INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

University of Alberta

CECR6: Evidence of Two Overlapping Open Reading Frames in the Cat Eye Syndrome Critical Region

by

Marie Diane Isabelle Mousseau 

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science

in

Molecular Biology and Genetics

Department of Biological Sciences

Edmonton, Alberta

Spring 2005

DEDICACE

Je voudrais dédier cette thèse à mes parents Pierre et Hélène pour leur support et leur amour infini.

ABSTRACT

The partial tetrasomy of chromosome 22 Cat Eye Syndrome Critical Region is responsible for heart, kidney and eye birth defect in humans. The sequence of the Cat Eye Syndrome Critical Region gene 6 (CECR6), one of 14 putative genes in that region, supports two overlapping open reading frames (ORFs). Bioinformatics and orthologous sequence comparison support the existence of both ORFs. Reverse transcription confirmed the presence of two different version of the *CECR6* mRNA: *CECR6*a which contains the large single exon ORF1 and *CECR6*b which splices at least once and prevents the translation of the larger ORF. Computer prediction programs elucidated the structure of the larger ORF as a transmembrane protein with multiple amino acid runs. This research has given weight to the possibility that *CECR6* could be the fifth example of overlapping genes in alternative reading frames in the human genome.

ACKNOWLEDGEMENTS

There are many people and organizations I would like to thank for having made this research possible: Dr Heather McDermid, my supervisor, for her thoroughness and her patience, for her love of science and of people. Drs Ross Hodgetts and Michael Hendzel for reviewing and making suggestions to this thesis. Past and present members of the McDermid lab, especially Stephanie Maier for being a great friend. Dr. Warren Gallin, Dr. Bryan Crawford, Dr. Wayne Materi and Lynn Podemski for valuable discussion. The Pilgrim, Wong, Nargang, Cox and Locke labs for equipment. Patricia Murray and Lisa Ostafichuk at the MBSU for their teaching. The Edmonton Valley Zoo for the Red Fronted Lemur blood. The Department of Biological Sciences for financial support and valuable teaching experience.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

CES	Cat Eye Syndrome
CECR6	Cat Eye Syndrome Critical Region 6
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	Messenger RNA
aa	Amino acid
tm	Transmembrane
bp	Base pair
kb	Kilobase pair
UTR	Untranslated region
UTRdb	UTR database
ORF	Open reading frame

CHAPTER 1. INTRODUCTION

Overlapping coding sequences on the same DNA strand in the human genome are few. Mathematically, if approximately 35,000 genes exist within the 3.2 billion base pairs of the human genome, there could be an average of 300,000 bases between genes (Veeramachaneni, Makalowski *et al.* 2004). However, genes within the human genome are not equally distributed: some DNA regions are rich in genes while others are gene deserts. Coding regions represents about 1.5% of the human genome (Mignone, Gissi *et al.* 2002). The constraints of sharing DNA sequence for overlapping genes seems too high in large eukaryote genomes, where most genes are found alone on a sequence, each having their own promoter region and transcribed in their own mRNA. The discovery of the existence of the *CECR6* gene in the Cat Eye Syndrome Critical Region (Footz, Brinkman-Mills *et al.* 2001) potentially adds a new example of overlapping coding sequences in the human genome. Using bioinformatics, the *CECR6* locus shows two possible open reading frames (ORF) on the same DNA strand. The possibility of the two ORFs being functional relies on their potentially joined transcription but independant translation. The research described in this thesis was aimed at proving the functionality of both open reading frames (ORF) of *CECR6* in the human Cat Eye Syndrome Critical Region.

1. Cat Eye Syndrome

1.1 Cat Eye Syndrome is used as a gene overexpression model

Cat eye syndrome (CES) is also referred to as Schmid-Fraccaro syndrome or Chromosome 22 partial tetrasomy (Online Mendelian Inheritance in Man -OMIM # 115470). CES is considered a model for gene overexpression in humans. Overexpression is a major cause of abnormal human development and pregnancy loss. While many deletion syndromes are well studied, there are few regions that lend themselves to the study of the overexpression of genes.

CES is a rare genetic disorder (1:50,000 to 1:150,000, OMIM) caused by supranumerary copies of the short arm and part of the long arm of chromosome 22, usually in the form of a partial tetrasomy creating a bisatellited, dicentric chromosome containing a duplication of the region 22q11.2. These supernumerary chromosomes can be inherited or appear *de novo* in an individual due to the unstable nature of this region of chromosome 22. The diagnosis is usually made using fluorescent in situ hybridisation (FISH) with a centromeric chromosome 22 marker. CES involves defects of the eye, heart, face, urogenital and skeletal systems and mental development that vary from patient to patient. The features of the disease are variable even within families where members with the same CES chromosome can have mild to severe phenotypes. The most characteristic abnormalities (Schinzel, Schmid *et al.* 1981) include anal atresia with fistula (absent anus), periauricular tags or pits (single small skin growth by the ear), kidney or heart defects and iris coloboma (visible cleft of the iris) to which the disease owes its name. However, coloboma of the iris is not particular to CES: it is diagnosed in

1 to 8 cases per 100,000 births world-wide (Gregory-Evans, Williams *et al.* 2004) and affects only about half of CES patients.

Most, if not all, of the features of this syndrome map to a 1 Mb critical region (the distal CES Critical Region), which is now cloned and sequenced (Footz, Brinkman-Mills *et al.* 2001). Fourteen putative genes have been identified in this region by the McDermid lab (Figure 1.1). Genes within this region that could be responsible for CES phenotypes would have to be dosage sensitive. Those include transcriptional regulators, receptors, signal transduction molecules and structural proteins. Based on expression during development and putative function from sequence motif analysis, the McDermid group is concentrating on the functional analysis of three of these genes: CECR1, a putative adenosine deaminase, CECR2, a chromatin remodelling protein and *CECR6*.

1.2 Preliminary characterisation of *CECR6*

CECR6 is a one exon gene with a putative leucine zipper near the predicted carboxyl end (Footz, Brinkman-Mills *et al.* 2001). *CECR6* mRNA is supported by ESTs in the NCBI database located in the unusually long 3' untranslated region (3'UTR, 3097 bp), which accounts for 62% of the *CECR6* mRNA length (4958 bp, Figure 1.2). The known 5'UTR accounts for only 2% (111 bp). Northern blot analysis done using one of the published ESTs as a probe has shown that the *CECR6* mRNA is ubiquitously expressed, with highest levels in brain and prostate but also in skin and heart (Figure 1.3, P. Brinkman-Mill, unpublished data, 1999). Initial *in situ* hybridisation of whole mount and sections of mouse embryos has revealed a diffuse general staining. No function or tertiary structures were deciphered using computer comparison programs.

More importantly, *CECR6* mRNA has the potential to code for two completely different novel proteins, which overlap in different reading frames. Thus one *CECR6* mRNA could translate two completely different proteins (Figure 1.2). Many examples of overlapping genes have been found in higher eukaryote genomes while the mechanism is widely used in prokaryote and small eukaryote genomes. The term overlapping is used at many levels in the literature. While not all overlapping genes share coding sequence on the same DNA strand, all share a part of their mRNA sequence, whether the UTRs or the introns on either strand of DNA.

2. Overlapping Genes in the Human Genome

While genes are usually linked to one locus, the term overlapping refers to genes with a part of their transcribed DNA sequence encompassing one another. Shared sequence between loci is a feature commonly found in organisms with very small genomes, originally thought to save space.

2.1 Gene overlap in bacteria: saving space or evolutionary drive?

Organisms have evolved numerous mechanisms to mask the effect of a mutation or to remove the mutation from the genetic pool. Gene overlap is said to be an antiredundant mechanism (Krakauer and Plotkin 2002). Antiredundant mechanisms in small organisms will respond to a point mutation by removing mutant genomes from the population and also include codon bias, co-ordinated expression of genes and checkpoint genes. Redundant mechanisms on the other hand include duplicated genes, correlated gene function and alternative metabolic pathways. These qualities increase the size of a

genome to mask the effect of a mutation and a single organism may contain both redundant and antiredundant properties.

Overlapping genes are a common feature of prokaryotic genomes. While the main function of overlap is usually thought to be conservation of space, new evidence obtained by comparing the number, spacing and conservation of overlapping genes in publicly available microbial genomes (Johnson and Chisholm 2004) leads to a different conclusion. The correlation between the number of overlapping genes and the size of genomes (related to the total number of genes) was very high (0.96), and consistent across Archaebacteria and Eubacteria, in both bacterial chromosomes and plasmids. While a third of all genes were shown to overlap, the frequency of overlap was not linked to the compactness of genomes calculated by the distance between genes and the percentage of coding sequence, nor the GC content of the genome. The only significant difference between overlapping and non-overlapping genes was the number of homologues in other species for each overlapping gene compared to single genes; overlapping genes have a 13% greater chance of having a homologue.

Overlapping genes also exist in large eukaryote genomes, such as mammals. Evidence of overlapping reading frames in the human genome has grown over the last couple of years. Genes can overlap in many ways: genes can be located on opposite DNA strands, single exon genes inserted in another gene's intron, or genes with coding regions overlapping each other on the same DNA strand. Interestingly, 84% of overlapping bacterial genes are located on the same DNA strand and use different reading frame (Johnson and Chisholm 2004). As seen in the following paragraphs, this is opposite to higher eukaryote genomes.

2.2 Genes with sequence overlapping on opposite DNA strands

The mouse and human genomes contain thousands of overlapping genes on opposite DNA strands (Veeramachaneni, Makalowski *et al.* 2004) categorised in different conformations: head to head (5' overlapping 5' region, 30% of cases), tail to tail (3' overlapping 3' region, 50 to 60% of cases) and embedded (16% of cases where no exons overlap and only 3% of cases where exons do overlap) (Figure 1.4). More importantly, only 10% of genes with exons overlapping have overlapping coding regions on different DNA strands.

Overlapping genes on different DNA strands are thought to evolve by many mechanisms: overprinting (creation of a novel gene from pre-existing DNA sequence) (Keese and Gibbs 1992) or by the loss of a polyadenylation site by one gene, and the presence by chance of another site within an opposite strand gene's sequence (Shintani, O'HUigin *et al.* 1999). The gene that lost the polyadenylation signal then extends to the new signal and overlaps the second gene. It is reasonable to think that these two hypotheses can also tentatively explain the creation of novel overlapping genes on the same DNA strand. ORFs can arise from pre-existing sequence in a limited number of mutational steps by the addition of a methionine codon, removal of a stop codon or small insertions or deletions of bases creating a frame shift. The loss of a polyadenylation site and the use of one further down the sequence, can occur for genes on the same strand, the difference being that this new longer mRNA will contain part or all of the sequence of the second gene in the right direction for translation. In higher eukaryote genomes, the occurrence of overlapping genes on the same DNA strand, compared to genes on opposite DNA strands, is rarely found, contrary to bacterial genomes.

2.3 Nested genes

There are few examples of genes present within another gene's intron. The following examples illustrate the different possibilities of overlap of genes on opposite strands sharing sequence: single exon genes and Overlapping Gene Groups. Early gene sequence analysis (Levinson, Kenwrick *et al.* 1990) identified the presence of an intronless gene on the opposite strand, within intron 22 of the human blood coagulation Factor VIII DNA sequence (Figure 1.5a). A similar overlap situation had already been described in Drosophila in 1986 and 1987 (Henikoff, Keene *et al.* 1986; Chen, Malone *et al.* 1987). The 2 kb transcript gene F8A (coagulation factor VIII- associated transcript 1) is present in the largest intron of factor VIII (39 kb) on the opposite strand. Although F8A is a single exon gene, multi-exonic genes can also be present within introns.

Overlapping Gene Group (OGG) (Karlin, Chen *et al.* 2002) refers to a multi exon gene within a large intron, usually the first or last of another gene with both genes on opposite DNA strands. There are 34 known OGGs on human chromosome 22 alone. Some human OGGs are conserved in mouse and even Drosophila. In some cases, there is a small overlap of coding regions between the two genes. An example of OGG is TIMP3 (tissue inhibitor metalloproteinase 3), located within an intron of SYN3 (synapsin-III) on chromosome 22 (Figure 1.5b).

2.4 Genes with coding sequence overlapping on the same DNA strand

Examples where most of the overlap occurs in the coding regions are rare and involve ORFs on the same DNA strand. The LGALS3 locus is a complex example where overlap between coding regions on the same DNA strand occurs. An internal promoter was found in the second intron of the human gene Galectin-3 (LGALS3) (Guittaut,

Charpentier *et al.* 2001). The promoter can drive the expression of three predicted ORFs which could result in multiple initiation sites and alternative transcripts of LGALS3 gene on the same strand. The third ORF (ORF3) is in frame with the known LGALS3 gene and would produce a truncated version of the protein. Subcellular localisation of each ORF tagged with EGFP showed that ORF3 is not translated. ORF1 and 2 are out of frame with LGALS3 and translated at the same efficiency from the same mRNA called *galig* (Galectin-3 internal gene). ORF1 and 2, while overlapping another gene's exon, are also overlapping each other on the same DNA strand.

The mature *galig* mRNA includes a portion of intron 2 and exons three to six of LGALS3. ORF1 and ORF2 correspond to alternative initiations of translation at different AUG codons within the *galig* transcript in different frames (Figure 1.6a). Using the luciferase protein tag for subcellular localisation showed that ORF1 and ORF2 localise differently within a cell in accordance with the properties of each predicted protein. ORF2 is hydrophobic and localises to the mitochondrial membrane while ORF1 is cytosolic. The LGALS3 and *galig* promoters were shown to be regulated differently as the mRNAs show different tissue specificity (Guittaut, Charpentier *et al.* 2001). The protein sequence of each of those ORFs was not recognized by any bioinformatics programs in 2000.

There are only three other known examples of alternative reading frames of a coding sequence. The initial example was the INK4a gene which codes for an inhibitor ($p16^{INK4a}$) of CDK4 and CDK6 and prevents exit from the G1 phase of the cell cycle. It was found (Quelle, Zindy *et al.* 1995) that the gene gives rise to two transcripts differing only in their first exons, E1α or E1β (Figure 1.6b). The predicted ORF created by E1β

used a different reading frame in exon 2 than that of $p16^{INK4a}$ leading to the discovery of a protein called $p19^{ARF}$ (alternative reading frame) which didn't share the amino acid sequence of $p16^{INK4a}$. $p19^{ARF}$ is also involved in cell cycle regulation which could explain their tight affiliation.

The most recent example of overlapping genes on the same DNA strand (Poulin, Brueschke *et al.* 2003) involves the fusion of an alternate DNA sequence, an independent 34 exons gene called MASK (multiple ankyrin repeats, single KH domain), to the beginning of the gene EIF4EBP3 (eukaryotic initiation factor 4^E binding protein 3) through an intermediate exon (E0). MASK flanks the 5' region of EIF4EBP3. The new splice variant uses the second exon of EIF4EBP3 in a different reading frame (Figure 1.6c). The alternative transcript is called MASK-$BP3^{ARF}$ and is not expressed like either of the original proteins.

The final case involves a functional relationship between the overlapping genes ALEX and XLαs/Gαs. A second ORF is possible within the large XL-exon of the XLαs/Gαs gene (Klemke, Kehlenbach *et al.* 2001). The new ORF called ALEX (alternative gene product encoded by the XL-exon) starts 32 nucleotides downstream of the XL-domain start site and terminates at the end of the XL-exon (Figure 1.6d). ALEX is conserved in rat, mouse and humans and contrary to the first paper where the alternative ORF was said not to share the same mRNA (Klemke, Kehlenbach *et al.* 2001), ALEX is translated from the same mRNA as XLαs (Abramowitz, Grenet *et al.* 2004). Furthermore, the ALEX protein binds the XL-domain of the XLαs protein *in vivo* suggesting that their arrangement on genomic DNA insures a similar expression pattern. An elongated version of the ALEX sequence found in some families, causes decreased

binding between XLαs and ALEX. This has been linked to brachydactyly and platelets and fibroblast disorder leading to increased bleeding in traumas (Freson, Jaeken *et al.* 2003).

2.5 The use of computers to predict overlapping genes

A downfall of the search for overlapping genes is the currently available bioinformatics programs for eukaryote genomes. "Gene finder" programs such as GENSCAN (Burge and Karlin 1997) can only recognise one gene in a region of one DNA strand even if many ORFs exist. This means that there could be many examples of overlapping genes in the human genome that will require a case by case elucidation. Available solutions to the limitations of these programs include the search for internal promoters (an ambiguous process) or the sequence disruption of the main ORF sequence prior to a gene search to allow the second ORF to be recognised. These methods are time consuming and do not allow work on many loci at a time. For example, careful sequence analysis on individual genes of the CES critical region was required to identify two large overlapping ORFs in the single exon gene *CECR6* (Footz, Brinkman-Mills *et al.* 2001). Lowering the threshold of putative gene sequences (codon usage and other gene characteristics can be different in a gene whose sequence evolution is driven by the gene it overlaps) and allowing overlapping results in the program output would allow the recognition of more overlapping genes. A new algorithm (MLOGD) was developed to detect parts of sequences in a pairwise sequence alignment that could be "double-coding" (Firth and Brown 2005). The alignment was optimised for small ORFs and is useful to detect overlapping genes in small genomes. Although this method could not detect three overlapping coding reading frames such as the LGALS3 locus, it produced results with

90% confidence when using optimal alignments. Results were lower for real sequence alignments. Characteristics of the sequence driving the expression of eukaryotic genes such as promoters, initiation codon domain and untranslated regions can be informative in the case of overlapping genes.

3. Initiation of RNA transcription and translation of proteins

The transcription of mRNA in eukaryotes is a complex and well regulated process. The transcription of an mRNA does not depend on the coding sequences it carries but rather on its promoter region and other regulatory elements that will allow its transcription. The presence of two functional reading frames within an mRNA therefore has no bearing on its transcription other than those aspects of its regulation that have evolved to ensure proper functioning of the translated protein(s).

3.1 RNA transcription is initiated downstream of a promoter site.

The eukaryotic promoter is a cis-acting DNA element located upstream of the start of the mRNA transcription. The promoter sequence, although performing the same basic function throughout eukaryotic genomes, is more variable than coding regions. Hence theoretical studies of the promoter sequence alone cannot lead to conclusive regulatory information (Wray, Hahn *et al.* 2003). The accepted general eukaryotic promoter model involves the presence of three DNA sequences: the TATA, CCAAT and GC boxes respectively located at 30, 100 and 200 bp from the start of transcription. Since these short sequences have a high probability of random appearance in the genome, promoter searches cannot lead to conclusive results on the beginning of a specific RNA.

Because of this, the only way to see if the two *CECR6* ORFs share the same mRNA is to experimentally find supporting RNA evidence in human cells.

Assuming that both *CECR6* ORFs are located on a single mRNA, the translation of each polypeptide must occur individually. Translation is initiated when a ribosome, scanning from the 5'cap of the mRNA, recognises a methionine as the initiation codon (Kozak 2001). Nearly 50% of human mRNA contain a methionine codon upstream of the translation initiation codon (Mignone, Gissi *et al.* 2002). These methionine codons must differ from the one chosen for initiation and in the case of two overlapping ORFs, both methionine initiation codons have to be recognised by a ribosome to produce two separate translation frames. The recognition of the second initiation codon requires the first one to be bypassed by the ribosome to start translation downstream in a different frame. This possibility is called "leaky scanning" and was hypothesised to explain the translation of the ALEX gene overlapping the first exon of the XLαs/Gαs gene (Kozak 2001). The initiation codon is recognised by the ribosome probably because of the "Kozak" sequence immediately surrounding the methionine.

3.2 A weak Kozak sequence on the first initiation sites could explain the leaky scanning model.

The Kozak consensus base pair sequence of 5'-GCCRCCATGG-3', where R is a purine (A or G), comes from a comparison of the base pairs surrounding the initiation site of a large number of vertebrate mRNAs (Kozak 1987a; Kozak 1987b). Some positions close to the initiation codon AUG were identical throughout the majority of the numerous genes pooled in databases and were deemed necessary to facilitate translation initiation. The most important positions (where the A of initiation codon AUG is in

position +1) are an R in position –3 and a G in position +4. This theory has been confirmed experimentally by targeted mutagenesis experiments (Kozak 1987) and the Kozak sequence is now used in commercial protein expression kits to ensure proper translation of the desired DNA sequence (Promega TNT® Quick Coupled Transcription/Translation Systems, catalogue number L1170).

Following the original conclusion on the role of the sequence surrounding the initiation codon, it was found that leaky scanning not only allowed translation of a second ORF but was responsible for regulating translation within one mRNA (Kozak 2001; Kozak 2002). Because the Kozak sequence, which facilitates the initial binding of the small ribosomal unit to the first methionine codon, is not as strong as it could be, the ribosome is sometimes able to reach the second initiation site. This was shown using the ALEX example (Kozak 2001).

The effect of upstream methionine codons on downstream regulation can account for the 5' untranslated region (UTR) potential for regulation. Upstream methionine codons are found in relatively long 5' UTRs and contain weak Kozak sequences (Rogozin, Kochetov *et al.* 2001; Xiong, Hsieh *et al.* 2001). This could decrease or regulate levels of translation compared to an mRNA with a short 5' UTR and no upstream AUG. The translation rate of the main ORF can also be affected by the presence of small ORFs in the 5' UTR. A review of the UTRdb (http://bighost.area.ba.cnr.it/BIG/UTRHome/) revealed several examples of mRNA with long 5'UTRs (between 1 and 3 kb) showing upstream ORFs (uORFs), including transcription factor Pax-5 (AF074913: 2 uORFs), Growth/differentiation factor 1 (GDF-1, M62302: 1 uORF) and the c-Myc proto-oncogene (AJ000928: 13 uORFs) (Mignone,

Gissi *et al.* 2002). Since almost all ribosomal attachment to the mRNA depends on the 5'cap (Kozak 2001), followed by a linear scanning of the mRNA, the presence of an upstream ORF decreases the chances of translation of the main ORF, thus decreasing the protein product levels. In a study of the *tat* mRNA in the human immunodeficiency virus (HIV) type 1 genome, it was found that the main ORF tat was followed by two ORFs *rev* and *nef*, which are not efficiently translated (Luukkonen, Tan *et al.* 1995). Using point mutation to introduce stop codons in the *tat* ORF, it was found that the efficiency of translation of the downstream ORF *rev* was inversely proportional to the length of the *tat* ORF as well as the distance between the two ORFs. The ribosome seems to be able to reinitiate translation if the distance between the two ORFs is smaller than 30 codons but longer distances allow the ribosome to fall off most of the time, hence the second ORF is translated at a reduced rate (as reviewed in (Mignone, Gissi *et al.* 2002)). This regulation is not applicable to mRNAs coding for overlapping ORFs, only to ORFs that are following each other. The difference is that in overlapping ORFs with alternative use of a coding sequence on a single mRNA, the second initiation codon can only be used by a ribosome that has not recognised the first initiation codon. While the 5'UTR can play a role in translation regulation, the 3'UTR also has regulatory functions. The *CECR6* mRNA was shown to have a very long 3'UTR (Footz, Brinkman-Mills *et al.* 2001), which could be responsible for important characteristics of the *CECR6* protein(s) expression.

3.3 UTRs play an important role in regulation

Untranslated sequences and primary and secondary structure of an RNA molecule can influence its transport, translation efficiency, subcellular localisation and degradation

(Kuersten and Goodwin 2003). Specific sequences within the UTRs can be binding sites involved in a variety of regulatory events (Mignone *et al.* 2002). The 5'UTR is usually more conserved than the 3'UTR because it contains regions that control the translation (Grzybowska, Wilczynska *et al.* 2001) but conserved domains have also been found in 3' UTRs. Within 21 different UTR structures listed (Mignone, Gissi *et al.* 2002), 14 belong exclusively in the 3' UTR. This is a result of the role of the 3'UTR. If the 5' UTR regulates translation, cis-acting elements of the 3'UTR are responsible for transport, degradation and subcellular localisation. Subcellular localisation of mRNA can be more energy efficient than protein transport since a single mRNA can be used to translate many copies of a protein, although diffusion of soluble proteins requires no energy from the cell. In fact, the fate of the mRNA depends more on the 3'UTR, which has more leeway in length and sequence. The only widespread constraint on the 3'UTR is to contain the polyadenylation signal to terminate the mRNA.

