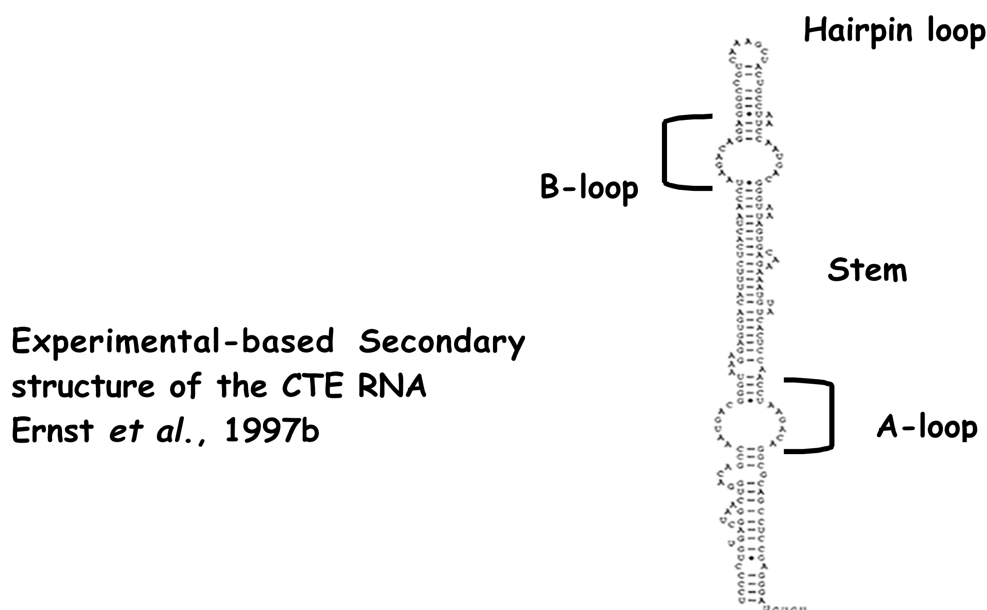


Figure 13. The experimental-based secondary structure of the CTE RNA

Chapter 4. The questions addressed in this PhD thesis

1) Main project:

Structural characterization of a protein/RNA complex: human protein TAP/ retroviral CTE RNA

The structure of TAP-N (in particular, TAP102-372) is now available (Liker *et al.*, 2000), which has structural and biochemical similarities with the U2B”-U2A’/U2snRNA spliceosomal complex structure (Price *et al.*, 1998). In both cases, an RNP-type RBD domain and a LRR domain come together to bind a specific RNA stem loop (figure 14). Although this finding suggests that TAP-N might interact with the CTE RNA in a similar mode, mutagenesis studies have indicated that this is unclear (Coburn *et al.*, 2001; Liker *et al.*, 2000). In the case of the spliceosomal complex, the α-helical face of the RBD motif sits on the concave β-surface of the LRR domain. Disruption of this domain-domain interaction abolishes U2snRNA binding (Price *et al.*, 1998). Mutagenesis studies show that this is not the case for TAP-N, though (figure 14). In addition, the RBD and LRR domains of TAP-N need to be in cis when binding to the CTE RNA. If mixed in trans, CTE RNA binding does not occur anymore (Liker *et al.*, 2000). These results suggest that TAP-N recognises the CTE RNA in a different mode and gives rise to the main question of this project:

How do the RBD and LRR domains of TAP-N come together to interact with the viral CTE RNA?

2) Side project

X-ray crystal structure and ligand-binding surface of REF1-II, a splice variant of the protein REF/Aly

The solution structure of the RBD domain of REF/Aly has recently been determined by NMR spectroscopy, but still no obvious structure-function relation has been addressed (Perez-Alvarado *et al.*, 2003). This leads to the question I have set forward to address in this section:

How is the structure of REF proteins related to their functional role?