The transcription, translation and regulation of mRNA in eukaryote genomes allows for the possibility of overlapping genes. In addition to the information gathered from a single sequence from computer programs and promoter, Kozak and UTR examination, comparison between orthologous sequences in a variety of species can be informative in the case of genes, overlapping or not.

4. Alignments with distantly or closely related species lead to different sequence information

How useful are sequence comparisons between species? The theory is that regions that are responsible for important biological processes will be conserved through

evolution. While bioinformatics results can be artefacts of the parameters or even of a particular DNA sequence for one species, the comparison of multiple species can reveal conserved DNA sequences or protein structural domains. Alignment of highly dissimilar sequences (Thomas, Touchman *et al.* 2003) can lead to many uncertainties because of the tendency of alignment programs such as Clustal W and MUSCLE to cluster matching regions and thus miss many small conserved regions. The most commonly used alignments that include the human sequence can be divided into three categories depending on the phylogenetic distances. Different information can be obtained from close relatives with very high sequence homology (human-primate), distant relatives with very little sequence homology (human-fish) and relatives in between (human-mouse).

4.1 Human and mouse sequence: eliminating false positives

Large-scale human-mouse sequence comparisons such as the CES Critical Region comparison (Footz, Brinkman-Mills *et al.* 2001), are popular because the phylogenetic distance between the two species is far enough to eliminate false positives while conserving interesting mammalian regulatory sequence. Such comparisons serve three purposes: identification of new genes, gene-regulatory elements and transcription factor binding sites within non coding regions. But the phylogenetic distance between human and mouse may not be appropriate to study regions that are highly conserved or too diverged since evolution is not homogeneous throughout chromosomes. More information can be obtained by alignments with other species or by "Phylogenetic footprinting" (Nobrega and Pennacchio 2003), which is a technique that uses multi-species sequence comparisons to give higher resolution of conserved regions of a domain, to the order of a few base pairs. The small region conserved in all species,

although not necessarily relevant, has the possibility to be a promoter or an enhancing sequence.

4.2 Primate sequences: phylogenetic shadowing shows regions less conserved

Comparison between closer species to humans (primates) or more distant species (fish, Drosophila) can lead to more precise results. The mutation rate, calculated by the rate of neutral mutations in different regions of the genome, can show many fold variation within a genome. Overall, the rate of independent genome evolution has been estimated at about 0.1 to 0.5% per million years and a study of over 5000 mammalian genes showed that the mutation rate was constant per year and similar within the genes studied (Kumar and Subramanian 2002). Primates share a common ancestor 6 to 8 million years old. Human, chimpanzee and gorilla share 98 to 99% of their sequence, both coding and non-coding (Hacia 2001). This high degree of conservation can point to important regions through the comparison of many primate species sequence using a technique named "phylogenetic shadowing" (Boffelli, McAuliffe *et al.* 2003). This process compiles the phylogenetic distance between every primate and the common ancestor of a phylogenetic tree. It assumes that regions that have regulatory importance will be conserved in all primates. Incorporation of all amino acid or nucleotide differences in a final alignment consensus identifies regions less conserved with at least one mutation present in one primate. Regions that do not show differences may have a regulatory function.

4.3 Human and fish sequences: conserved regions are of importance

Fish is the most distant vertebrate from humans with a common ancestor dating back 400 to 450 million years. The zebrafish and fugu fish genomes are respectively in

the process or are considered completely sequenced (Aparicio, Chapman *et al.* 2002). Sequencing of the fugu genome has lead to the identification of 1000 new human genes in conserved regions between the two species (Nobrega and Pennacchio 2003). Conserved regions that do not have the potential to be coding regions could correspond to important regulatory sequences. An advantage of the smaller size of the fish genomes (365 million base pairs for fugu), is that those regulatory regions will be located closer to the genes they affect, simplifying the analysis (Gilligan, Brenner *et al.* 2002).

4.4 Overlapping gene patterns can be conserved between species

A study of the NCBI human (April 2003, 34 604 genes) and mouse (March 2003, 33 936 genes) genome assemblies lead to the identification of 774 (2.2%) overlapping gene pairs on opposite DNA strands in humans and 578 (1.7%) overlapping gene pairs on opposite DNA strands in mouse (Veeramachaneni, Makalowski *et al.* 2004). Of the human overlapping pairs, mouse orthologous genes were found for both genes in only 255 (33%) cases and were found to be overlapped in 95 (12%) cases while others (150 pairs) were located on the same genome assembly contig. Only 10 human overlapping gene pairs were found in different mouse contigs, suggesting that the close position in the mouse genome contributed to eventual overlap of the genes. In parallel, of the 578 mouse overlapping gene pairs, only 240 had two human orthologs and 95, predictably, were overlapping in humans. In 144 cases, the human homologues were located on the same human contig and did not overlap. In only one case were the genes on different contigs. The mouse overlap could have occurred after the divergence of the mouse and human ancestors or could have been lost in humans. This dynamic of gene overlap between human and mouse does not favour nor falsify an evolution towards overlapping. In

addition, a comparison of the degree of conservation between human and mouse genes overlapping on opposite DNA strands did not show a greater average of conservation than single gene orthologs. This was true for both UTR and coding region overlap. Looking at patterns of overlap between the 95 human and mouse orthologous overlapping gene pairs, a "significant fraction" showed different overlap patterns. This differs from an earlier report that claimed that all overlapping gene pairs they had looked at showed the exact same overlap pattern (Shendure and Church 2002). Using the SRR gene example, the 2004 paper blames unfinished 3' and 5' UTR sequencing for this discrepancy. The human SRR gene's 3'UTR overlaps with the 3'UTR and coding region of FLJ10534. In mouse, the Srr 3'UTR and coding regions overlap only with the 3'UTR of LOC193029. This reversed pattern of overlap can be explained by the short 3'UTR of both mouse genes that would show a different pattern of overlap if they were as long as their human homologues. Although this explanation reflects the overlapping gene evolutionary hypothesis related to the 3'end overlap (Shintani, O'HUigin *et al.* 1999), this explanation cannot account for every pattern of overlap nor the potential loss of overlap between human and mouse gene pairs. Other theories such as overprinting to generate novel genes (Keese and Gibbs 1992) or chromosome rearrangement, also have supporting evidence in the 2004 dataset, which suggest a variety of mechanisms in the evolution of overlapping genes.

Limiting the phylogenetic comparisons to just close or distantly related species will not reveal all regulatory or coding regions. The combined use of closely related species (human-primates) and distant species (human-fish) gives information on primate specific regulatory regions or novel genes, as well as regulatory regions located far from

known genes in primates but that are close to the gene in question in fish. Comparison between orthologous overlapping gene regions can give information on the evolution and function of the overlap. The human chromosome 22 pericentromeric CES Critical Region shows conserved synteny with a region on mouse chromosome 6 where all genes but one (CECR1) have a mouse orthologue.

5. *CECR6*: a locus with overlapping coding sequences in the CES Critical Region?

In light of the growing interest in overlapping genes, gaining information on the possible functionality of both predicted ORFs was important for further research. The analysis was done through sequencing and comparison of orthologous *CECR6* sequences in a variety of species, the use of computer prediction programs and reverse transcription RNA analysis. Orthologous sequence comparison did not rule out the functionality of the second ORF and reverse transcription confirmed the presence of two different versions of the *CECR6* mRNA: *CECR6*a which contained the large single exon ORF and *CECR6*b which splices at least once to create a possible second reading frame in the coding region while eliminating the translation of the larger ORF. Computer prediction programs elucidated the structure of the larger ORF as a transmembrane (tm) protein with multiple amino acid runs. This research has given weight to the possibility that *CECR6* could be the fourth example of overlapping genes in alternative reading frames in the human genome.

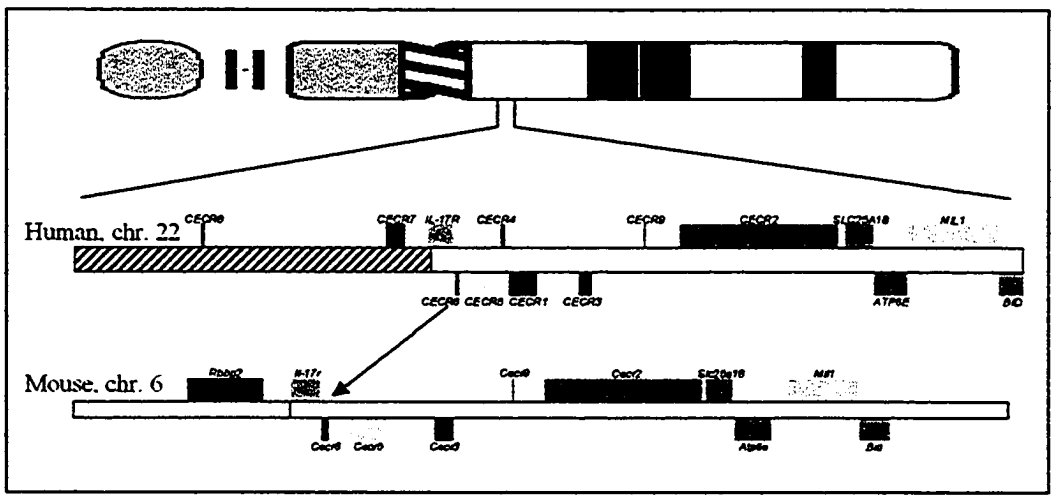


Figure 1.1: Cat Eye Syndrome Critical Region, narrowed down to one megabase of DNA on human chromosome 22, has a region of conserved synteny on mouse chromosome 6. The *CECR6* locus (in green) is located near the pericentromeric region of chromosome 22 on the minus strand (orientation towards the centromere). CECR6 is flanked by IL-17R (on the opposite strand) and by CECR5. Adapted from Footz *et al*, 2001.

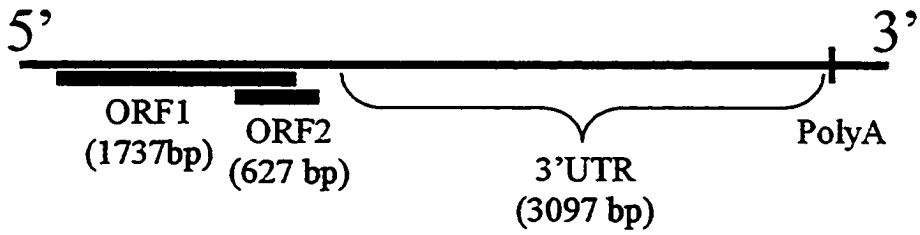


Figure 1.2: Organisation of ORF1 (green) and ORF2 (red) on the *CECR6* locus (blue). ORF1, the larger ORF at 1737 base pairs, starts upstream of the smaller ORF2 (627 bp). There is a 90% DNA sequence overlap of ORF2, whose termination codon is located 69 bp downstream of the ORF1 stop. The 3'UTR located between the ORF2 termination codon and the polyadenylation signal (PolyA) is unusually long (3097 bp) and contains no repeats. The ORFs were predicted using the ORF Finder function in Genetool version 1.0.

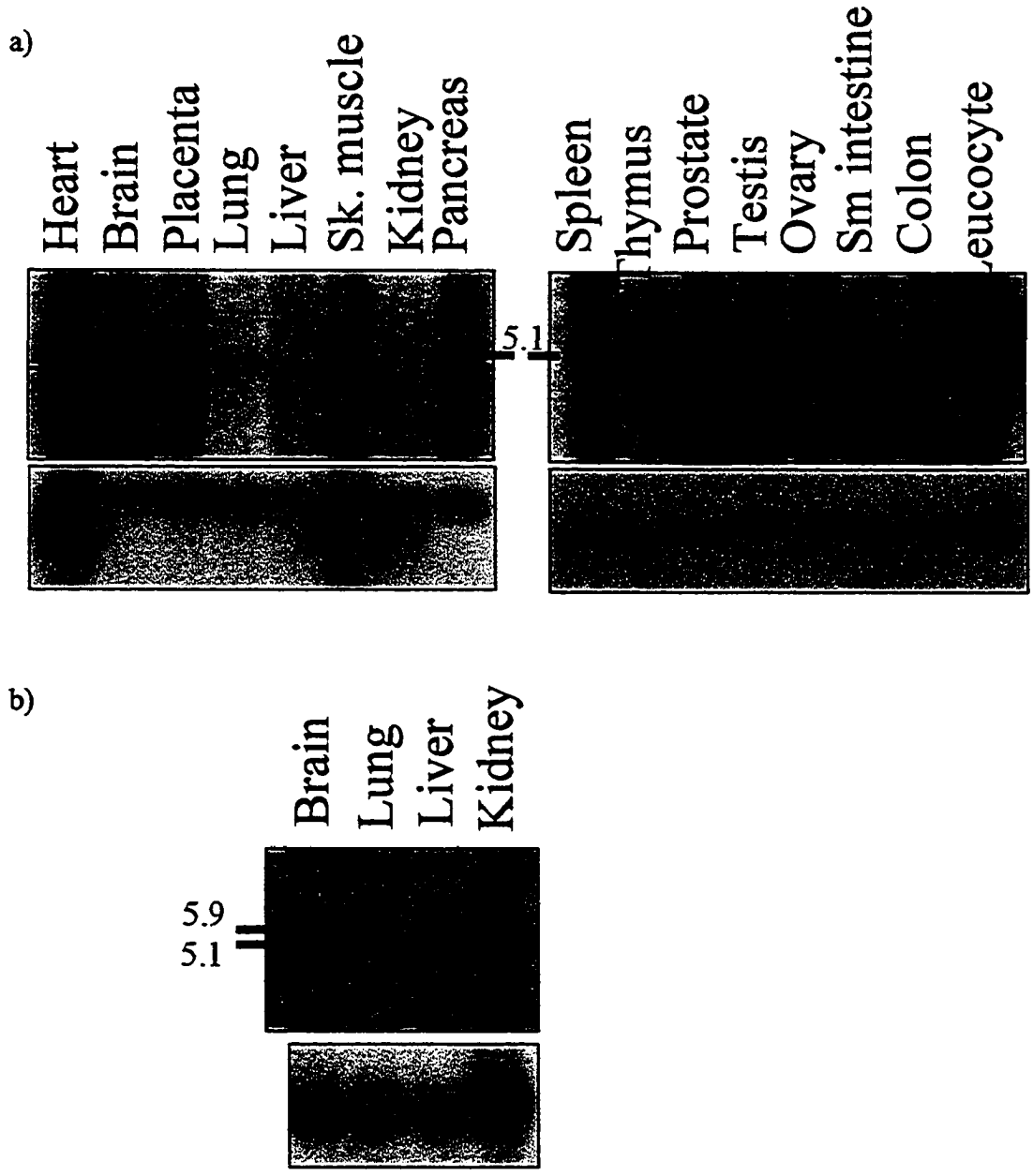


Figure 1.3: Adult and foetal human CECR6 RNA hybidization analysis. a) Stronger signal for the CECR6 mRNA (running at 5.1 kb) can be seen in adult heart, brain, prostate, testis and periferal blood leucocytes. b) The hybridization to foetal total RNA shows increased amounts of the CECR6 mRNA in brain and a larger band (5.9 kb instead of 5.1 kb) in liver. The larger band may also be present in brain which the thickness of the band would mask. The probe used to hybridize the total RNA ran on this agarose RNAse free gel is located at the end of the 3'UTR, which will pick up all versions of the CECR6 mRNA. The lanes were also probed using the beta-actin sequence as a loading control which causes thicker bands to be seen in heart and skeletal muscle due to cross-hybridization of the probe with alpha-actin. This Northern analysis was performed by P. Brinkman-Mill in 1999.

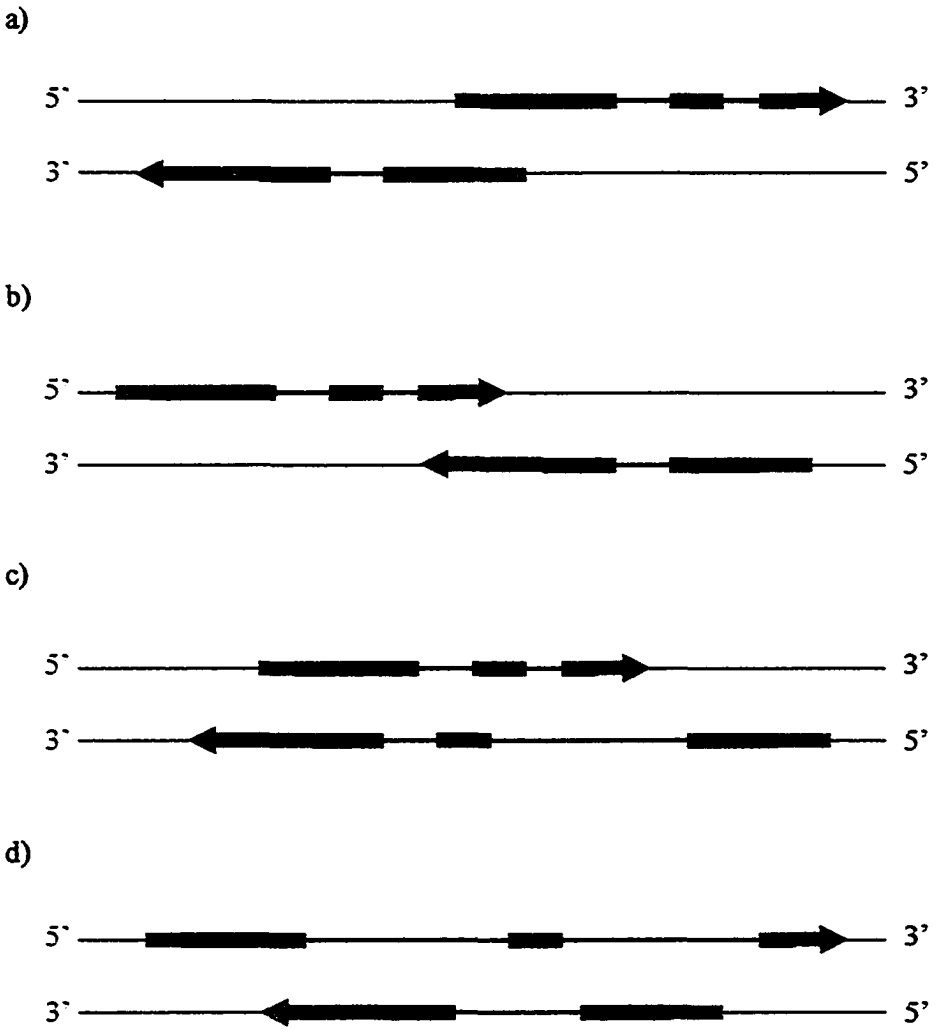
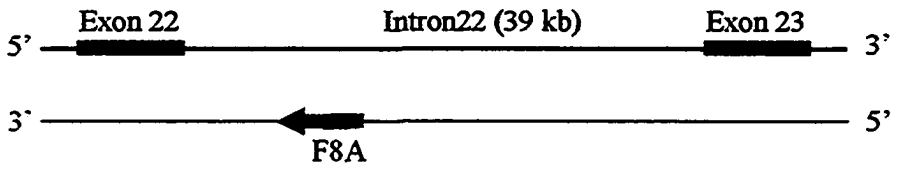


Figure 1.4: Overlapping ORFs on opposite DNA strands. Genes can overlap a) head to head, b) tail to tail, or can be embedded with c) exon overlap or d) no exon overlap. Thicker lines represent the regions contained in the mature mRNA transcript with dark lines representing coding region and lighter lines, UTRs. Thinner lines between dark boxes are introns. The arrow points in the direction of the transcript and stops at the polyadenylation signal. Promoter sequences are omitted but would be located upstream of the boxes, sometimes overlapping the other gene region. Adapted from Makalowska *et al*, (2004).

a) Factor 8



b) SYN3

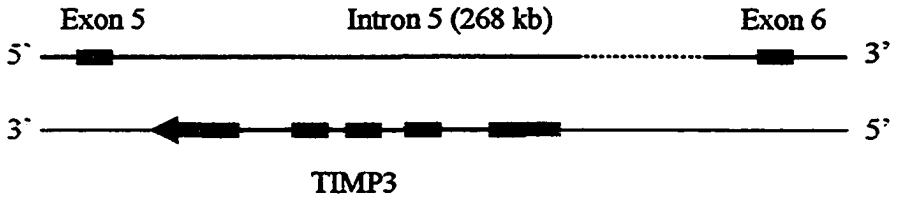
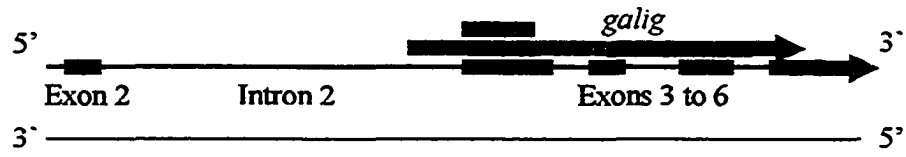
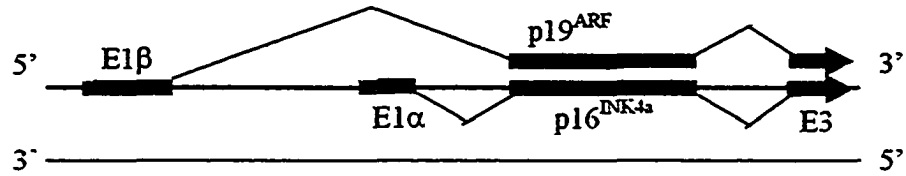


Figure 1.5: ORFs nested in inton of a multi exonic gene. Genes within other genes introns usually overlap on different DNA strand. a) The gene F8A (coagulation factor VIII-associated intronic transcript 1), is present in the largest intron (39 kb) of F8 (coagulation factor VIII) on human chromosome Xq28. b) The five exon gene TIMP3 is located the 268 kb intron 5 of SYN3 on human chromosome 22 (adapted from Karlin, Chen et al. 2002). Thicker lines represent mature mRNA transcript sequence with black lines as coding regions and grey lines, UTRs. Figures not to scale.

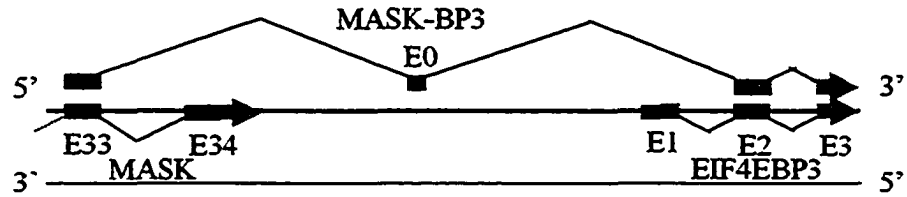
a) LGALS3



b) INK4a



c) EIF4EBP3



d) XLαs

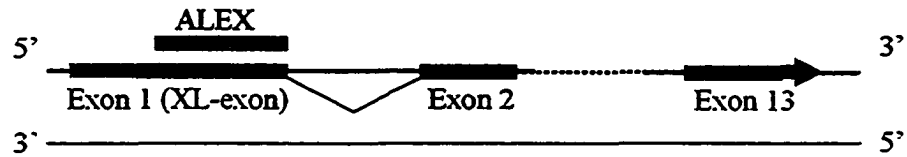


Figure 1.6: Four examples of overlapping coding regions using alternative reading frames on the same DNA strand. a) The *galig* transcript spans intron 2 to exon 6 of LGALS3 on the same DNA strand. ORF1 and ORF2 are translated from the *galig* mRNA. b) Alternate use of the first exons E1α or E1β in the INK4a gene leads to two different proteins $p16^{INK4a}$ or $p19^{ARF}$ which have alternative reading frame in the second exon. c) Alternative splicing between two separate genes MASK and EIF4EBP3 lead to the novel protein MASK-BP3 which uses E2 in a different frame. d) The single exon gene ALEX spans the XL-domain of XLαs. The ALEX protein, translated in a different reading frame, binds to XLαs and regulates its action. Overlapping genes are represented on each side of the 5'3' strand. Thicker lines represent mature mRNA transcript sequence with black lines as coding regions and grey lines, UTRs.

CHAPTER 2: MATERIAL AND METHODS

1. DNA isolation

The *CECR6* orthologous DNA region was obtained from PACs and BACs containing respectively human and mouse DNA contigs, and from genomic DNA of different mammalian species. All centrifugation steps were done at room temperature and maximum speed (35,000 g) in an Eppendorf microcentrifuge.

1.1 Plasmid DNA from bacteria (miniprep procedure)

PAC (P1-based artificial chromosome) 143i13 (AC005300) and PAC 109l3 (AC006946) each contain the full human *CECR6* genomic sequence. BAC (bacterial artificial chromosome) 541L22 (NT_039382) includes the *CECR6* mouse genomic DNA region. To recover the vectors from the bacterial stocks, overnight cultures grown in 1.5 µL LB medium were spun for 30 seconds and the pellet was re-suspended in 100 µL of Alkaline solution I (50 mM glucose, 20 mM Tris-Cl pH 8.0 and 10mM EDTA pH 8.0). 200 µL of Alkaline solution II (for 1 mL made fresh: 880µL ddH_2O, 100 µL of 10% SDS and 20 µL of 10M NaOH) was added and the tubes were placed on ice after being inverted several times. 150 µL of Alkaline solution III (for 100 mL: 60 mL of 5 M KOAc, 11.5 mL glacial acetic acid and 28.5 mL ddH_2O) was added and the tubes were incubated on ice for 5 minutes followed by centrifugation for 5 minutes. 5 µL of 10mg/mL RNase A was added to the supernatant followed by 30 minutes of incubation at 37°C. These steps were then followed by a phenol-chloroform extraction (Shambrook J. 2001).

A solution of phenol: chloroform: isoamyl alcohol (25:24:1) was added to the plasmid mix. The solution was then mixed and spun for 2 minutes. An equal volume of chloroform: isoamyl (24:1) was added to the aqueous layer. Again, the solution was mixed and spun for 2 minutes. Two volumes of 95% EtOH were added to the aqueous phase to precipitate the DNA. The solution was mixed and placed at –20°C for 30 minutes to increase the DNA yield and spun for 15 minutes. The pellet was then washed with 1 mL of 70% EtOH, spun for 5 minutes, and dried 10 minutes. The pellet was dissolved in 10-50μL of ddH_2O depending on the pellet size.

1.2 Mammalian DNA

1.2.1 Lemur blood

The Edmonton Valley Zoo graciously provided 1 mL of Red Fronted lemur blood. DNA was extracted using 12% DTAB (12% dodecyltrimethylammonium bromide, 2.25 M NaCl, 150mM Tris-HCl pH 8.6, 75 mM EDTA at room temperature), and 5% CTAB (5% cetyltrimethylammonium bromide, 0.4 M NaCl at room temperature) following a modified protocol (Gustincich, Manfioletti *et al.* 1991).

1.2.2 Rhesus monkey DNA

Rhesus Monkey DNA was purchased from Clontech (6860-1).

1.2.3 Other mammal DNA

Chimp, gorilla, orangutan, gibbon, rat and rabbit DNA were donated by the Diane Cox lab at the University of Alberta, Department of Medical Genetics.

2. DNA sequencing

DNA sequencing was performed on an ABI Prism 377 DNA sequencer at the Molecular Biology Service Unit of the Department of Biological Sciences.

1.1 Sequencing reaction

Sequencing reactions used Amersham Biosciences DYEnamic™ ET Dye Terminator Kit (MegaBACE™). This cycle sequencing kit worked well in high-salt conditions for "less pure template". The high processivity of the DNA polymerase decreased the cycle time and the dye labelled nucleotides were suitable for long templates (over 500 bp of accurate sequence) and GC rich regions. Half of the reaction prescribed by the Amersham Biosciences protocol was used (4 μL of DYE ET solution, 1 μL of 2 pmol specific primer (Table 2.1), and 4.5 μL of DNA to sequence). 0.5 μL of Pellet Paint® Co-Precipitant by Novagen was added to the reaction to stain the DNA pellet blue during the ethanol precipitation phase.

1.2 Sequencing PCR program

A three-step program was used on a PCR machine (PTC 200 DNA Engine™ system Peltier Thermal Cycler by MJ Research) to add fluorescent nucleotides: 95°C for 3 minutes, then 30 cycles of 95°C for 30 seconds, 50°C for 30 seconds, 60°C for 1 minute.

1.3 Preparing the PCR reaction for sequencing

In order to pellet the dye labelled DNA, 40 μL of 95% EtOH and 1 μL of 3 M NaOAc/EDTA was added to the PCR reaction before 30 minutes of incubation at -20°C.

The solution was then spun at maximum speed for 15 minutes at room temperature. The supernatant was decanted and the pellet washed with 200 μL of 70% EtOH and spun at maximum speed for 5 minutes at room temperature. The supernatant was decanted and the pellet dried at room temperature for 10 minutes. The pellet was re-suspended in 1.5 μL of formamide used also as loading buffer.

1.4 Sequencing gel analysis

Sequencing gel data was computed using the program ABI Prism 377-96 Data Collection from PE Biosystems version 2.6. The gel analyses were performed with ABI Prism DNA Sequencing Analysis Software version 3.4.

3. DNA amplification by PCR

3.1 Primer design

All primers were designed from sequences suggested by the computer program Genetool versions 1 and 2 by BioTools Incorporated (Table 2.1).

3.2 PCR reaction

Dr. Michael Pickard from the Department of Biological Sciences, University of Alberta, provided the Taq polymerase and PFU polymerase. Phylogeny PCR was done using Invitrogen Taq.

3.2.1 General PCR reaction

The total volume of the general reaction was 50 μL, consisting of 5 μL of 10xPCR buffer, 4 pmol of each primer, 1 μL of 10mM dNTP, 2 μL of DMSO, and 1 to

10 ng of DNA. For the "hotstart", 1 μL of Taq polymerase: PFU (25:1) was added during the 80°C phase of the PCR program. This late addition of the enzyme prevents its degradation due to high temperature of the first phase. The hotstart was used in touchdown PCR to maximise chances of a single PCR product.

3.2.2 Enhancer system

Invitrogen PCRx Enhancer System was used where the general PCR failed to amplify a product. This system increases primer specificity and allows amplification of GC-rich regions or problematic templates (stable secondary structures for example). The addition of the PCRx Amplification Buffer and of the PCRx Enhancer Solution in lieu of the 10x PCR buffer and the DMSO respectively, were the only differences in the Invitrogen protocol.

3.2.3 Colony PCR reaction

Bacterial colonies were picked and lysed mechanically using a small piston in 100 μL TE pH 8.0 (10mM Tris and 1mM EDTA) to re-suspend the plasmid DNA within other cellular particles. Two microlitres of the suspension was used in a 25 μL PCR reaction. Other reaction components included: 4 pmol of each primer, 2.5 μL of 10xPCR buffer minus $MgCl_2$ by Invitrogen (y02028), 0.75 μL of magnesium chloride, 0.5 μL of 10 mM dNTP and 0.5 μL of Taq polymerase: PFU (24:1).

3.3 PCR programs

All PCR reactions were done on the PTC 200 DNA Engine™ system Peltier Thermal Cycler by MJ Research. The acronym A.T. used in the following sections stands for annealing temperature, which differs for every set of primers.

3.3.1 Touchdown PCR

Most PCR reactions were performed using a touchdown PCR program to increase the specificity of the primers. The program started with 95°C for 2 minutes followed by a 80°C stall to add TAQ polymerase (to perform the "hotstart"). This was followed by 10 cycles of 95°C for 30 seconds, then the A.T. +6°C in the first cycle, decreasing by 0.6°C every cycle thereafter, for 30 seconds, and 72°C for 2 minutes. The DNA was then amplified by 30 cycles of 95°C for 30 seconds, A.T. for 30 seconds and 72°C for 2 minutes. The program ended with 72°C for 10 minutes and 4°C ongoing to preserve the PCR solution.

3.3.2 Gradient PCR

This PCR cycle was used to attempt many different A.T. at the same time to troubleshoot primer pairs: 95°C for 3 minutes, 80°C stall, 35 cycles of 94°C for 15 seconds, a 10 to 30°C range of A.T. for 20 seconds and 72°C for 2 min. The program ended with 72°C for 10 minutes and 4°C ongoing.

3.3.3 Colony PCR

The colony PCR was used to amplify bacterial DNA: 95°C for 3 minutes, 36 cycles of 94°C for 15 seconds, A.T. for 20 seconds, and 72°C for 2 minutes. The program ended with 59°C for 20 seconds, 72°C for 10 minutes and 4°C ongoing.

3.4 DNA resolution

All DNA samples were resolved on 1% agarose in 1x TBE gels. The gels were run at 10V per centimetre of length.

Samples were fractionned with 10x Stop buffer (50% glycerol, 10 mM Tris pH 7.5, 100 mM EDTA, 0.1% Bromophenol Blue, 0.1% Xylene Glycol, 0.1% Orange G). The ladder was made with 10x Stop buffer and 10 kb+ ladder from Invitrogen.

4. RNA manipulations

4.1 RNA source

Human tissue total RNA was extracted by previous lab members. Foetal tissue and adult brain and prostate tissue total RNAs were purchased from Clontech.

4.2 Reverse Transcription Polymerase Chain Reaction (RT-PCR)

4.2.1 Reverse Transcription

The reverse transcription was performed using Invitrogen Thermoscript™ RT-PCR System (11146-016). Differences from the Invitrogen protocol included using twice the amount of DNaseI to ensure no DNA contamination and the use of 0.8 µL instead of 1 µL of Thermoscript. The Gene Specific Primer (GSP) program was 42°C for 30

minutes, and 50°C, 53°C, 55°C, 57°C, 60°C each for 10 minutes. The program ended with 85°C for 5 minutes.

4.2.2 Primers

All primers were designed using the Genetool 2 computer program. A reverse primer in the 3'UTR near the *CECR6* ORFs was used for the production of cDNA from RNA, as *CECR6* 3'UTR is too long to use the Oligo dT primer provided with the kit. Primers for the following PCR reaction are listed in Table 2.1b.

4.2.3 Polymerase Chain Reaction

The touchdown PCR program described in 3.1.1 was used to amplify the cDNA sequence, using 40 cycles instead of 30 to allow more product to be amplified.

5. Bioinformatics programs

The world wide web offers a variety of small programs tailored for specific molecular predictions. Most of these programs run directly on the web without installation need. Larger programs with multiple tasks such as Genetool must be purchased and installed on particular computers. Internet search engines include Genamics (http://genamics.com/index.htm) and Google for scientists (http://scholar.google.com/) .

5.1 Database searches for similar sequences

The most commonly used database search program is BLAST (basic local alignment search tool). It is a mathematical algorithm that searches for similar small "words" (3 letters for amino acids and 11 for nucleic acids) in the query sequence and databases to come up with similarity scores. The BLAST output shows sequence alignments of the query and the database using the Smith-Waterman algorithm. The National Center for

Biotechnology Information (NCBI) at the US National Library of Medicine has dedicated a powerful computer system to run BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) but the program is available through other servers. BLAST2SEQ is one of the programs offered by NCBI and identifies the most conserved region between two sequences (Tatusova and Madden 1999).

5.2 Pair-wise and multiple sequence alignment program

Pairs of sequences can be aligned using three methods: the Dot matrix method which places each sequence on the two axes of a graph and plots the similarities (http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html), the word method used in the BLAST algorithm and the Dynamic programming method. The dynamic programming method can perform the most accurate alignment available by comparing every pair of characters in both sequences using optimisation parameters. Match and mismatched characters and gaps are created so that the match between identical characters is the maximum possible.

Multiple sequence alignment (MSA) can use the dynamic programming method for a limited amount of sequences (only three long or up to eight short sequences) due to the computational burden of a multi-dimension matrix with predetermined scores. Programs that use the progressive method of MSA can handle a greater number of sequences by first building a tree from a pair-wise sequence comparison and then building the MSA, starting with the most related sequences and progressively adding less related sequences. The CLUSTAL programs, using this method, were introduced in 1988 and have been progressively refined (Mount, 2001). The program works best on closely related sequences and provides a good indication of conserved domains. The most recent

CLUSTALW (W for weight) allows global MSA of DNA and proteins (http://www.ebi.ac.uk/clustalw/) and was used to produce all the alignments in this work. Sequence similarities was then highlighted by running the CLUSTALW output through Boxshade 3.21 (http://www.ch.embnet.org/software/BOX_form.html).

5.3 DNA sequence utility programs

5.3.1 Genetool version 1 and 2

The program Genetool version 1 and 2 by BioTool Incorporated includes many sequence utility features. The multiple sequence alignment tool, which I used to merge newly sequenced pieces together, uses the FastLSA (Fast linear space alignment) program algorithm. Genetic Reference Point Logistics (GRPL) a method related to logistic regression (used for pattern recognition) and developed at the University of Alberta (Hooper, Zhang *et al.* 2000), is used for ORF prediction.

5.3.2 Other ORF finder programs

ORF finders use more than the presence of initiation and termination codons to predict potential ORFs. For example, the Hidden Markov Model (HMM) was first used to recognise linguistic patterns and is used in DNA sequence identity because even if coding sequences are "hidden" in a DNA sequence, it contains certain characteristics such as codon bias and length of ORF (Stormo 2000). This model used in the program GENSCAN (http://genes.mit.edu/GENSCAN.html) amongst others, is not made to recognise potential overlapping reading frames.

5.4 Protein sequence utility programs

5.4.1 Primary and secondary structure prediction

Programs used to predict *CECR6* putative proteins structure can be found on the proteomic server Expasy (Expert Protein Analysis System: http://au.expasy.org/) by the Swiss Institute of Bioinformatics (SIB). They include PredictProtein (http://cubic.bioc.columbia.edu/predictprotein/), 3D-PSSM (http://www.sbg.bio.ic.ac.uk/~3dpssm/) and Threader (http://bioinf.cs.ucl.ac.uk/threader/threader.html). Prediction of subcellular localisation was done using the program PSORT (http://www.psort.org/). Translation of the nucleotide sequence into amino acids was done using the Expasy translate tool (http://au.expasy.org/tools/dna.html) and the BCM Search Launcher (http://searchlauncher.bcm.tmc.edu/) also used for sequence formatting.

5.4.2 Prediction of membrane spanning domains

Most programs are found on Expasy (Appel, Bairoch *et al.* 1994). They include: HMMTOP (Tusnady and Simon 2001) (http://www.enzim.hu/hmmtop/), SOSUI (Hirokawa, Boon-Chieng *et al.* 1998) (http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html), TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/) , SMART (http://smart.embl-heidelberg.de/), Split Server (http://split.pmfst.hr/split/4/) and CDART from NCBI (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps). Differences in predictions are due to different algorithms and threshold. The Hidden Markov Model TMHMM (Sonnhammer, von Heijne *et al.* 1998) is considered the most accurate (Moller,

Croning *et al.* 2001) although all of these programs only provide the topology of the protein and not the 3D structure (von Heijne 1999).

a)

Primer Name	5'3' Primer DNA sequence
Phylo-AF	cgcgccccgggatcggcaggac
Phylo-BR	taggtgacagaaggtagaagacaagcg
Phylo-CF	cgcggccggatgcttcctgggcac
Phylo-DR	ctgaaaagggttacctcttccaaata
Phylo-ER	gtatattgctggtctccctccaagc
Phylo-FR	gcggaaaccctaaagccaagaacag
Phylo-HF	cctacctggcctggcttatctact
Phylo-IF	gcttatctactccatcgccttcac
Phylo-JF	cttccgtctcaccatggcgctgtc
Phylo-KF	gcgctgtcggtgcccctgctctac
Phylo-LR	ggtctccctccaagccgtcctcac
Phylo-MR	gcacctacctccctcaccgttaac
Phylo-NR	ggtgggcaaagcaagaagcagagg
Phylo-OF (mouse)	gcccagcgaacagcatgcacaacc
Phylo-PF (mouse)	ttgcgtacctggcctggctcatct
Phylo-QF (mouse)	ggctcatctactccatcgctttca
Phylo-RF (mouse)	ggctcggcgggtcccctgctcctg
Phylo-SR (mouse)	ggtctccctccaagcagtctctac
Phylo-TR (mouse)	gggcccttcttcctcacggtatat

b)

Primer Name	5'3' Primer DNA sequence
F14790 (F2)	ttcaccgcctgccccctctc
F14833 (F1)	cagtccagtggctccagtcc
R16445 (R1)	agcaagggcacgtccaccag
R3	gcgagggtgaggaagtagacg
FM1 (m)	gagctagagacttcatgttc
FM2 (m)	gcctggcctgatacgctttc
RM1 (m)	ctcgctaatggcacgcacta

Table 2.1: DNA oligonucleotides (primers) used to amplify parts of the *CECR6* locus in a) the phylogenetic experiments and b) the RT-PCR on human (primers Phylo-AF to NR) or mouse (primers Phylo-OF to Phylo-TR) total RNA. Primers were purchased desalted from either Invitrogen or Qiagen. Primer stock was diluted to 200nM and then to 4nM for PCR reaction or 2nM for sequencing.

CHAPTER 3: RESULTS

As previously published (Footz, Brinkman-Mills *et al.* 2001), the *CECR6* mRNA sequence supports two alternate overlapping open reading frames (ORF) in one predicted coding region (Figure 1.2). This idea came from the fact that a long stretch of the human *CECR6* mRNA sequenced by the McDermid group shows very few termination codons (UAA, UAG or UGA) in two different reading frames (1 and 3 in Figure 3.1a) while the third frame (2 in Figure 3.1a) shows termination codons scattered throughout the region (Figure 3.1a and 3.1b). Initiation codons (AUG) are also present in frames +1 and +3 leading to two potential ORFs with 90% of ORF2 overlapping ORF1 on the same DNA strand. The mouse *CECR6*, in comparison is lacking an initiation codon in the ORF2 frame and thus seems to only support one ORF (Figure 3.2a and 3.2b). The two ORFs have the potential to belong to two separate overlapping genes leading to two potentially functional proteins. The first potential coding sequence, ORF1 spans 1737 base pairs, which corrsponds to a 578 amino acids polypeptide. ORF2 mesures 627 base pairs translated into 209 amino acids.

1. Putative protein sequence of ORF1: a membrane protein with multiple amino acid runs

While ORF2 did not show any sequence or structure characteristics that could help identify its function, primary and secondary structure analysis of ORF1 has lead to the following findings.

1.1 Multiple Amino Acid Runs

The predicted protein sequence of ORF1 is highly unusual; with numerous four to eight amino acid repeats such as polyproline, polyhistidine, polyglycine and polycysteine stretches. To be significant, not due to random chance, the amino acid repeat must be over five amino acids long in a protein of 400 residues (Karlin *et al*, 2001). A significant stretch is called a "multiple amino acid run" and a non-significant stretch, a multiplet. The predicted ORF1 contains both. The stretches of more than three amino acids can be summarised using the single letter amino acid notation (Table 3.1): $P_4A_3G_5G_5S_3D_3S_3S_3G_3A_3C_5R_3L_3V_4H_4G_3A_5L_3A_7P_8$ (Figure 3.3). This notation does not take into account repeats separated by only a few nucleotides (G_5RRG_5 could have the same properties as a G_{10} repeat for example) or regions rich in one amino acid (HNH_2LH_4).

The amino acid repeats found in *CECR6* ORF1 are not encoded by a single DNA codon the way polyglutamine diseases correspond to CAG triplets which would indicate polymerase slippage (Petruska, Hartenstine *et al.* 1998). Instead, the eight proline repeat for example, is composed of three different codons (CCT, CCA, CCG) and reads: cct/cca/cca/cct/ccg/cca/cca/cct. The seven alanine repeat (GCT, GCC, GCA, GCG) reads: gcc/gcg/gcc/gca/gcg/gct/gca. Due to this use of multiple codons and the codon third base "wobble", ORF2 in frame +3 shows very few repeated amino acids.

1.2 *CECR6*a is a membrane protein

*CECR6*a amino acid sequence is not recognised by any tertiary structure programs such as 3D PSSM (Kelley, MacCallum *et al.* 2000) or Threader (Jones, Tress *et al.* 1999). This lead to further bioinformatic analysis showing that *CECR6*a in fact contains a

number of transmembrane (tm) domains. It is unfortunately very difficult to decipher the number of domains without extensive protein tagging experiments. The computer programs CDART (Geer, Domrachev *et al.* 2002), TMHMM (Figure 3.4) (Sonnhammer, von Heijne *et al.* 1998) and SMART (Ponting, Schultz *et al.* 1999) predict four strong tm domains while SOSUI (Hirokawa, Boon-Chieng *et al.* 1998) predicts seven, and Split Server (Juretic, Zoranic *et al.* 2002) predicts six. Differences in predictions are due to different algorithms and threshold. The consensus between all programs seems to be that four tm domains are strongly supported and two more are less supported by most programs.

The presence of positively charged amino acids in three of the predicted tm domains throughout *CECR6*a is noticeable because as little as one charged amino acid can confer voltage sensitivity to the tm domain (Figure 3.3). The presence of a single proline residue in three of the predicted tm domains is also significant as prolines can facilitate the helical structure needed for tm because of the chemical structure of the amino acid.

By comparing amino acid repeats and tm domains (Figure 3.3), a negative correlation can be seen. In fact no multiple amino acid run and only two multiplet are found in predicted tm domains. To understand the importance of a sequence such as a multiple amino acid run or a predicted tm domain, comparison between the orthologous *CECR6* sequence of many species is used to outline conserved regions and even the evolution of a characteristic (lengthening of a repeat, appearance of a tm domain).

1.3 ORF2, a putative soluble protein

ORF2 amino acid sequence requirements may be secondary to that of ORF1 since no secondary or tertiary structures can be found using computer programs. The only conclusion as to a characteristic of the putative ORF2 polypeptide is that it is probably soluble (SOSUI).

2. Phylogenetic comparison of orthologous *CECR6* sequences

In order to understand the mutations that created ORF2 and to gain insight into the function of ORF1, a comparison of the *CECR6* locus in a variety of species was made using available NCBI database sequence (human, olive baboon, cow, mouse, and three fishes) and by sequencing a part of the *CECR6* locus of several species (six primates, rabbit and rat). The initiation and termination codons, the conservation of the reading frames and the degree of conservation of different regions of the two ORFs across species were studied to address the existence of both ORF1 and ORF2 in humans.

2.1 ORF1 and ORF2 in mammals

2.1.1 ORF1 is conserved in mouse, baboon and humans

The larger open reading frame ORF1 is conserved in mouse, baboon and human. Comparison between the human (NCBI accession number NM031890), olive baboon (AC091672) and mouse (NM033567) *CECR6* orthologous sequences on the NCBI database show that the human putative protein sequence shows 86% identity (only identical amino acids are included) to the mouse sequence (Figure 3.5). When amino acids that differ but have the same properties are taken into account the percentage rises to 89% similarity between human and mouse. The human sequence also shows 98%

identity to the baboon predicted protein which is consistent with other evolutionary data (Hacia 2001). Phylogenetic alignments using a larger number of sequences (Figure 3.6) in the following section will demonstrate this similarity.

2.1.2 ORF2 initiation codon is primate specific

The smaller ORF2 (209 amino acids) is conserved in baboon and human (89% similarity at the amino acid level) but is thought not to be present in mouse due to the loss of the initiation methionine codon, replaced by a point mutation to produce a leucine. In order to see if this mutation was particular to mouse or if ORF2 was a primate novelty, I used primers (Table 2.1a and Figure 3.6c) made to conserved regions in human and in mouse to amplify and sequence the *CECR6* ORF2 region in various mammalian species (Figure 3.7). Some sequences were obtained from the NCBI database to complement the alignment. The multiple sequence alignment of primates, cow (AAFC01729699), rat (NM_033567), rabbit and mouse amino acid sequence in frame +3 (Figure 3.6b) shows that the *CECR6* ORF2 initiation codon is conserved in primates only, which supports the potential use of that codon but can also be due to the phylogenetic closeness of primates. The Red Fronted Lemur, phylogenetically furthest from humans in this alignment, shows the methionine in position 1 of ORF2. It appears that the first methionine of ORF2 is not conserved in cow, rodents and rabbits in frame +3 (Figure 3.6b). Three methionine codons are found in frame +1 in the region overlapping ORF2 (Figure 3.6a). A point mutation in the codon changes the amino acid to leucine in all four species. According to the Tree of Life Web Project (http://tolweb.org/tree/phylogeny.html) the sub-groups lagomorphs (rabbits, hares and picas) and rodents (including murinae, porcupine and squirrel) are the two branches of the group "Glires", thus rabbit, mouse and rat are

closely related and the same leucine codon can be expected. Cows, in the Artiodactyla are closer to Cetacea than to primates or glires. Importantly, there is no other methionine in the reading frame to take over the start of transcription, but by comparing the rest of the ORF2 amino acid sequence, no termination codon can be found disrupting the possible ORF and the frame is conserved in all species studied.

2.1.3 The region unique to ORF2 is only conserved in primates

Most of the ORF2 region overlaps with ORF1. Only the 3' end of ORF2, after the ORF1 stop, is not shared. The evolution of this region is important for the existence of ORF2 since a DNA comparison of the 3' untranslated region between human and mouse reveals very low conservation. If this region mimicked the 3'UTR properties in mouse, then the lack of conservation would indicate that ORF2 was not present. Conservation of the region between ORF1 and ORF2 termination codons in primates but not in mouse (Figure 3.8) indicates that ORF2 cannot exist in mouse but is a possibility in primates. Comparison of the region between the two protein stop codons in any frame shows that the mouse sequence similarity to the human sequence ends with the ORF1 stop codon (Figure 3.8b) and the region between ORF1 and ORF2 stop codons behaves like the rest of the 3'UTR sequence (low sequence conservation and frame shifts). The human and baboon sequence alignment shows conservation of that region (Figure 3.8a). This similarity could be due to the close relationship of those species but doesn't refute the existence of ORF2 in primates.

We hypothesised that if ORF2 was translated in primates, the last third of ORF1 would be more conserved than the first two thirds of ORF1 because a mutation to the last third of ORF1 would have an effect on two genes and thus be less likely to have a neutral

effect. There is no evidence of the last third of ORF1 being more conserved in primates due to the already high nucleotide sequence conservation (Figure 3.6a and 3.6b). It is more interesting to look at conservation of the same region in the ORF2 frame because more silent mutations can occur in the third base pair of each codon due to the redundancy of the genetic code. These mutations, neutral in frame +1, would be the first base pair of the codon triplets in ORF2's frame +3. The alignment (Figure 3.6b) shows more mutations than in frame +1 but the distribution of the mutations does reveal a definite ORF2 conservation constraint. Using sequences available on the database (human and olive baboon), the full sequence of the *CECR6* ORF1 and ORF2 coding region can be compared. The repartition of differences in frame +1 (Figure 3.5) and frame +3 (Figure 3.9) shows that the mutation rate seems to be the same throughout the region.

2.2 ORF1 is present in fish.

Recent additions to the NCBI sequence database of large genomic clones and cDNAs of a variety of species have facilitated the search for *CECR6* homologues. The zebrafish genomic clone BX950180 (September 2004), the tetraodon cDNAs CR671181.2 and CR664705.2 (December 2004) and the channel catfish partial cDNA CV989369.1 (December 2004), all support homologous sequence to human *CECR6* ORF1. The predicted ORF is about 100 base pairs shorter in the fish, due to some gaps within the sequence and initiation codon, fifty to sixty base pairs into the human ORF1 sequence. No multiple amino acid runs and very few multiplets are found in the fish sequence, meaning that those were positively selected in the mammalian sequences.

Analysis by TMHMM v.2 revealed that all the putative fish proteins had multiple transmembrane domains (Figure 3.4).

The putative protein sequence alignment shows that regions predicted to be tm domains by the TMHMM program in the fish proteins (Figure 3.10) correspond to the predictions on the mammalian proteins. These domains are also very conserved, up to 95% conservation in a twenty amino acid sequence, considering the large phylogenetic distance between mammals and fish. While human ORF1 show four strongly and two less predicted tm domains, the zebrafish and the tetraodon sequences have six strongly predicted tm domains. Interestingly, the less strongly predicted domains in humans correspond to the two extra fish tm domains, a possible clue as to the number of real tm domains in human ORF1. Another clue on functional domains of the ORF1 protein lies in the region between the fourth and fifth fish tm domains, which shows strong conservation but does not correspond to tm domains. This conserved region between fish and mammals is not recognised by computer programs as a known domain.

2.3 *CECR6* mRNA has conserved regions in the large 3' UTR

There are over 3 kb between the ORF1 stop codon and the polyA signal. According to AceView (NCBI), this length is in the top 5% in size in the available 3'UTR database. Three regions of approximately 60 bases appear to be of importance as they are 80% conserved between the *CECR6* human and mouse mRNA sequence but no know pattern could be recognised in the URTdb nor significant sequence similarity or folds could be seen. (Figure 3.11).

3. RNA analysis reveals an intron in the *CECR6* mRNA, producing an alternate mRNA

3.1 Expressed Sequence Tag (EST) in the NCBI database

The NCBI program AceView (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly) counts 40 cDNA clones supporting the *CECR6* mRNA in the NCBI database as of February 2005. Most of those clones only span the 3' UTR due to its very long nature, which complicates reverse transcription of the *CECR6* coding region by companies using an oligo dT oligonucleotide priming on the polyadenylation tail of the RNA. The 4973 base pair sequence of the human mRNA (accession number AF307451), including the coding region was obtained by Polly Brinkman-Mills in 2000 in the McDermid lab by conducting 5' RACE analysis. It is still the only sequence of a *CECR6* transcript containing the ORF1 and ORF2 coding region available on the NCBI database.

There are two categories of ESTs which differ from the typical *CECR6* intronless gene (Figure 3.12): ESTs on the opposite strand and ESTs showing spliced out sequence. Three opposite strand ESTs were found up to 10 kb upstream of *CECR6* (BC021739, BF223077, BI561230). No genes have been predicted in this region. However, the PIP analysis (Footz, Brinkman-Mills *et al.* 2001) showed conserved sequence between human and mouse, upstream of *CECR6* (Figure 3.13), which correlate to the ESTs locations. Further analysis by reverse transcription of these regions could lead to the discovery of a new gene in the CES critical region on the opposite strand of *CECR6*.

Two clones were found to correspond to the *CECR6* mRNA sequence but were missing regions of DNA. Clone BB647465 is a 400bp RIKEN clone and is composed of 3 small human DNA fragments with boundaries that do not correlate to exon/intron

consensus junction sequence. This could be an aberrant clone due to the cloning methods used to create RIKEN clones which allowed fragments from different areas to rearrange together or pieces of the original sequence to be deleted. Random digestion and cloning can lead to many inserts coming together in a single vector which are then picked up as one single aberrant sequence.

Clone AK095609 spans the entire *CECR6* mRNA region but is missing 1108 bp. The clone was found in a full insert sequence library done by 5' oligocapping sequencing. The reverse transcription is done from the 5'cap of the mRNA which in the case of the *CECR6* mRNA, is located closer to the coding region than the polyadenyl tail and allowed the *CECR6* coding region to be included. The spliced out piece start and ends with the expected GT/AG exon/intron boundary. The first 100 bp exon lies upstream of the ORF1 start and contains sequence upstream of the published 5'UTR start. The second exon spans the last third of ORF1 and the 3'UTR of the *CECR6* mRNA. This clone cannot support ORF1 and the splice cannot be an ORF1 splice because the clone doesn't contain the ORF1 initiation codon or its first two thirds. There are two possibilities as to the translated ORF supported by this clone: ORF2 or a truncated version of ORF1 (trORF1), using one of the three methionines present in the region where both ORF overlap. The trORF1 using the first methionine present after the splice would lead to a single tm domain protein with two amino acid runs (A_7P_8). This methionine used for trORF1 is also present in the mouse *CECR6* ORF1 sequence. Confirmation of the existence of this spliced version of the *CECR6* mRNA would lead to additional information as to the use and regulation of the two potential overlapping

ORFs. In addition, clues to the potential translation efficiency of each ORF can be obtained by studying the region surrounding the initiation codon,

3.2 Kozak sequence

The Kozak sequence is a nucleotide consensus (5'-GCCRCCATGG-3') in the eukaryotic translation initiation region that was determined by comparing a wide variety of sequences (Kozak 1987). The closer an initiation region is to the Kozak consensus, the better the ribosome recognition and the stronger the initiation should be. The conservation in some positions seem to be more important than in others, notably positions –3 (R) and +4 (G), where A is in position +1. Comparison of the regions around the ORF1 and ORF2 putative initiation codons as well as the region surrounding the methionine codons present in ORF1 (Figure 3.1b) to the optimal Kozak sequence reveals that ORF1 and ORF2 do not show strong Kozak sequences. Of the three methionine codons found within the ORF1 sequence present on the *CECR6* mRNA containing the intron, the strongest Kozak sequence is found around the first methionine leading to a 223 amino acid trORF1. This methionine is conserved in baboon, mouse and zebrafish but not tetraodon (Figure 3.10). The two other methionines (labelled #2 and #3, Table 3.2) show a very weak Kozak sequence and would lead to very small ORFs (166 and 179 amino acids respectively) thus will not be considered as potential ORFs further.

3.3 The *CECR6* intron is present in humans

To follow the AceView program notation, the unspliced version of the *CECR6* mRNA was called *CECR6*a and all the spliced versions, *CECR6*b.

I designed PCR primers (Table 2.1b) to test the existence of the AK095609 intron. Total RNA from human foetal brain and foetal liver showed strong bands on

CECR6 human Northerns (Figure 1.3) and total RNAs for these tissues were available in both human and mouse, thus they were chosen to look for the *CECR6*b mRNA isoform. Human brain was also the tissue used to produce clone AK095609 that showed the *CECR6*b splice. Reverse transcription PCR (RT-PCR) using the F1-R3 primer pair (Figure 3.14b) gave the two expected bands (with and without the predicted intron, Figure 3.14a, panel A) using human foetal brain total RNA. RT-PCR on foetal liver using F1-R3 leads to the same banding pattern but DNA bands were consistently too faint to sequence (data not shown). Both *CECR6*a and *CECR6*b are present at the same time in the foetal brain and liver tissue. All RT-PCR results were confirmed by sequencing. A fainter band migrating at about 1000 bp was seen in the foetal brain lane which proved to be impossible to sequence. It was probably an artefact of the PCR reaction.

In the same tissue, RT-PCR using a different forward primer F2 (50 base pairs upstream of F1) did not produce the bigger band (no splice) but rather only the smaller band (splice) (Figure 3.14a, panel B). This experiment was repeated with the same result, with primers R1 and R3 interchangeable. Since both versions are present in the tissue, the F2 primer is not able to hybridise to *CECR6*a, which could mean that *CECR6*a and *CECR6*b have different mRNA start sites. Since the primers are about 50 bp apart, the start of the two mRNA isoforms must be at least this distance apart.

The RT-PCR done on human foetal liver total RNA also gave insight into the importance of the first exon: the sequence of the liver band is missing 31 base pairs compared to the brain sequence. The difference in base pairs is due to an alternate use of the splice site acceptor on *CECR6*b second exon (Figure 3.15), which can be visualised

by the further migration of the liver band compared to the brain band on the RT-PCR gel (Figure 3.14a, panel B).

The *CECR6*b first exon probably is non-coding 5'UTR for two reasons. The translation frame is not conserved between foetal brain and liver, the splice acceptors being 31 bp apart, a number not divisible by three. As well, there are no methionine codons in the ORF1 or ORF2 frames. The ORF2 translated region therefore seems to be within the second exon, which would also be true for the possibility of trORF1 (Figure 3.16c). Nevertheless, there are also no stop codons present in this first exon, which leads me to think that the available *CECR6*b sequence may be the end of a larger transcript, even though clone AK095609 comes from a 5'oligocapping sequencing experiment. This is a possibility due to the very few clones in the database that include the coding sequence of *CECR6* and to the experimental techniques used to confirm the existence of the *CECR6*b splice that did not allow sequence investigation outside of the PCR primer region.

Assuming that the available *CECR6*b mRNA sequence is complete, there are two translation frames possible due to two methionines present in the second exon, one in frame +1 and one in frame +3 (Figure 3.1a). The ORF1 frame +1 leads to 223 amino acids corresponding to the last third of the original ORF1. The ORF2 sequence is unaffected by the *CECR6*b splice, and its putative translation could be advantaged by the absence of the ORF1 initiation codon. This second possibility would make the *CECR6* locus another of the few examples of genes with overlapping ORFs on the same strand.

3.4 The intron does not exist in mouse

TrORF1 is theoretically possible in mouse, the *CECR6* sequence containing many methionine codons (Figure 3.2a). RT-PCR using mouse specific primers (Table 2.1b) in the same area as F1, F2 and R3 in humans lead to a single PCR product that corresponds to the unspliced version of the mRNA (Figure 3.17). This result was confirmed by sequencing confirming that the mouse only produces *CECR6*a mRNAs (Figure 3.18c).

The existence of the *CECR6* intron in human but not mouse can be correlated with the phylogeny results, confirming that ORF2 cannot exist in mouse but is supported in primates. This observation is not refuted by the experiments in this thesis and will be discussed in the next section.

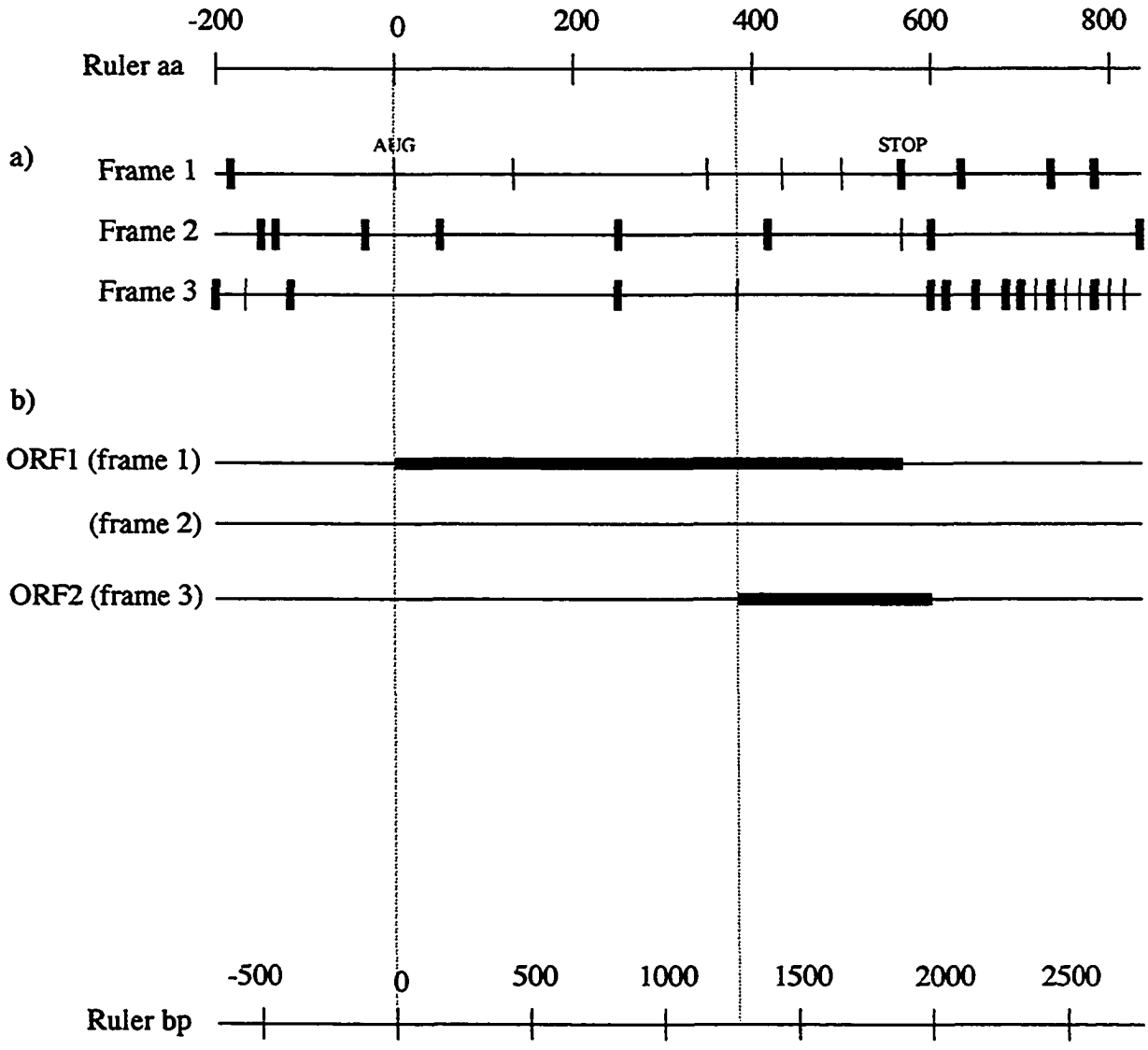


Figure 3.1: Correlation between position of the initiation codons (AUG represented by a thin line) and the termination codons (thicker line) in the three reading frames in the *CECR6* region and predicted ORFs a) Frame 1 and 3 both show a region delimitated by an initiation and a termination codon, b) which lead to the potential ORFs. The rulers' origin (top in amino acid, bottom in base pairs) is set at the initiation codon of ORF1.

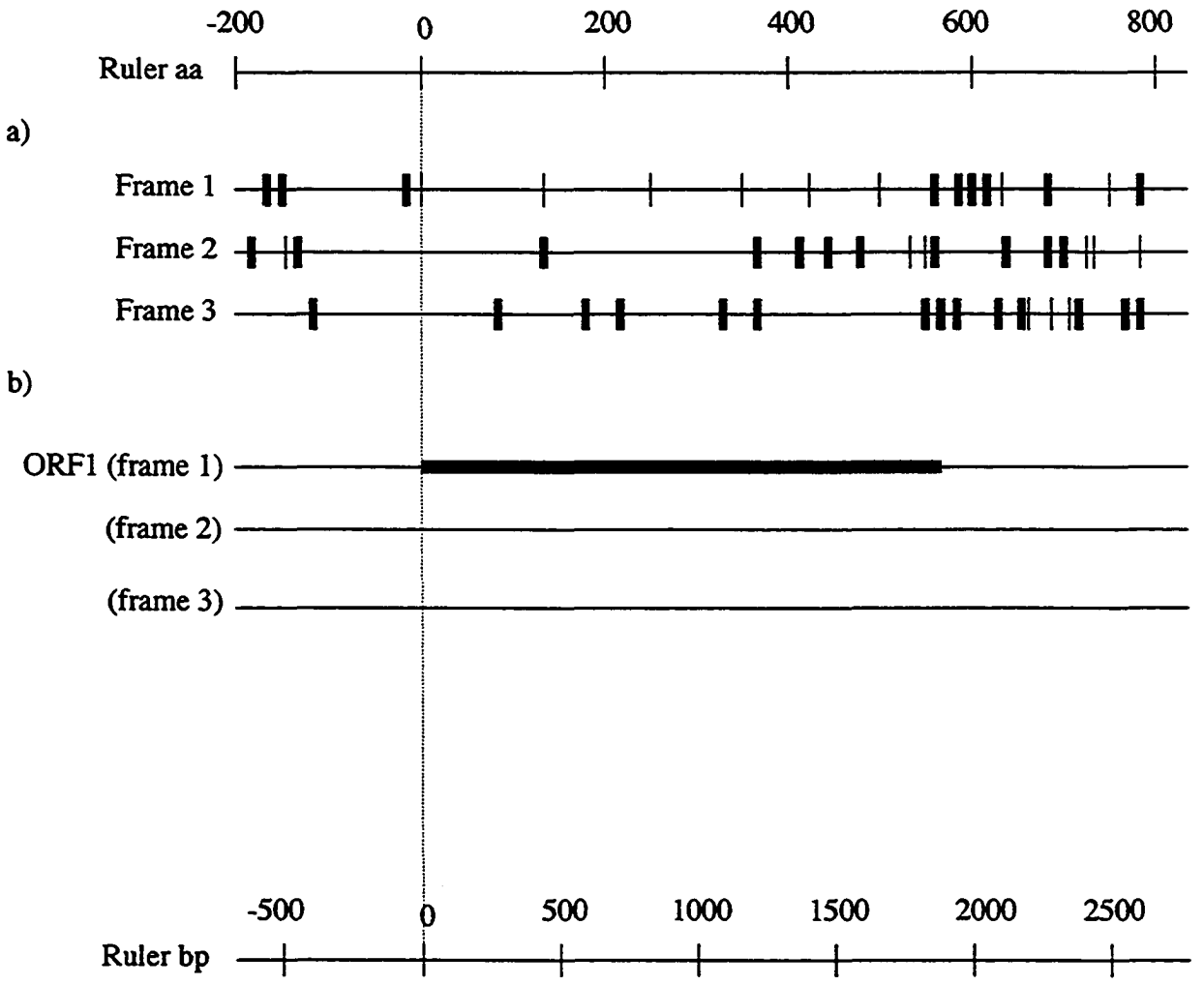


Figure 3.2: Analysis of the mouse *Cecr6* locus DNA and protein sequence. a) Correlation is shown between initiation codons (AUG represented by a thin line) and termination codons (thicker line) in the three reading frames in the *Cecr6* region and b) predicted ORFs. Frame 1 alone shows a region delimitated by an initiation and a termination codon, which lead to the predicted ORF1. Frame 1 also shows many other initiation codons within ORF1 which could lead to a smaller ORF in frame 1. The rulers' origin (top in amino acid, bottom in base pairs) is set at the initiation codon of ORF1.

MRPALGHPRSVSSASGSFPPPPAAARLQPLFLRG
GSFRGRRGSGDSSTSTSTSR**GGGGG**RR**GGGGG**SP
SSSTGAEREDDDESLSVSKPLVPNAALLGPPAQV
GAPAGPAPVAFSSSAATSSSTSTPTSSCSMTAAD
FGGGAAAGAVGGPGSRSAGGAGGTGTGSGASCCP
CCCCCGCPDRPGRRGRRRGCAPSPRCRWGYQALS
VVLLLAQGGLL**D**LYLIAVTDLYWCSWIAT**D**LVVV
VGWAIFFAKNSRGRRGGAASGAHNHHLHHHHAAP
PLHLPAPSAATAGAKARGARGGAGGAGGGLG**AAA**
AAGEFAFAYLAWLIYSIAFT**P**KVVLILGTSILDL
IELRAPFGTTGFRLTMALSV**P**LLYSLV**R**AISEAG
APPGSAGPLLLQPQRHRAAGCFLGTCLDLLDSFT
LVELMLEGRVPLPAHL**R**YLLIAVYFLTLAS**P**VLW
LYELN**AAAAAAA**SWGQASGPGSCSRLLRLLGGCL
VDVPLLALRCLLVVSYQQPLSIFMLKNLFFLGCR
GLEALEGCWDRGNRASPSRARGGYGAPPSA**PPPP**
PPPPQGGSQLGHCISENEGGAHGYVNTLAVASQN

Figure 3.3: Relationship between amino acid repeats (in red) and predicted tm domains (underlined) in human ORF1. There is a negative correlation between tm domain and repeats as no multiple amino acid runs (in bold) and only two multiplets are found in a tm domain. Noticeable residues (single letters in bold) within tm domains are proline (P) that accentuate the helical structure of the region and the positively charged residues arginine (R) and aspartic acid (D) that can render the tm domain sensitive to different polarisation of the cellular membrane.

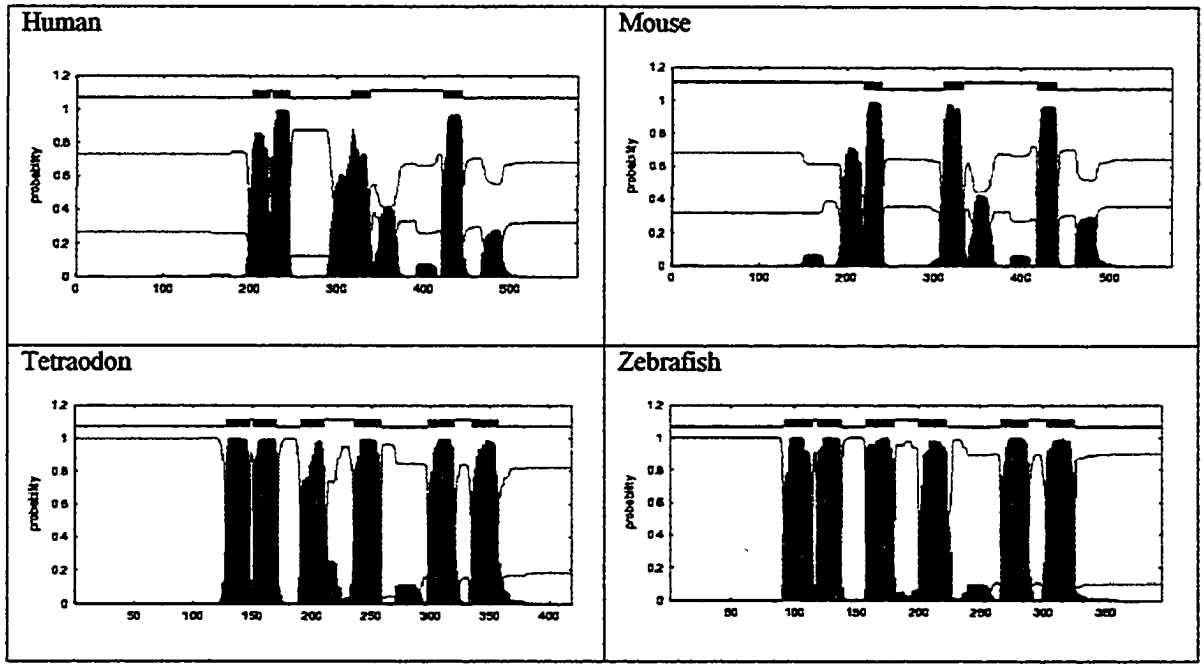


Figure 3.4: Results of the TMHMM v.2 analysis on predicted protein sequences on different mammal and fish species. According to this program, human CECR6a has 4 strong tm domains (probability of 0.8 to 1.0), mouse has 3, and tetraodon and zebrafish show 6. The probability of the amino acid sequence corresponding to a transmembrane domain is represented by the size of the peaks measured on the y-axis while the x-axis indicates the position on the protein. The program also predicts the position of the amino acid loops around the cellular membrane where the blue and pink lines indicate respectively inside and outside regions of the organelle.

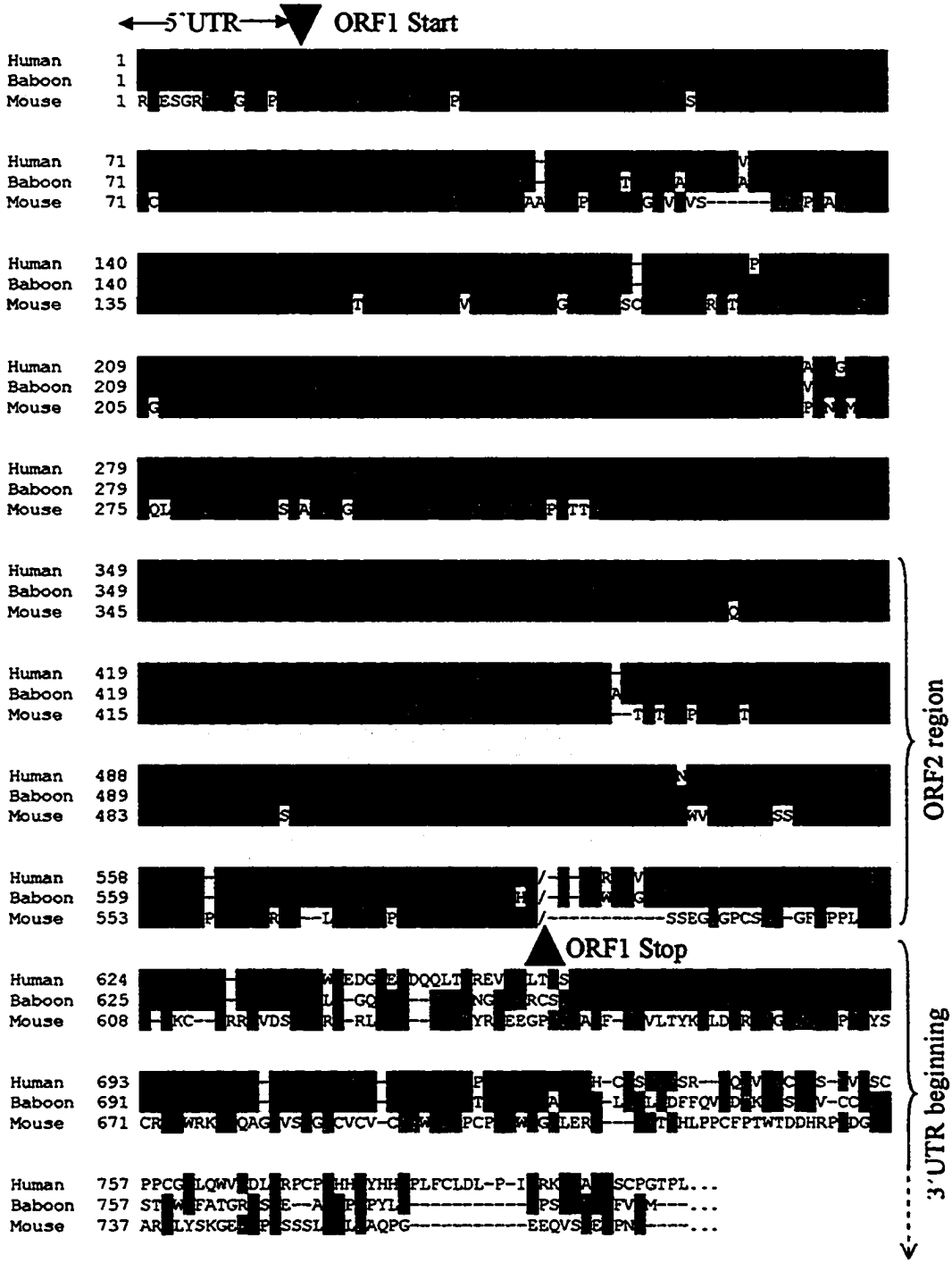
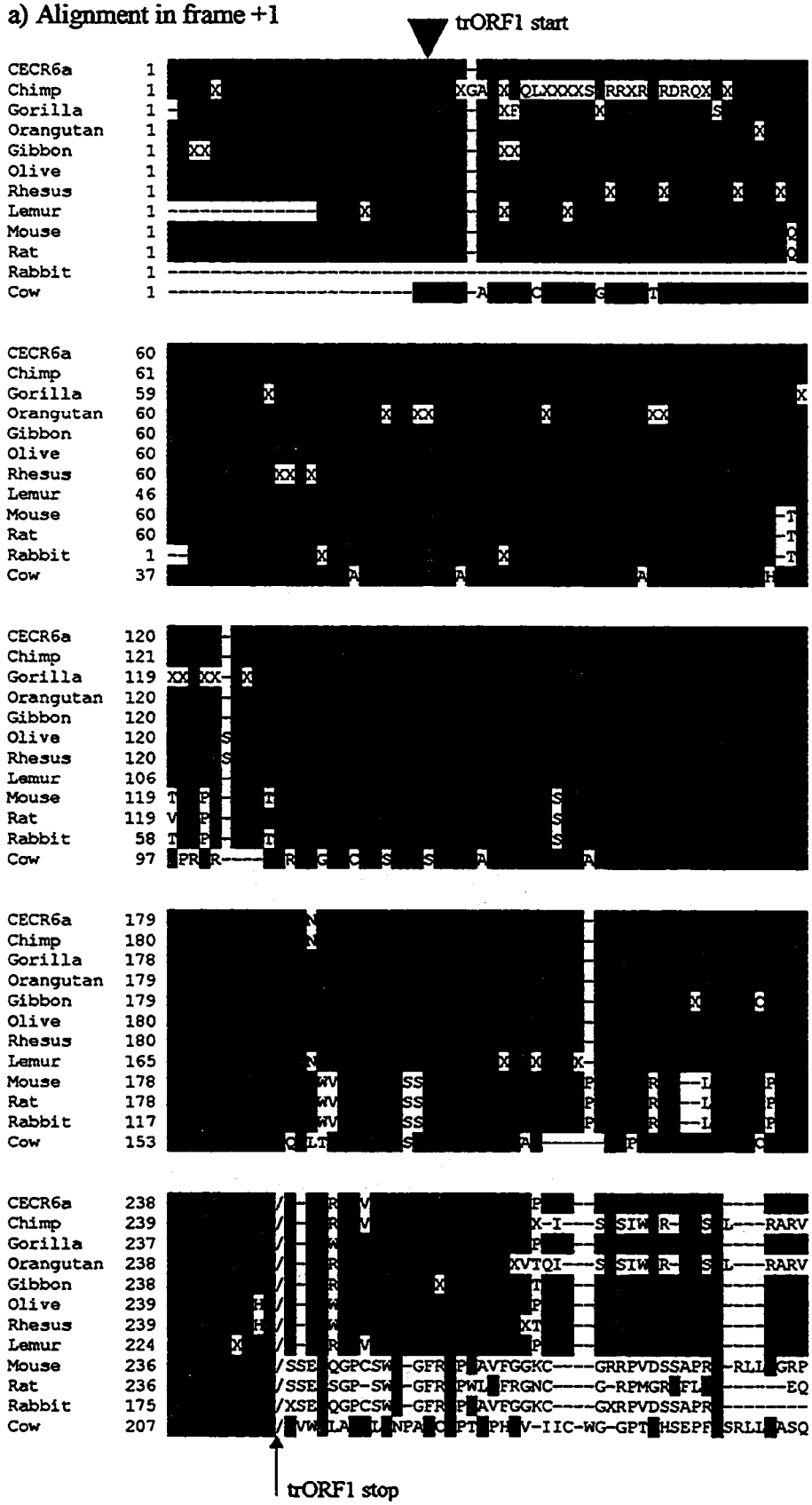
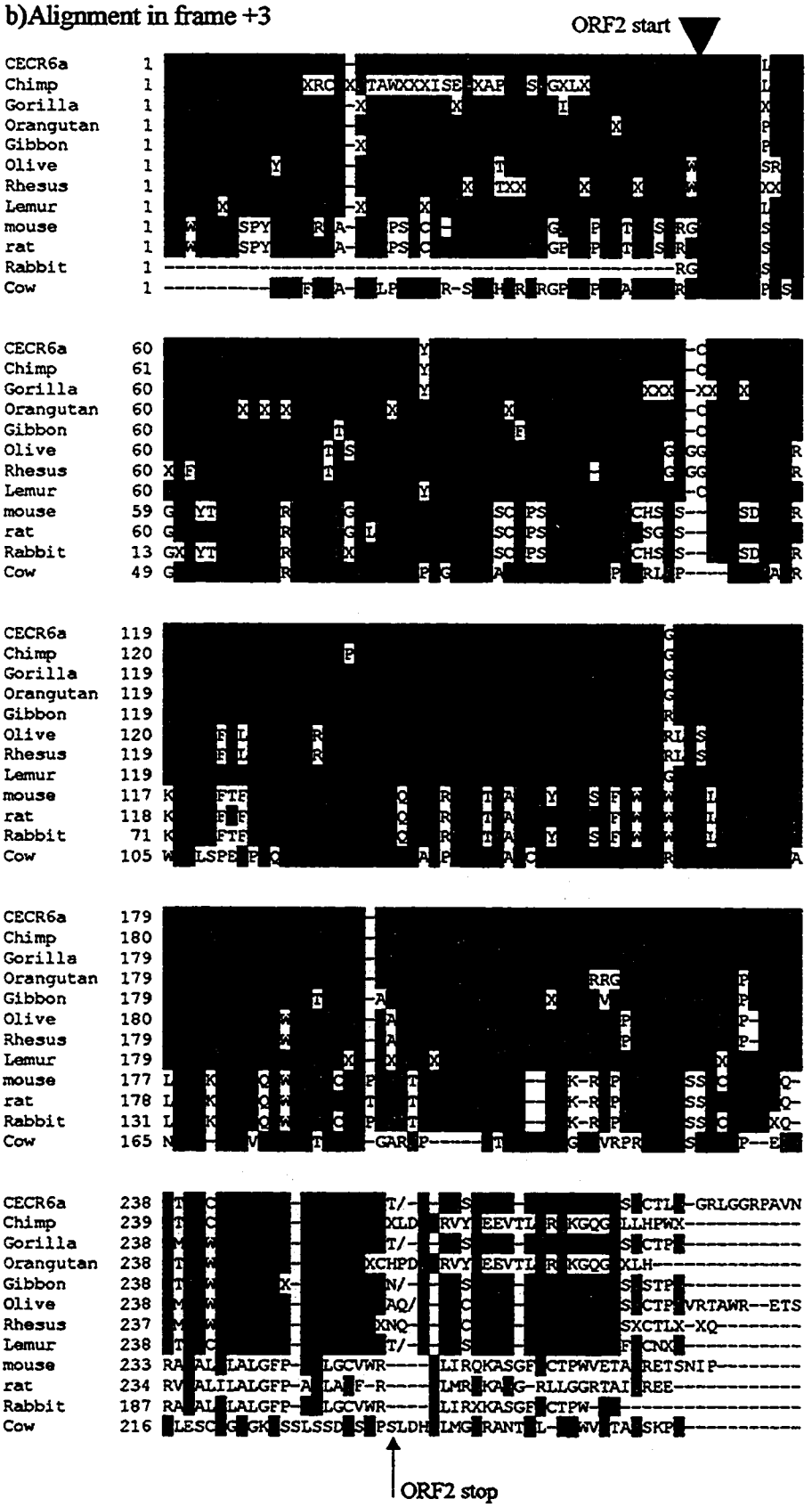


Figure 3.5: Comparison of human, baboon and mouse amino acid sequence in frame +1 of the *CECR6* ORF1 coding region. Conservation between the three sequences is high in the ORF1 region with a total of ten human/baboon mutations and long stretches of human mouse identity (up to about 100 amino acids). The 3'UTR sequences, artificially translated for the purpose of comparing identity, show less similarity than the ORF1 region, the mouse sequence showing no conservation of the frame immediately after the termination codon. Black shading occurs if two out three sequences with identical amino acids, while grey involves similar amino acids.

a) Alignment in frame +1



b)Alignment in frame +3



c) Location of primers on the CECR6 genomic DNA

Figure 3.6: Phylogenetic alignment of the carboxyl end of ORF1 shows conservation within primates and mouse, rat and rabbits. Multiple alignment at the amino acid level of primates, rodents, rabbit and cow sequences a) in frame +1 and b) in frame +3 The sequence are less conserved in both frame past the stop codon. c) Primers used for the PCR and subsequent sequencing flanking the region containing ORF2 in humans and its equivalent in mouse. Primers were made to human or mouse DNA sequence in this region and degenerate PCR was performed on a variety of species. The inverted triangle in a) points to a conserved initiation codon downstream of ORF1 initiation codon that leads to a truncated ORF1 (trORF1, see text) and in b) to the ORF2 initiation codon not conserved in rabbit, cow or rodents. The human, olive baboon, cow, rat and mouse sequences are taken from the NCBI database. Frame change in the last line of the alignment can be attributed to sequence errors in the database (cow, olive baboon) or to the overall quality of the available chimp sequence. The red arrow in all alignments represents the ORF stop codon (/).

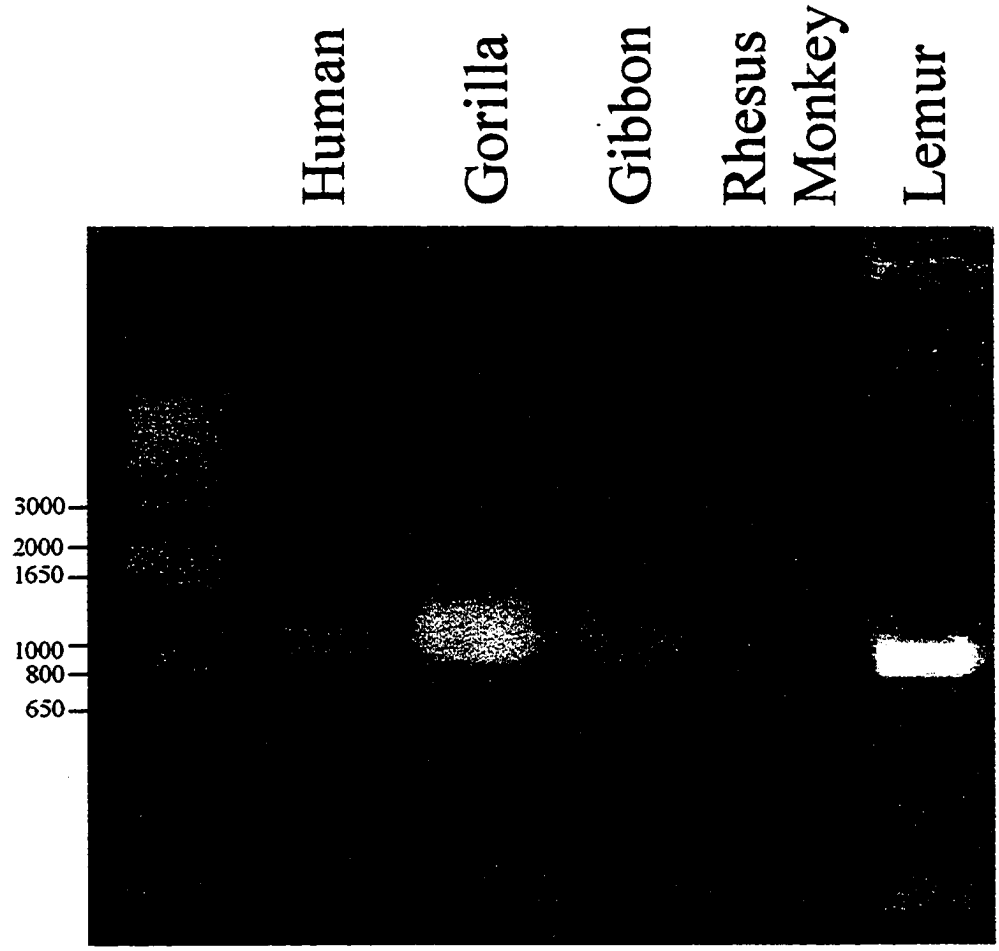
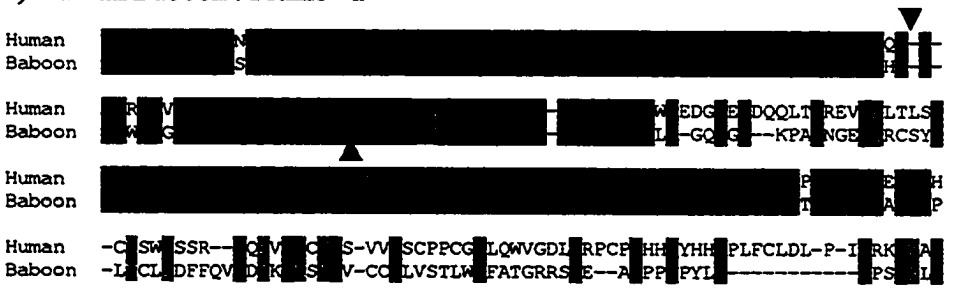
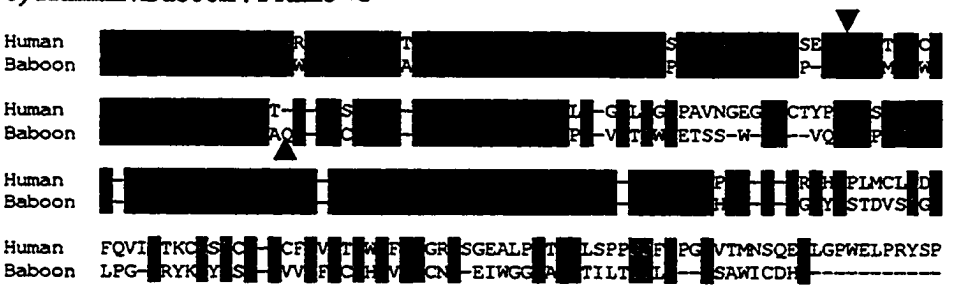


Figure 3.7: Example of various primate amplification product of the CECR6 ORF2 region. Primers used were inferred from regions of similarity between the human and baboon sequences available in the published database. Bands in more than three single lanes from each organism were each sequenced a number of times to build the phylogenetic alignment (Figure 3.6). The human band pattern was used as a positive control for amplification and sequencing.

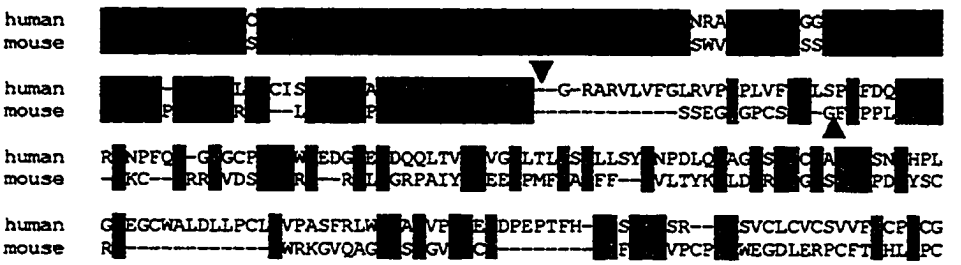
a) Human/Baboon : Frame +1



b) Human /Baboon : Frame +3



c) Human/Mouse : Frame +1



d) Human/Mouse : Frame +3

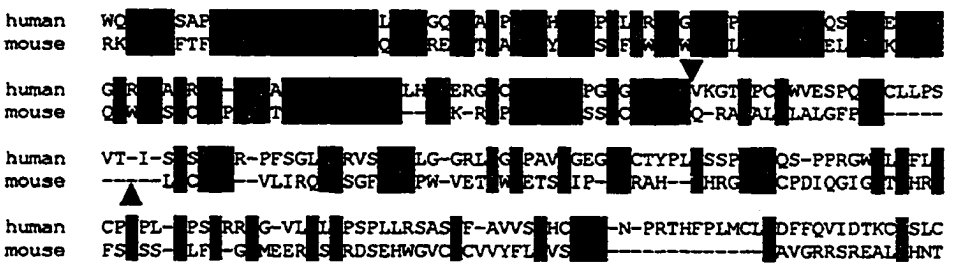


Figure 3.8: Comparison of the carboxylic end of ORF1 (frame +1) and ORF2 (+3). The downward triangle indicates the end of the ORF1 predicted protein while the upward triangle, the ORF2 stop is 23 amino acids away. The position of an ORF stop outside of its frame (ORF1 stop in frame +3 for example) points to the amino acid using the closest nucleotides. a) b) The region unique to ORF2 (between the two triangles) is conserved between human and baboon in both frames with baboon ORF2 using a different termination codon. c) d) Mouse sequences in both frames show no sequence conservation past the ORF1 stop as well as no potential termination codon for ORF2 in frame +3. Alignment containing baboon or mouse was done separately to allow shading (black for identity, grey for similarity) and frame shifts to be viewed easily.

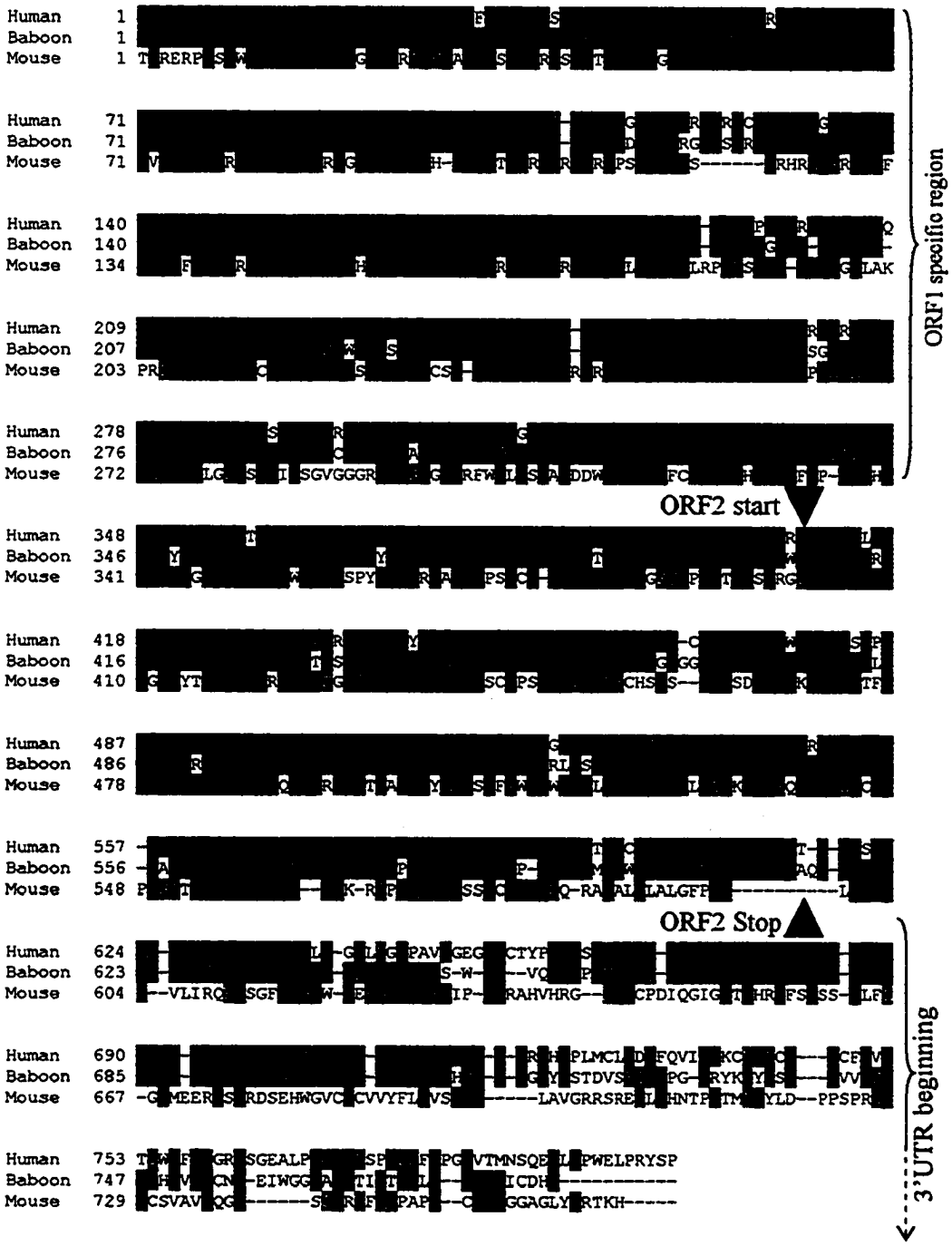


Figure 3.9: Comparison of human versus mouse amino acid sequence in frame +3 in the *CECR6* ORF2 coding region. More mutation between human, baboon and mouse can be seen in this frame than in frame +1 (Figure 3.5), approximately one every ten amino acids for human/mouse. The mutation rate does not seem to change in the ORF2 region. In fact, the baboon sequence shows less mutations (15) in the first two thirds of the sequence, corresponding to ORF1 alone, than in the ORF2 last third (18). The Boxshade program parameter was set at 0.5 thus two of the three sequences have to match in order to be shaded. This causes the third line (mouse) to have less shading (match) than the two others, especially in the 3'UTR, but does not allow the mouse sequence to have an influence on the human/baboon shading.

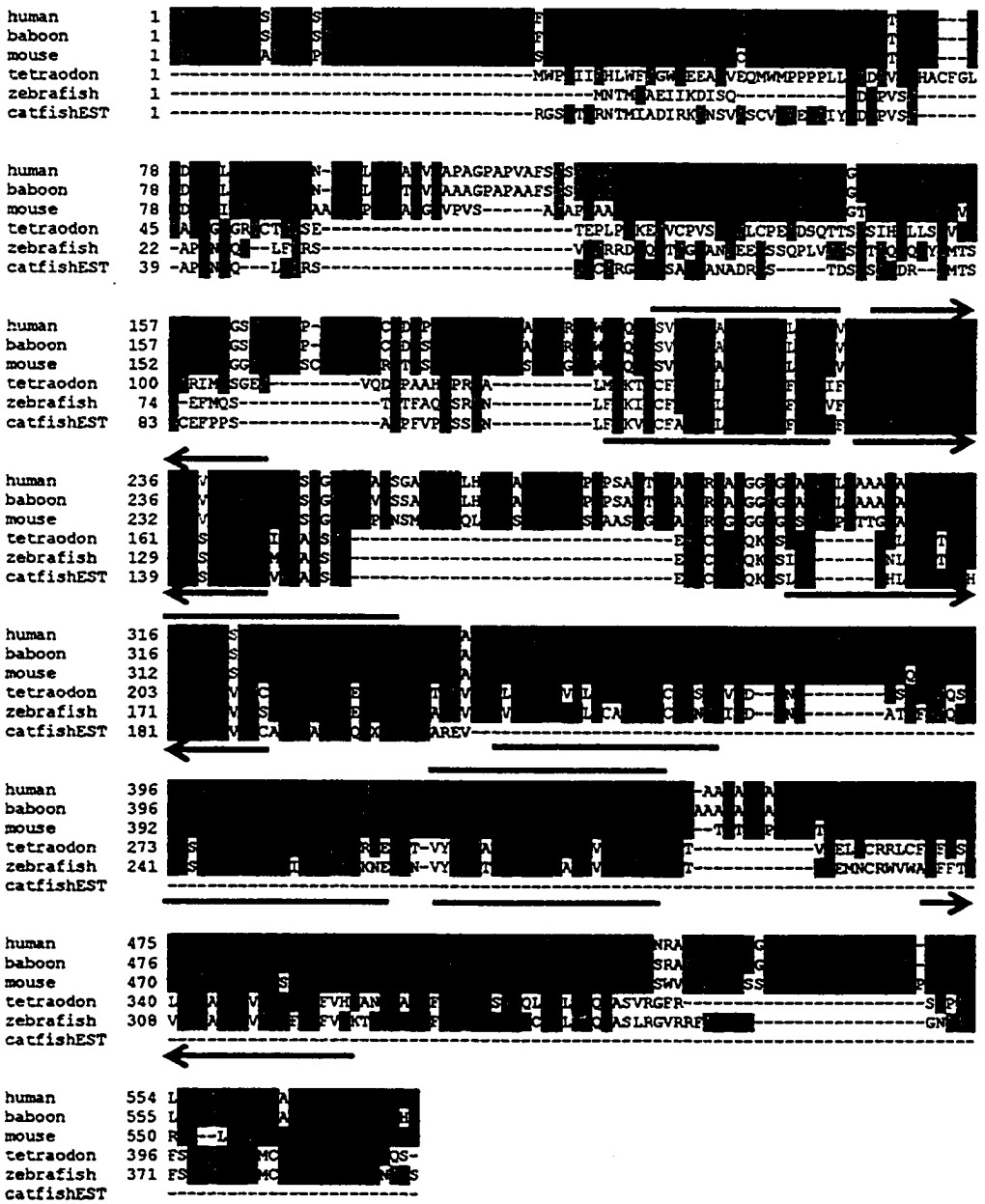
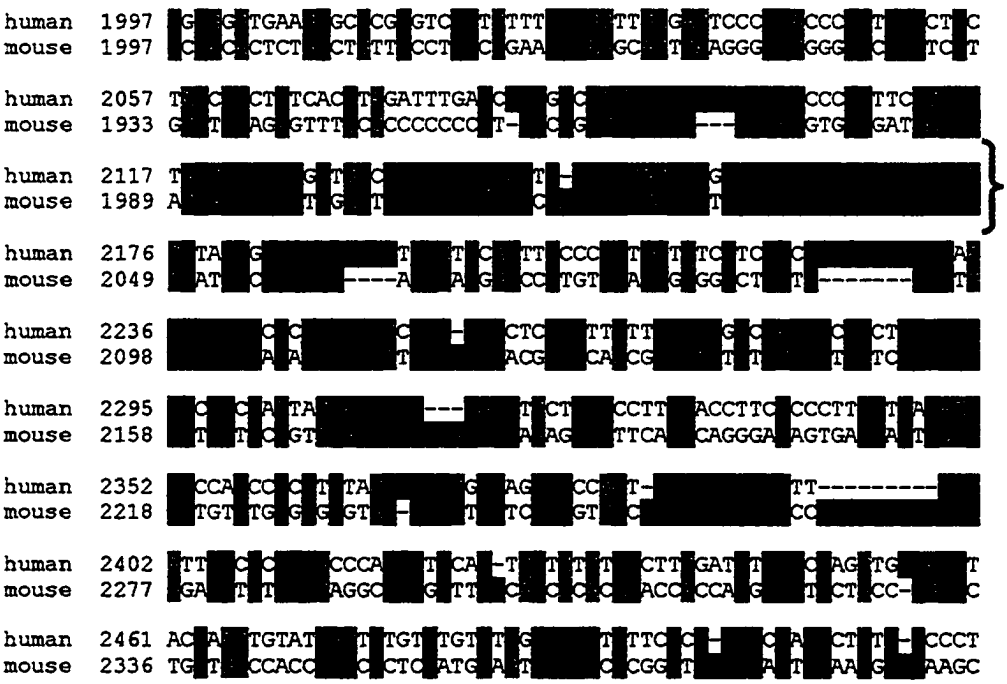
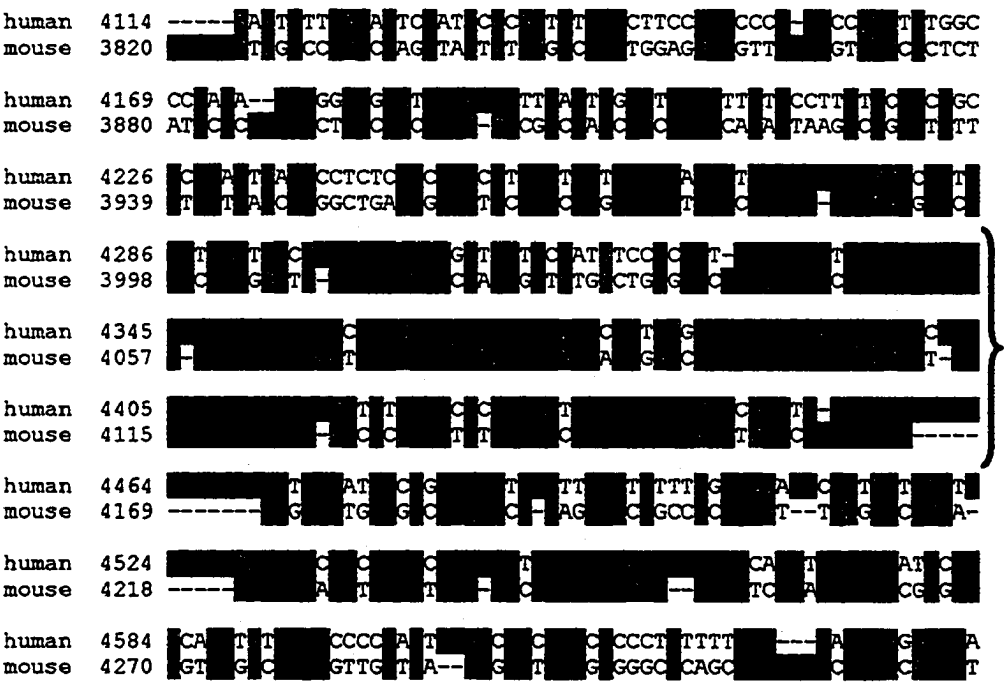


Figure 3.10: Multiple alignment of ORF1 predicted amino acid sequence in mammalian and fish species. Four predicted mammalian tm domains are illustrated by a blue line over the sequence alignment and the six fish tm domains, by a red line under, with individual sequence prediction following the same colour pattern. These tm regions as well as the region underlined in green show higher conservation, than the rest of the protein, up to 95% across species. This conservation of a sequence between tms could indicate a functional domain. Sequences used are found on the NCBI database. The incomplete catfish EST is the only available catfish sequence of the CECR6 locus, and is predicted to have all three tm domains its sequence includes, by the TMHMM v.2 program. No splice versions of the fish CECR6 mRNA are found in the database. Arrows indicate that the tm domain continues to the next line.

a)



b)



c)

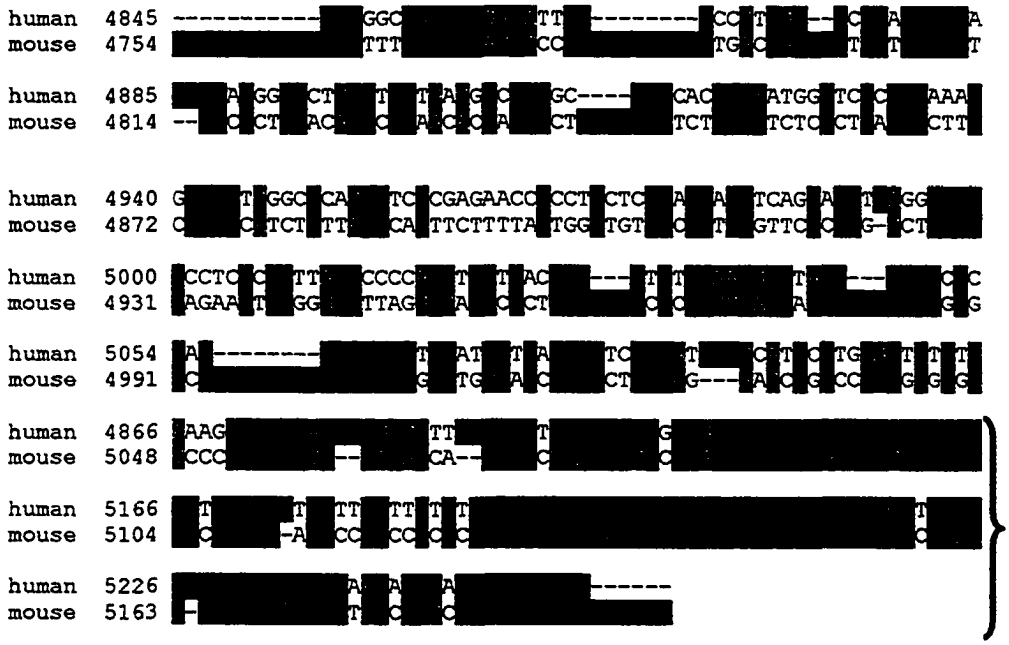


Figure 3.11: Human and mouse sequence comparison of the three conserved regions in the 3'UTR of the CECR6 mRNA. The three regions identified by the brackets are between 50 and 100 bp long and show 80% identity. The regions are spread out through the mRNA: a) a region directly after the two ORFs, b) a region near the 3' end of the mRNA and c) the polyadenylation signal region. The regions were initially identified by comparing the human and mouse published mRNA sequences using the program BLAST2SEQ (Tatusova and Madden 1999). Alignments were done in three parts using Clustal W: the conserved sequence and the regions upstream and downstream were aligned separately to circumvent the program's tendency to provide the most spread out alignment. Aligning the whole sequence at once would not have revealed the conserved regions. The three parts were put together before using Boxshade to shade conserved nucleotides. Numbers next to the sequence correspond to the nucleotide position in the human or mouse *CECR6* mRNA.

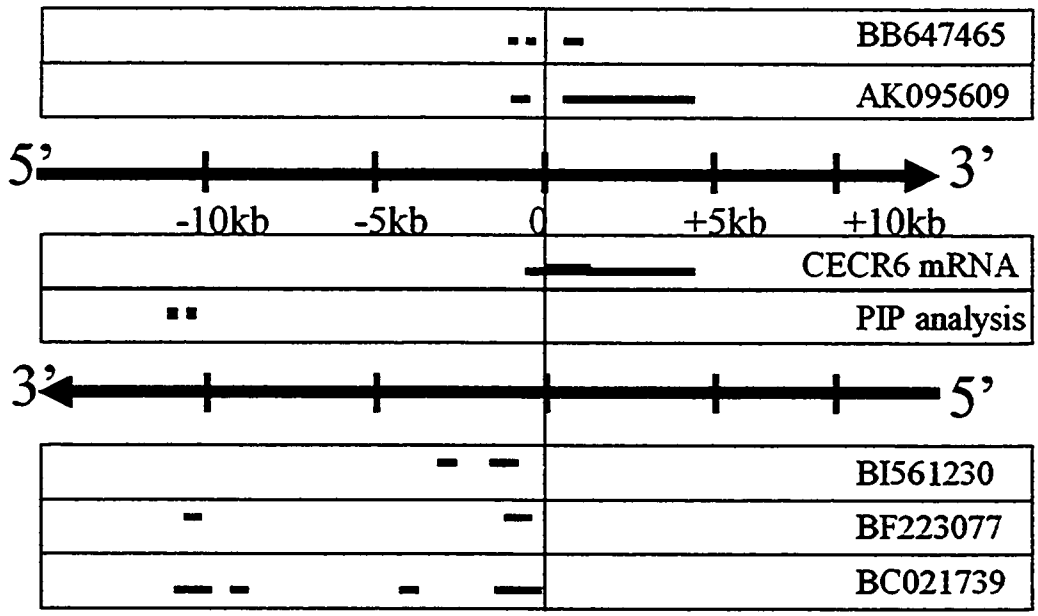


Figure 3.12: Representation of the unexpected CECR6 EST clones in the NCBI database. Clones above the 5'3' arrow include the RIKEN clone (BB647465 in green) without proper GT/AG splice junctions and the clone showing the proper CECR6b splice (AK095609 in pink). Clones below the 3'5' reverse arrow include the clone on the opposite DNA strand from CECR6 and do not show proper splice junctions. Each clone's approximate position (coloured lines) can be compared to the beginning of ORF1 (in blue, between the arrows) on the arrows' DNA ladder. The PIP analysis (Figure 3.13) predicted a sequence similarity around –10 kb represented in dark blue.

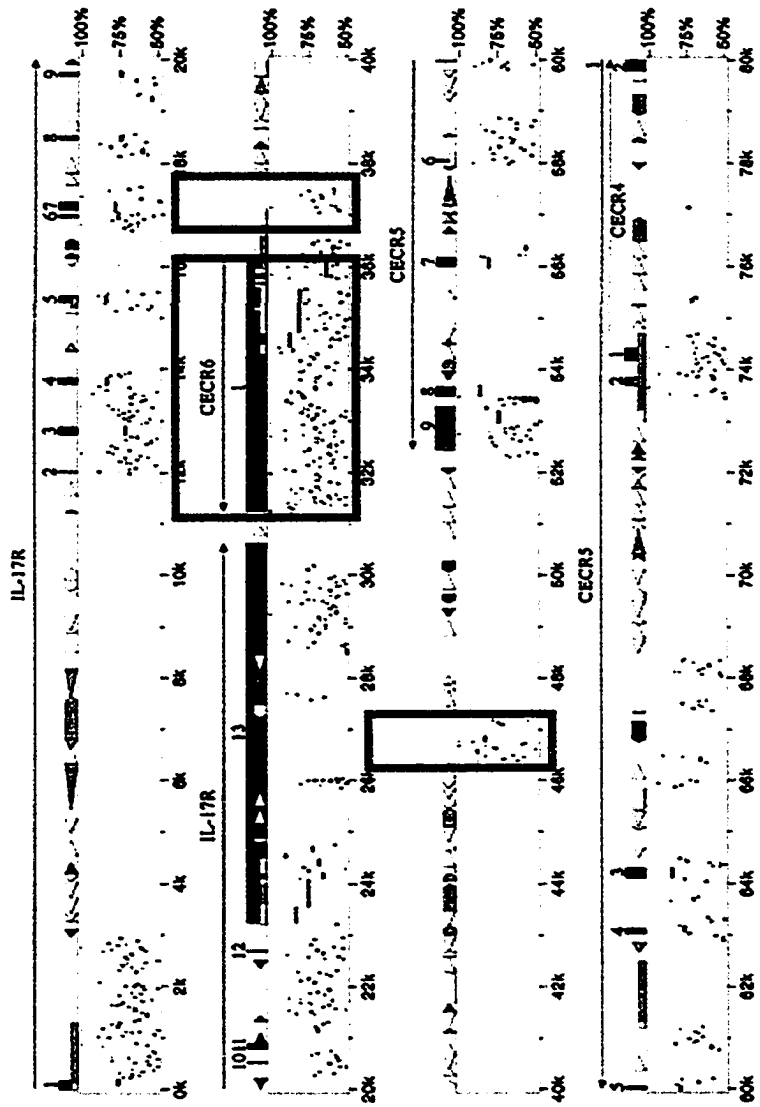
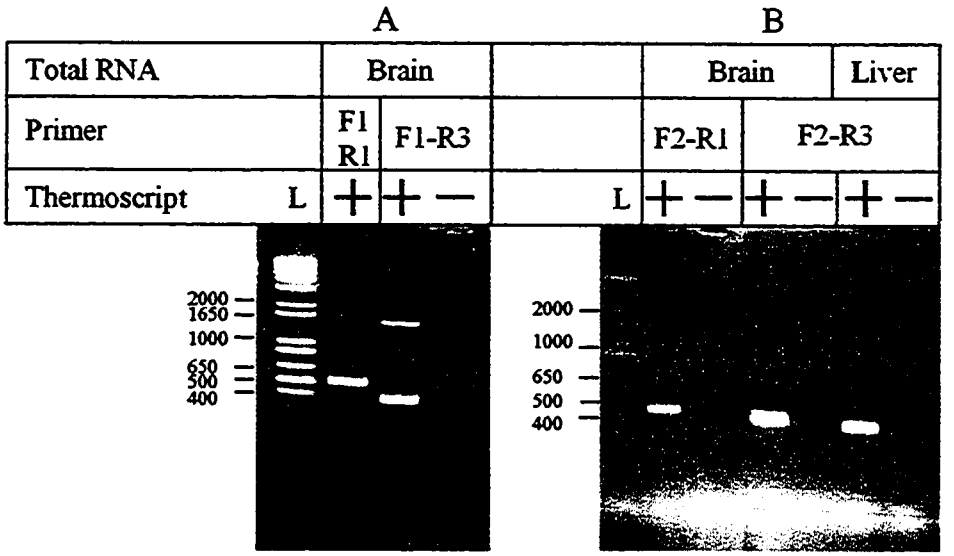


Figure 3.13: Similarity analysis between the human and mouse homologous regions surrounding the *CECR6* locus. The red boxes indicate regions of conservation potentially related to the CECR6 locus. The thicker black lines represent coding regions of the CES Critical Region DNA. CECR6 is flanked by IL-17R and CECR5. DNA scale in kb starts at IL-17R exon 1. Regions of similarity between humans and mouse are indicated by dashes under the sequence (scale of 50 to 100%). Adapted from Footz *et al*, 2001.

a)



b)

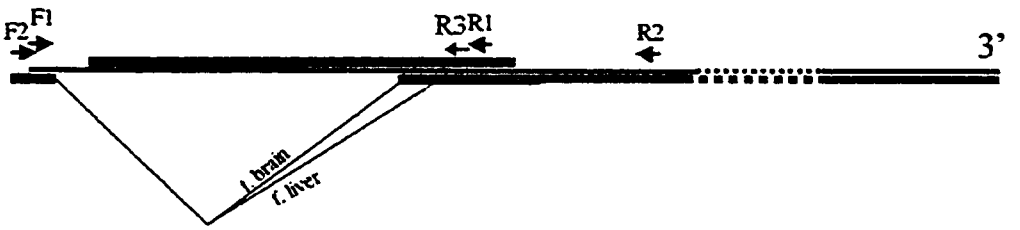


Figure 3.14: The CECR6 mRNA can be spliced. Reverse transcription PCR a) using the forward primer F1 (panel A) leads to two bands corresponding to an unspliced version (*CECR6a*) and a spliced mRNA (*CECR6b*) in human foetal brain total RNA, which was confirmed by sequencing. RT-PCR using the forward primer F2, 50 bp upstream of F1 does not yield the unspliced version (larger band) of the *CECR6* mRNA (panel B). Reverse primer R1 gives bands 100 bp bigger than with reverse primer R3. Bands smaller than the expected 500 bp in the first lane and the 800 bp in the second lane did not correspond to any known sequences. b) Further sequencing of both human foetal brain and liver bands showed a 31 base pairs difference due to alternative splicing (black lines). The *CECR6a* mRNA (in blue) starts between F1 and F2. ORF1 is represented in green and *CECR6b* in red.

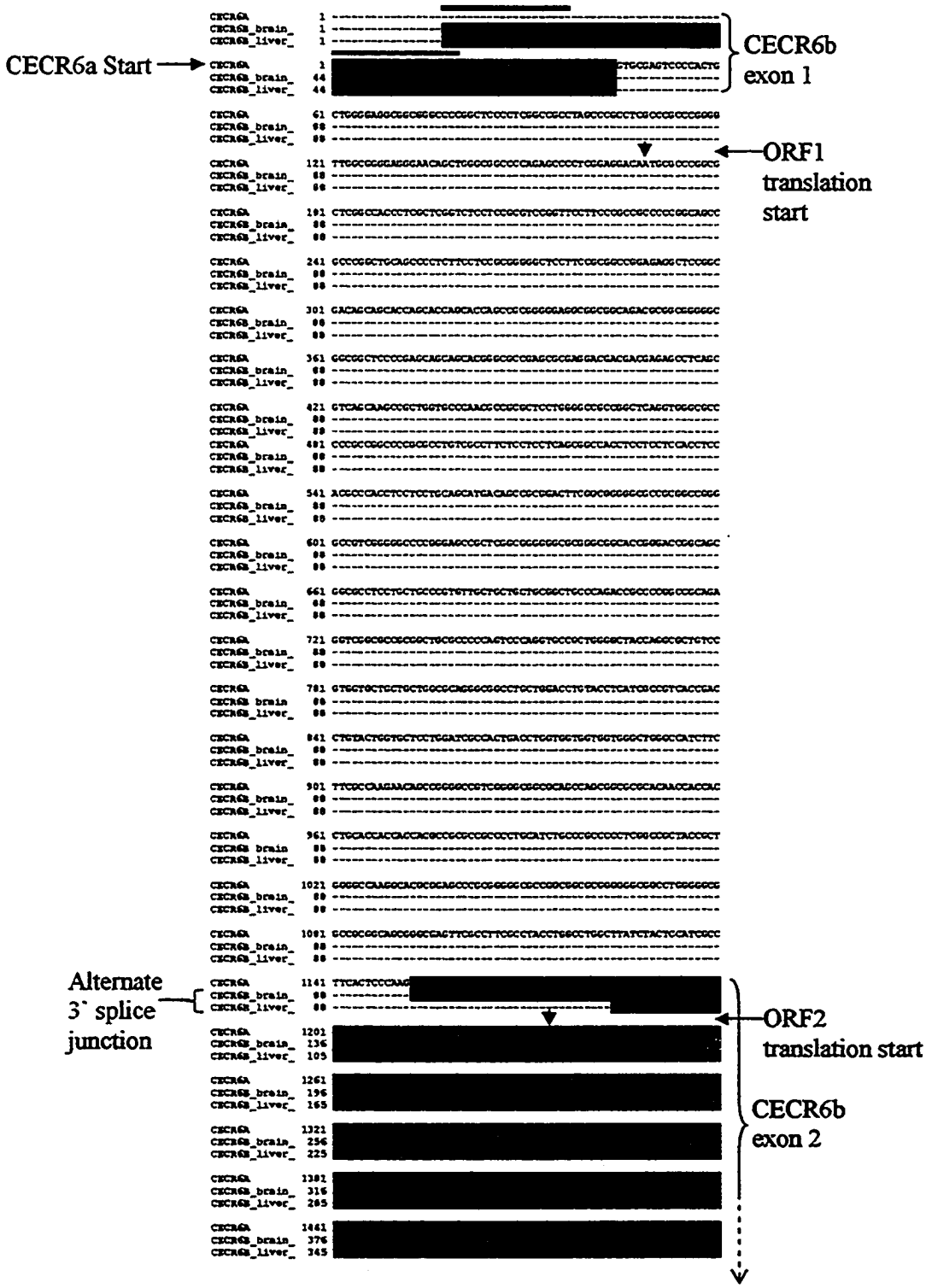


Figure 3.15: Nucleic acid alignment of the CECR6a and CECR6b mRNA. The known CECR6a mRNA starts 44 bases downstream of the known CECR6b mRNA. CECR6b shows the same 5' splice junction but a different 3' splice junction specific to either fetal brain or fetal liver mRNA. The two versions of the spliced intron follow the GT donor-AG acceptor splice DNA sequence consensus. The 3' end of the coding region and the 3'UTR has been omitted for legibility but would show a perfect alignment for all three mRNA variants. The black line represents F2 primer and in grey, F1.

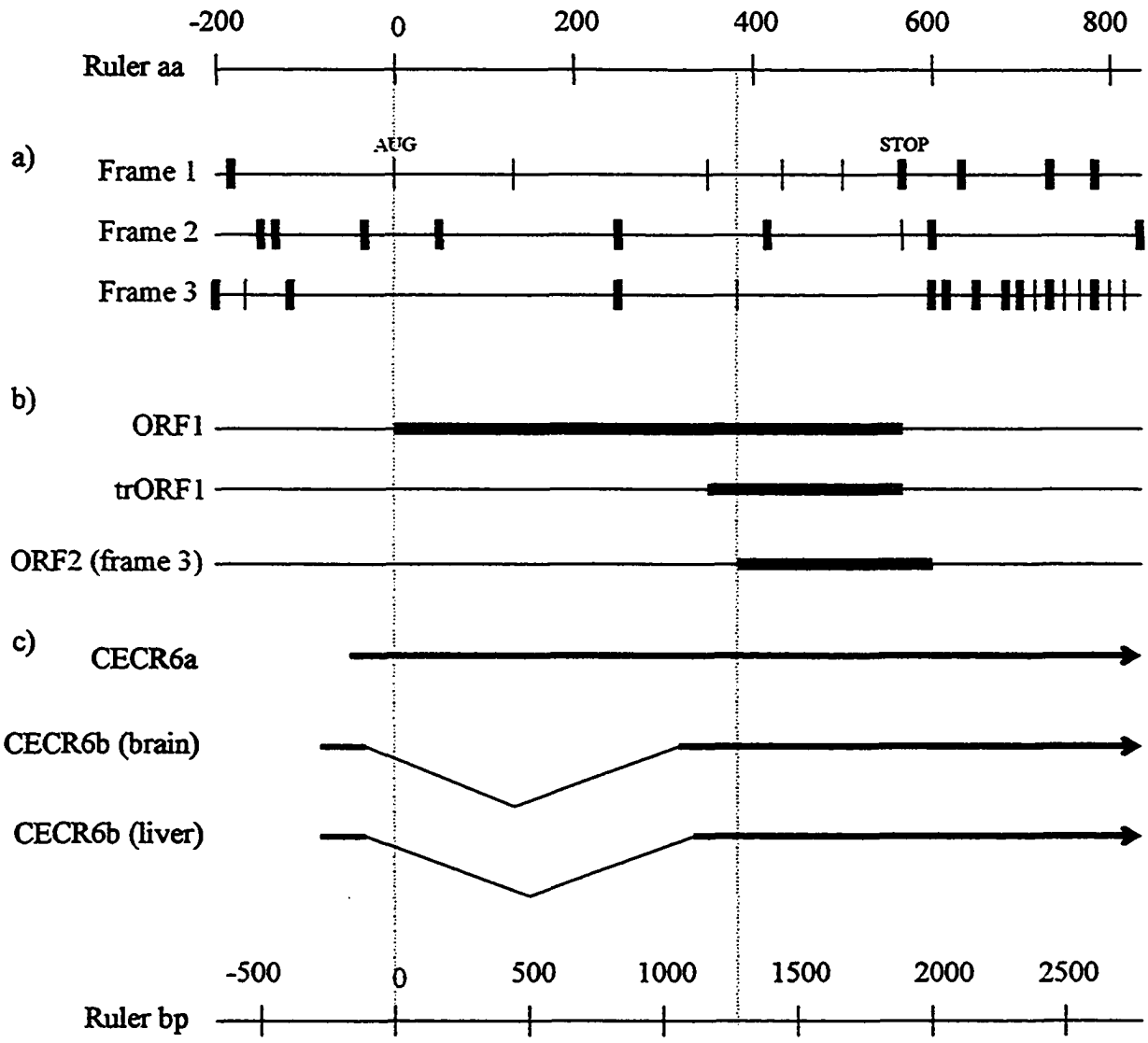


Figure 3.16: ORF1 is not present on *CECR6b* a) The position of the initiation and termination codons lead to b) two potential ORFs in the region span by *CECR6b*: ORF2 and a truncated version of ORF1: trORF1. c) Comparison between the two available mRNA sequences shows that the *CECR6b* mRNA starts upstream of *CECR6a*. Neither trORF1 nor ORF2 show potential coding sequence in the region upstream of the splice site on the *CECR6b* mRNA. The rulers' origin (top in amino acid, bottom in base pairs) is set at the initiation codon of ORF1.

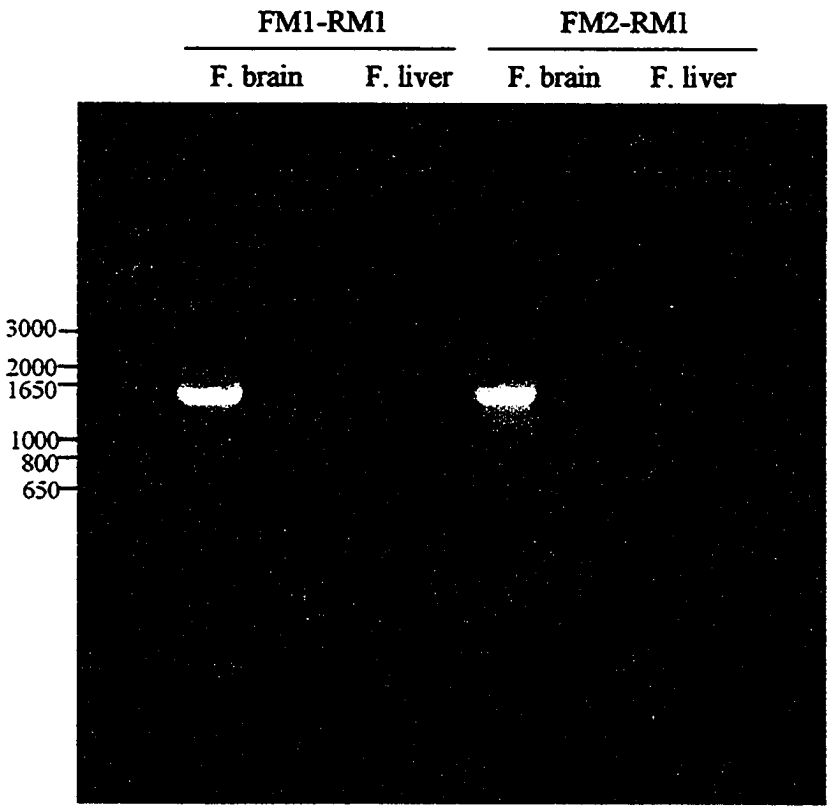


Figure 3.17: RT-PCR on mouse total RNA does not show a spliced version of the CECR6 mRNA. Two different forward primers in regions comparable to human F1 and F2 primers were used on mouse foetal brain and foetal liver total RNA. A strong band is seen in foetal brain around 1600 bp, using both forward primers and a faint band is seen using FM1-RM1 foetal liver but this last band could not be confirmed by sequencing. The smaller band running around 800 bp in the first foetal brain lane is too large to correspond to the splice version of CECR6b (500 bp) and sequencing confirmed it was a PCR artefact, similar to that in Figure 3.14. RT-PCR reactions used as controls lacked thermoscript and are indicated by a minus sign while positive results were expected on lanes marked with a plus sign. DNA contamination was assessed in a PCR containing no RT product (H_2O lane).

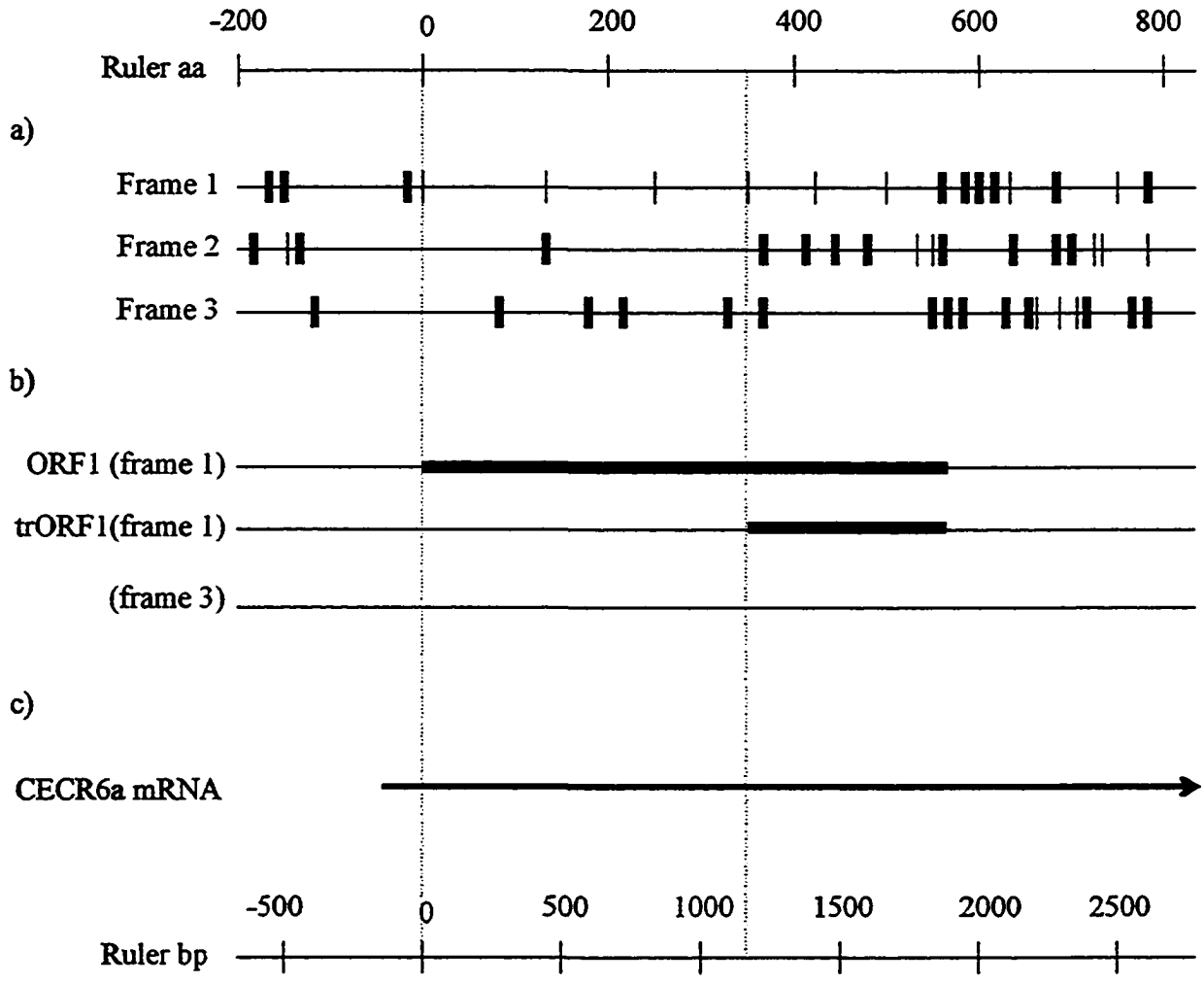


Figure 3.18: Analysis of the mouse *Cecr6* locus DNA, protein sequence and mRNA.
a) Correlation is shown between initiation codons (AUG represented by a thin line) and termination codons (thicker line) in the three reading frames in the *Cecr6* region, b) predicted ORFs and c) available mRNA sequence. The only *CECR6* mRNA found in mouse is unspliced and supports ORF1 and a truncated ORF corresponding to the human trORF1. The rulers' origin (top in amino acid, bottom in base pairs) is set at the initiation codon of ORF1.

Amino Acid	3 letter abbreviation	1 letter abbreviation	RNA codon	Property
Alanine	Ala	A	GCA,GCC,GCG,GCU	Hphobic
Cysteine	Cys	C	UGC,UGU	Hphobic
Aspartic Acid	Asp	D	GAC,GAU	-
Glutamic Acid	Glu	E	GAA,GAG	-
Phenylalanine	Phe	F	UUC,UUU	Hphobic
Glycine	Gly	G	GGA,GGC,GGG,GGU	PolarU
Histidine	His	H	CAC,CAU	PolarU
Isoleucine	Ile	I	AUA,AUC,AUU	Hphobic
Lysine	Lys	K	AAA,AAG	+
Leucine	Leu	L	UUA,UUG,CUA,CUC,CUG, CUU	Hphobic
Methionine	Met	M	AUG	Hphobic
Asparagine	Asn	N	AAC,AAU	PolarU
Proline	Pro	P	CCA,CCC,CCG,CCU	PolarU
Glutamine	Gln	Q	CAA,CAG	PolarU
Arginine	Arg	R	CGA,CGC,CGG,CGU,AGA, AGG	+
Serine	Ser	S	UCA,UCC,UCG,UCU,AGC, AGU	PolarU
Threonine	Thr	T	ACA,ACC,ACG,ACU	PolarU
Valine	Val	V	GUA,GUC,GUG,GUU	Hphobic
Tryptophan	Trp	W	UGG	Hphobic
Tyrosine	Tyr	Y	UAC,UAU	PolarU
STOP		-	UAA,UAG,UGA	

Table 3.1: The nucleic acid codon universal code includes 20 different amino acids and three Stop codons. The 64 possible nucleotide triplets show redundancy, with more than one triplet corresponding to one amino acid. Each amino acids can be abbreviated using a one or three letter code and shows one of four chemical property: hydrophobicity (Hphobic), polar uncharge (PolarU), or charged acidic (-) or basic (+).

	Position									
	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4
Kozak	**G**	**C**	**C**	**R**	**C**	**C**	**A**	**T**	**G**	**G**
ORF1	A	G	G	**A**	**C**	A	**A**	**T**	**G**	C
trORF1	C	T	**C**	**A**	**C**	**C**	**A**	**T**	**G**	**G**
Met #2	**G**	A	G	C	T	G	**A**	**T**	**G**	C
Met #3	A	T	**C**	T	T	**C**	**A**	**T**	**G**	C
ORF2	**G**	G	**C**	C	G	G	**A**	**T**	**G**	C

Table 3.2: Comparison of the region surrounding ORF1 and ORF2 initiation codons to the Kozak sequence (in bold). The nucleotide positions for ORF1 and ORF2 in bold correspond to the Kozak sequence for optimum initiation, the most important being positions –3 and +4. While ORF1 and ORF2 sequences shows only two similarities (not counting the initiation codon ATG), trORF1 shows five including the two most important positions. The two other methionine codons (labelled #2 and #3) present in frame +1 on the *CECR6b* mRNA have very weak Kozak sequences (one or two matches respectively) and lead to very small ORFs (166 and 79 amino acids respectively).

CHAPTER 4: DISCUSSION

The overall objective of my thesis has been to determine if the *CECR6* locus could be the fifth example in the human genome of overlapping genes on the same strand using alternative reading frames of a coding region. Using available computer programs, phylogenetic analysis and experiments at the RNA level, I have reached the conclusion that the evidence compiled in my research support the existence of two overlapping ORFs in primates but not in mouse or fish.

1. *CECR6*: a very long 3'UTR

The *CECR6* locus showed interesting characteristics even without the presence of the second ORF. The *CECR6* mRNA 3'UTR is estimated in the top 5% for length with 3097 base pairs in the human sequence and 3113 bp in the mouse sequence. In the UTRdb (database, http://bighost.area.ba.cnr.it/BIG/UTRHome/), the 3'UTR can vary greatly in length between 20 to 8500 bp with an average of 1027 bp in human UTR sequences (Grzybowska, Wilczynska *et al.* 2001). The above average length of the *CECR6* 3'UTR in mice and in humans, and the three small (60bp) conserved domains suggest that it has a function. Sequence analysis of the three conserved regions failed to provide information on the regulation or subcellular localisation of the *CECR6* mRNA. The role of these domains might be deduced from subcellular localisation experiments of the *CECR6* mRNA as 3'UTR motifs are usually responsible for localisation or regulation. Intracellular localisation of the CECR6 mRNA in mouse or primate cell cultures using a

single strand probe could be disrupted by introducing vectors with 3'UTR sequence containing mutations in each of the three domains.

2. ORF1: a number of significant multiple amino acid runs

CECR6 shows multiple repeated amino acids that can be divided in two groups. Statistically significant amino acid repeats (at least 5 of the same amino acid in a row for a 400 amino acid long protein) are termed amino acid runs (Karlin, Brocchieri *et al.* 2002). Notation for amino acid repeats is the single letter amino acid code followed by the number of repeats in subscript (P_8 for example is the notation for nine prolines repeated in a row). The *CECR6* amino acid run sequence in humans is $G_5G_5C_5A_5A_7P_8$. Repeats shorter than the significant number of amino acids are termed multiplets. The multiplet sequence of *CECR6* is $P_4A_3S_3D_3S_3S_3G_3A_3R_3L_3V_4H_4G_3L_3$. Additional information on individual amino acids is available in a table of the different amino acids with their abbreviations and characteristics (Table 3.1).

According to a proteomic analysis completed on five complete eukaryotic genomes focussing on multiple amino acid runs (Karlin, Brocchieri *et al.* 2002), multiple amino acid runs in specific eukaryotic genes are not present in prokaryote homologues but could correspond to "recent evolution linked to complex evolution of heart and brain development". In light of this vague statement, I believe that the correspondence between a higher brain function or a heart with more chambers with repeated amino acids could simply be due to an evolutionary correlation without interdependence. The positive selection for amino acid repeats is found in ORF1. The different fish *CECR6* sequences available do not show conservation of any of the multiple amino acid runs nor the

multiplets discussed below. The only multiplet that is found in the tetraodon *CECR6* ORF1 sequence is a proline P_4 repeat that aligns with the second human glycine G_5 run. Those two amino acids are chemically very different and the significance of this multiplet is probably very low. In the human genome, 20% of all genes contain at least one significant run, most predominantly glutamate (19.8%), proline (18.2%) and leucine (19%), but only 1.9% had more then one run. This last number increases to 7% in Drosophila. Cysteine was one of the least represented with only 0.4% frequency of runs in the human genome. Other hydrophobic amino acids (isoleucine, valine, methionine, phenylalanine, and tryptophan) also exhibit amino acid runs at very low probability, with the exception of leucine (19%) and alanine (16.9%). ORF1 shows some unusual hydrophobic repeats such as a polycystidine and a polyvaline repeat. Alanine in high frequency leads to alpha-helix stability and flexible hydrophobic properties. That ORF1 has two of these runs located before the third and after the fourth tm strongly supports tm domains in humans.

Multiple amino acid runs are selected for by evolution and therefore must have a specific function (Karlin, Brocchieri *et al.* 2002). Codon usage in fact varies within the run which discredits the idea of small duplications or "strand slippage" in *CECR6*. Frequency and length of the amino acid runs depend on their property. Polar uncharged (glycine, histidine, asparagine, proline, glutamine, serine, threonine and tyrosine) and some hydrophobic amino acids are far more prevalent in human proteins than charged residues (aspartate and glutamate both acidic (-), and lysine and arginine both basic (+), 65 to 80% versus 2 to 7%). ORF1 amino acid repeats have representatives from each category and shows less charged amino acid repeats (D_3R_3 are the only charged

multiplets). Known polar uncharged amino acid runs are also ten times longer than those of all other amino acids, with a maximum length measured at 206 identical amino acids in a row (neither the amino acid nor the protein was indicated). ORF1 amino acid runs are short by comparison but numerous, which is not common. Other unusual sequence properties associated with multiple amino acid runs are charge clusters, alternative charge runs, histidine patterns and multiplets. The ORF1 sequence only contains the latter: an unusually large number of multiplets. Combining the amino acid runs and the multiplets reveals 20 amino acid stretches of 3 or more repeats. This does not include regions rich in certain amino acids like glycine (11 in a 22 amino acid sequence) or leucine (10 of 21) and amino acid repeats such as RPGRRGRRRG or TSSSTSTPTSSCS.

Multiple amino acid repeats seem to be common in disease genes. More than 40% of the 192 genes containing one stretch or more of repeated amino acids (2% of the database searched in 2002) are associated with diseases (Karlin, Chen *et al.* 2002). The authors mention syndrome related protein(s) in the CES Critical Region with many multiplets, which presumably is *CECR6*. Most commonly known are polyglutamine diseases like Huntington's but the Karlin group's comparative genome wide bioinformatic analysis has found new cases involving other amino acids such as glycine, serine, proline and glutamate. Examples include cancer related proteins, proteases and voltage-gated calcium and potassium channel protein. On the other hand, multiple amino acid runs are never found in some varieties of protein classes including metabolic enzymes, structural proteins, housekeeping proteins, chaperones, and degradation and DNA repair proteins. Interestingly, calcium and potassium channel proteins but not metal

transporters have multiple amino acid runs. In addition, 80% of Drosophila proteins containing amino acid runs are found in developmental and transcription regulation.

The potential for *CECR6* to cause diseases in this light is not very relevant to Cat Eye Syndrome in that CES is the result of a duplication of the CES Critical Region. Diseases associated with multiple amino acid runs are caused by the variation in the length of those repeats. The increase in length of amino acids can be due to polymerase slippage leading to the instability of the region. The different class of proteins associated with amino acid runs do provide an insight on one potential prediction of the CECR6 protein: the amino acid runs along with the multiple transmembrane domains containing charged amino acid discussed in the next section suggest the CECR6 protein could be a voltage-gated channel. The importance of the multiple amino acid repeats as functional elements in the *CECR6* ORF1 sequence also depends on their position compared to the predicted transmembrane domains. A hydrophobic amino acid stretch could be conserved because it increases the hydrophobicity of a tm domain. No multiple amino acid runs are found in predicted tm domains of ORF1 but repeats that are predicted to be outside of the tm domains could be a functional part of the protein. For example the ORF1 polyproline stretch is located near the carboxyl terminus of the protein and could be indicative of protein binding and represent a functional domain.

3. ORF1 is a membrane protein

I discovered that ORF1 is a tm protein and that this property is conserved in primates, mouse and fish. While six tm domains are strongly predicted in fish using the TMHMM program, the human and mouse proteins show four strongly predicted domains

and two weaker ones. Other programs predict between four and seven domains (this last one, SOSUI, which predicts seven tm domains, includes the weakest possibilities of a tm domain) and are consistent with each other as to the regions involved.

Membrane proteins are an important part of the proteome: 25% of genes are estimated to code for membrane proteins in the human genome and this number is consistent across organisms (Stevens and Arkin 2000). Tm proteins are known to have no tertiary structures and to be highly conserved. The number of tm domains influences greatly the function of the protein: shape of the pore produced (if any) (Ling, Wang *et al.* 1999), location of amino and carboxyl ends inside or outside the organelle, length of the loops or position of functional domains (Dewji and Singer 1997; Lehmann, Chiesa *et al.* 1997). There are two well studied families: four tm domains proteins including tetraspanins (Boucheix *et al*, 2001) and seven tm domains proteins including olfactory or hormone receptors, usually part of a G-coupled protein pathway (Pierce, Premont *et al.* 2002). These two groups include members that show homology in sequence and length of loops. CECR6 ORF1 shows no sequence similarity to any membrane protein families. The CECR6 tm domains were predicted based solely on their amino acid hydrophobicity pattern. Other protein groups have different numbers of tms, are not as uniform and contain smaller subclasses within the "x-tm" group (Miraglia, Godfrey *et al.* 1997). For example, some six tm domain proteins are associated with voltage-gated cation channels such as potassium, sodium and calcium (Grabowski and Black 2001; Sigworth 2003). The channel is a homotetramer of six tm domain proteins with each protein contributing a "gate". The region between the fifth and sixth tm domains is highly conserved throughout species and is called the P domain (for pore). There is no conservation seen in that region

of CECR6 ORF1 and the sixth domain is one of the least conserved (along with the fourth) in primates and mouse.

Tm domain structure is not well understood. The simplified conclusion is that between 20 and 30 consecutive amino acids form an alpha-helix that spans the membrane but membrane fluidity and other factors can cause deviations to this model (Baldwin 1993). The composition of amino acids in a tm domain also influences the properties of the protein. For example, single charged amino acids within a tm domain can cause the domain to react to voltage, displacing it vertically within the membrane. Another effect is caused by proline residues, which distort the hydrogen backbone of a protein (MacArthur and Thornton 1991). While no prolines are found in soluble alpha helices, 25% of tm domains contain a proline (Yohannan, Faham *et al.* 2004). These prolines could be selected for because of their "hinge" effect in the tm domain or the facilitation of the helical structure. Prolines may also inhibit the preliminary stages of folding the polypeptide prior to insertion into the membrane. Experimental mutation of the proline residue to alanine did not affect the original "kink" conformation of the tm domain, suggesting that the rest of the domain has evolved to keep the bent proline effect (Yohannan, Faham *et al.* 2004). Nevertheless, introduction of prolines in tm helices by point mutation *in vitro* leads to non-functional proteins while more neutral substitution with alanine produced functional proteins, suggesting that the native proline residues play a specific role in the tm domains. Proline residues were mostly tolerated near the end of the tm domains, implying that the distortion of a single proline is not localised (Yohannan, Yang *et al.* 2004).

The number of transmembrane domains predicted for *CECR6* varies between species where six strong tm domains can be seen in the fish sequences available while primates and mouse sequences show four strong domains (p=0.9 to 1) and two less supported (p=0.2 to 0.4). Two out of four main domains in human *CECR6* contain a proline residue, which is comparable to the literature. Three out of six total tm domains have a positively charged amino acid. The voltage sensibility of the ORF1 tm domains may be an important factor to consider in further functional analysis of the protein. Two of the three charged amino acids are conserved in the fish sequences. Further comparisons between the predicted ORF1 tm domains in the various species show that the tms are predicted in the same vicinity (the less conserved domains in mammals matching two of the domains in fish) even though the fish sequence is a about 100 amino acids shorter than the human. Conservation in the tm area is higher and the amino acids gaps in the fish sequence are located outside of the tm domains.

The real number of tm domains in a membrane protein can usually not be ascertained without biochemical data since many factors must be taken into account such as lipid composition of the membrane and pH (Ash, Zlomislic *et al.* 2004). The sequence conservation between human, mouse and fish sequences suggest that *CECR6* is a six tm domain protein, where the less supported domains in the human and mouse sequences are real tm domains. It is also possible that the fish had six tm domains and two of them were lost before the mouse and human evolved separately. The low score of two or three tm domains in the human and mouse sequence respectively could be attributed to divergence of early shared sequence. To complicate matters, hydrophobic domains do not have to span the membrane but instead can associate with one side of the membrane, perhaps

because of a regulatory function. For example, out of the ten hydrophobic regions of the PS1 protein, only six span the membrane (Lehmann, Chiesa *et al.* 1997).

Computer analysis of the ORF1 sequence alone will not yield definitive results as to the number of tm domains of the protein. There are some ways to determine number of tm domains in a protein experimentally. Experimental work on membrane proteins can be complicated by some of their inherent attributes. Their hydrophobicity and their need to aggregate and multimerize causes them to precipitate out of most buffers and makes them difficult to bind to nylon blotting substrates. Membrane proteins also stick to labware and are mostly present at low levels in cells. While crystallography is the method of choice to study most protein conformations, membrane protein inherent qualities do highly complicate the process, if only because their proper folding depends on insertion in the bilayer membrane (Selinsky 2003). In 1999, only 15 three dimensional membrane proteins structures had been solved. The Nobel prize in chemistry was awarded in 1988 to Johann Deisenhofer, Robert Huber and Hartmut Michel "for the determination of the three-dimensional structure of a photosynthetic reaction centre" and in 2003 Peter Agre and Roderick MacKinnon "for discoveries concerning channels in cell membranes" (http://nobelprize.org/). In both cases, the elucidation of the three dimensional structure of the membrane protein led to the discovery of their function.

Molecular biology techniques are better adapted for membrane proteins. For example, expression of the recombinant CECR6 ORF1 protein in liposomes can be useful if the ORF1 protein contains a fluorescent tag (such as the green fluorescent protein) inserted next to a predicted domain. Chemical treatment of the liposomes, degrading amino acid loops located outside of the lipid membrane will cause loss of fluorescence if

the tagged loop is outside the membrane. Fluorescence will indicate that the GFP tagged loop was protected inside the liposome. Sequential experiments, tagging each amino acid sequence between predicted tm domains, will provide the number of real tm domains. Molecular biology companies are now addressing membrane protein research needs with different kits and protocols. For example, expression of membrane proteins require the use of lipid micelles to circumvent the aggregation of the product, which are now available through Promega (Canine Pancreatic Microsomal Membranes, catalogue number Y4041). Millipore has published technical data in their catalogue showing that passivation of labware (blocking porous sites with a reagent) increases yield (protocol number PC1001EN00).

4. Phylogenetic comparison of *CECR6*

Analysis of the predicted polypeptide sequence of a gene doesn't always lead to knowledge of its structure or function. As biologists, comparison between species comes naturally. Conserved DNA regions between species can be the tool needed to direct further analysis. For example, a study of the 1.8 Mb "greater CFTR region" (cyctic fibrosis transmembrane conductance regulator and nine other genes) on human chromosome 7 and it orthologous regions in 12 other species, showed multiple small (on average of 58 bp) multi-species conserved sequences (MCS) not detectable by pairwise alignment (Thomas, Touchman *et al.* 2003). Comparison of various regions of the *CECR6* locus between vertebrate species indicated that although the ORF1 protein sequence and putative structure were conserved even in the fish species, ORF2 was not

found in a variety of fish, nor in mouse, rat, rabbit or cow. In fact, ORF2 characteristics were found only in primates, regardless of whether the mRNA is spliced or not. For example the first AUG that could act as an initiation codon in frame +3 was not present outside of the primate group. The primate group was also the only one to show conservation of nucleotide and amino acid sequence past the ORF1 termination codon to a termination codon present in the ORF2 reading frame, 69 bases downstream. Although this can be attributed to the innate conservation between primates, the region following ORF2 termination codon was not conserved, the baboon sequence even missing nine nucleotides present in the human sequence. In comparison, the sequence following ORF1 termination codon in mouse shows a gap of twelve amino acids and no frame conservation of the sequence compared to primates. This rules out the use of the region between ORF1 and ORF2 termination codons in mouse.

The analysis of the Kozak sequence surrounding the different initiation codons shows that, in accordance with the Kozak overlapping gene theory (Kozak 2001), ORF1 shows a weak Kozak sequence which could allow the ribosomes to skip the beginning of ORF1 to initiate translation of ORF2 in the unspliced *CECR6a* mRNA. The possible initiation codon for ORF2 doesn't show a perfect Kozak sequence, the sequence surrounding the first methionine available in frame +1 (trORF1) after the splice site of *CECR6b* shows a stronger Kozak than in frame +3. The relation between ORF1 and ORF2 Kozak sequences are only relevant if both reading frames are complete on the mRNA which only occurs on *CECR6*a. The intron spliced out in *CECR6*b removes most of ORF1. It is possible then that ORF2 is translated in both *CECR6*a and *CECR6*b. There can also be regulation occurring between the two mRNA isoforms where *CECR6*a would

only allow ORF1 to be translated and in *CECR6*b ORF1 is not translatable and so trORF1 or ORF2 would be translated.

5. The new *CECR6* splice variant

The RT-PCR performed on human and mouse tissue confirmed the *CECR6* mRNA splice version obtained from the FLJ EST collection (Ota, Suzuki *et al.* 2004) and revealed a new splice version of *CECR6*b missing 31 bases. The same experiment (primer location and tissue source) performed on mouse RNA did not give a spliced version, thus the two versions of *CECR6*b are not present in mouse in the tissues studied. The sequence spanned by the *CECR6*b intron includes the ORF1 start. This new splice version in humans does not produce an ORF spanning both sides of the intron, leave trORF1 and ORF2 as possibilities entirely within exon 2.

The splice can be seen as a regulatory element. Short 5'UTRs that do not contain uORFs allow higher translation efficiency of the first ORF (Mignone, Gissi *et al.* 2002). A single nucleotide 5'UTR has been showed to be enough to allow translation of an ORF in mammalian cells (Hughes and Andrews 1997). The main function of the CECR6b splice could be to increase the translation efficiency of trORF1 or ORF2 by decreasing the 5'UTR length and disrupting ORF1.

Using different forward primers in the 5'UTR, I established that the 5'UTR of *CECR6*b starts upstream of the *CECR6*a transcription start because the F2 primer only amplified the *CECR6*b mRNA while the F1 primer located 50 bp further downstream amplified both versions in the same tissues (Figure 3.14). This suggests that *CECR6*a and *CECR6*b are two different mRNAs regulated by two different promoters. These two

putative promoters and thus two starts of transcription could also accounts for the splice difference. The transcription start of *CECR6*b could be close to that of *CECR6*a but either mRNA could also include more exons upstream of the *CECR6* locus, with the known exons being the 3' end of the transcript. Interestingly, the closest TATAA nucleotide sequence, indicative of a promoter site, was found 3500 bp upstream of the predicted CECR6a and CECR6b transcription start sites. This could indicate that the promoter is located significantly upstream of the published sequence. In order to find regions upstream of the known 5'UTR, 5'RACE has been used. This experiment was performed during the initial mRNA sequencing by P. Brinkman-Mills (unpublished) and did not show sequence upstream of the published mRNA. *CECR6* ORF1 was complete and the mRNA was considered complete. Since this initial experiment did not recognize *CECR6b*, new 5'RACE experiments could uncover sequence upstream of the *CECR6b* 5'UTR while the *CECR6a* mRNA may be complete.

In light of the work done on the human and mouse *CECR6* mRNA, there are three possibilities regarding the sequence of *CECR6*b (Figure 4.1a). Assuming that the sequence of the *CECR6*b EST published (Ota, Suzuki *et al.* 2004) is complete, the first possibility of a translated product is trORF1. Because of the presence of a methionine in position 356 of ORF1, present after the splice site, it is possible that the *CECR6*b splice site allows for a smaller version of the ORF1 predicted protein. The methionine codon is also present in baboon, mouse and zebrafish but is not present in tetraodon. This way, the ORF1 truncated protein would contain a single predicted transmembrane domain and would share the ORF1 stop. This version would not fall in the category of overlapping genes in alternative reading frames.

The second possibility, assuming again that the available *CECR6*b mRNA sequence is complete, is that mutations that allowed splicing of the *CECR6* mRNA and the presence of a methionine in a different frame than ORF1, also created a small novel gene that uses an alternative reading frame overlapping the ORF1 gene (Figure 4.1b). In this version, ORF2 does not share the ORF1 amino acid sequence and termination codon. The presence of the methionine and splice site seems to be correlated in primates although this could be a coincidence. The mouse *CECR6* mRNA seems to contain no splice sites and a leucine codon instead of the methionine present at the beginning of the predicted ORF2 in primates. The structure of the ORF2 protein does not share homology with any known protein in the database and the extent of sequence or function prediction is that ORF2 is probably a soluble protein. The fact that ORF2 does not have homologous proteins in the database is to be expected since ORF1 is a conserved protein and ORF2 is using the same nucleotide sequence which has evolved according to ORF1 requirements. It could be surprising to think that the ORF2 protein has a purpose in the human proteome under these circumstances but we can refer back to the ALEX protein example (Klemke, Kehlenbach *et al.* 2001). The ALEX protein overlaps the XLαs gene's largest exon but is translated in a different reading frame. It is the example closest to the *CECR6* situation. Not only is ALEX translated in mouse, rat and humans, but the protein also binds and regulates the XLαs protein. It is conceivable that ORF2 could have a function related to ORF1, but this interaction would have appeared *de novo* in primates.

It is interesting to compare the Kozak sequence of the truncated ORF1 and ORF2 since both are always present on the *CECR6* mRNA. Since the trORF1 Kozak sequence

is strong, this gives more weight to the one transmembrane domain protein translated from the truncated ORF1 protein rather than the use of the ORF2 start site.

The last explanation for the *CECR6*b splice site is an adaptation of the Keese and Gibbs hypothesis (Keese and Gibbs 1992), which explained overlapping genes on different DNA strands by the loss of polyadenylation signal in one gene that created a longer 3'UTR or even extra coding sequence to the next available polyadenylation signal, sometimes present by chance in a gene on the opposite strand. Since the polyadenylation signal is not a palindromic site (the basic sequence is AATAAA), the site used on one strand cannot be used by an mRNA on the other strand. The new splice version of the *CECR6* mRNA I have found could be the last exon of a gene originally located upstream of the *CECR6* locus on the same DNA strand, that lost its polyadenylation signal in primates and so uses the *CECR6* mRNA as its 3'end (Figure 4.1c). This version implies that the *CECR6*b sequence found in the 5'cap EST sequencing paper represents only the 3' end of this larger gene. Initial sequence homology analysis between mouse and humans using the PIP program (Footz, Brinkman-Mills *et al.* 2001) does not identify any genes in the CES Critical Region that could splice to the *CECR6*b mRNA. The upstream parallel gene is CECR5, a nine exon gene located 24 kb away. Although it is possible that CECR5 splices to CECR6, the change in reading frames with the two different splice acceptors in human foetal brain and liver do not support this last theory. Further analysis using 5' RACE discussed earlier is necessary to test this last theory as the exact transcription start of both mRNA isoforms has yet to be determined.

6. Future work

Future work on the *CECR6* locus should concentrate on the functional significance of the three ORFs. Additionally, the function of ORF1, a possible membrane protein, could be dosage sensitive and produce negative effects as a result of the CES tetrasomy. Examples of defects due to overexpression of membrane proteins include neuropilin, a highly conserved membrane protein in frog, chicken and mouse expressed in the cardiac and nervous systems as well as in the limbs during development. Overexpression of this protein in a transgenic mouse line caused phenotypes such as extra digits, malformed heart, nerve and blood vessels (Kitsukawa, Shimono *et al.* 1995). Correct spatiotemporal expression of neuropilin was deemed essential for normal development.

Any protein experiments on the CECR6 protein(s), whether looking for functionality or localization, will be based on the translation of each ORF *in vivo*. Determining which ORFs are expressed will require using antibodies targeted to each of the ORFs to differentiate between the three hypotheses on functional ORFs described in the previous section. Size and occurrence of bands on a protein gel probed by Western analysis, especially the distance the ORF2 band migrates will be of interest to determine the existence of the ORF2 protein. Antibodies to two regions near the amino end of ORF1, outside tm domains, were ordered during this project. The antibodies were raised to small polypeptides (ten to fifteen amino acids long on the ORF1 sequence) and bound to KLH proteins. The KLH conjugated proteins were injected into rabbits to cause a strong immune reaction and increase the recovery of the targeted antibodies. Preliminary testing of the antibodies has not led to conclusive results. There is inconsistent specificity

and the background is very high. This can be due to the difficult extraction of ORF1, typical of membrane proteins, from human tissue protein medleys but could be attributable to the antibodies themselves as very few peptide antibodies have led to conclusive results in this lab. Antibodies to longer peptides or to the whole ORF1 protein may lead to more specificity. Antibodies to ORF2 and to the last third of ORF1 (where trORF1 is located) are also important for further research on the *CECR6* locus.

Functionality of membrane proteins will depend on the number of real tm domains as well as the type of membrane it spans (nuclear, endoplasmic reticulum, etc). Addressing the number of tm domains in CECR6 ORF1 was discussed earlier, with the use of commercial kits to express ORF1 in vitro in the presence of liposomes. This experiment will be complemented by the localization experiments following.

Since there are few bioinformatics leads on the function of each proteins, tests not specific to membrane or soluble proteins may shed light on protein targets or interaction. Sub-localisation of membrane proteins such as ORF1 and trORF1 as well as soluble proteins like ORF2, in mammalian cells cultures can be seen using the previously described antibodies to examine the native protein localization. Sub-cellular localization can lead to important information regarding the function. Localization can be seen using both membrane fractionation and immunofluorescence microscopy. Membrane fractionation can be used to separate cellular membranes from soluble proteins but more detailed fractionation can help separate different compartments of the cell. A GFP or luciferase tagged recombinant protein can be expressed in a mammalian cell system to also look at localisation but large protein tags can exclude the proper insertion of the protein in the membrane. In the case of the Down Syndrome Critical Region gene 2

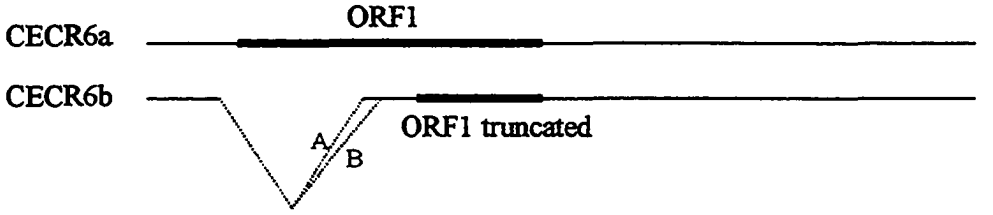
(DSCR2), sub-cellular localization identified this protein to be soluble and targeted to cytoplasmic compartments. Using membrane fractionation assays, DSCR2 was shown to be in the soluble fraction even with its the two predicted tm domains (Vesa, Brown *et al.* 2005).

In light of the ALEX protein example, interaction studies between ORF1 and ORF2 could be very informative, if both ORFs are translated at the same time. A pull down assay was used to prove that Human Rad51 and Rad52 purified proteins interact with each other as well as with two minichromosome maintenance proteins (MCM2 and 3 respectively). These interactions confirmed the double strand break DNA repair model proposed and shed light on the recruitment of the MCM proteins in this model (Shukla, Navadgi *et al.* 2005). It is possible that CECR6 ORF2 acts as a regulator of the action of ORF1 by binding it. This would not explain the spliced mRNA where ORF2 is translated alone unless trORF1 and ORF2 are both translated from CECR6b and ORF2 can bind the last third of ORF1. Assuming that all three proteins are translated, pull down assays will require extracts of one of the three proteins to be bound to a column in order to recover the second protein from a protein tissue or cell culture extract. Information on binding partners, whether between the three putative CECR6 proteins or with other native proteins, would shed light on the function of the CECR6 locus. Obtaining information on the function of the CECR6 proteins will further the CES Critical Region studies, but will also be a valuable example of overlapping genes on the same DNA strand in the human genome.

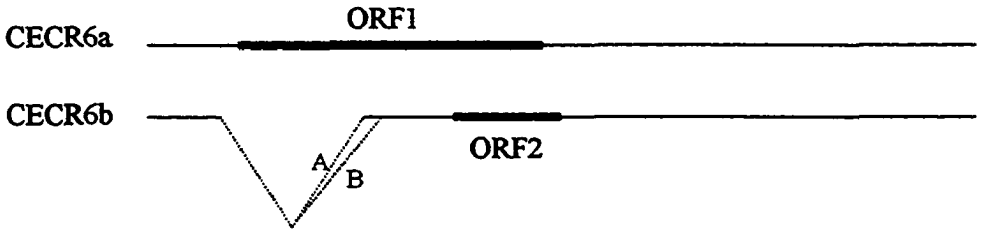
7. Conclusion

I have hypothesised that the *CECR6* locus in the CES Critical Region contains two genes in different reading frames with coding sequences overlapping on the same strand. While ORF1 is a well conserved transmembrane protein with many amino acid runs, ORF2 amino acid sequence cannot be assigned a tertiary structure. Examples of overlapping genes on the same strand in the human genome are few but evidence compiled in this research supports two alternate proteins, ORF1 in all higher vertebrates tested and trORF1 or ORF2 in primates.

a) CECR6b allows the translation of a truncated version of ORF1



b) CECR6b enables ORF2 to be translated



c) The known CECR6b could be the end of a longer mRNA

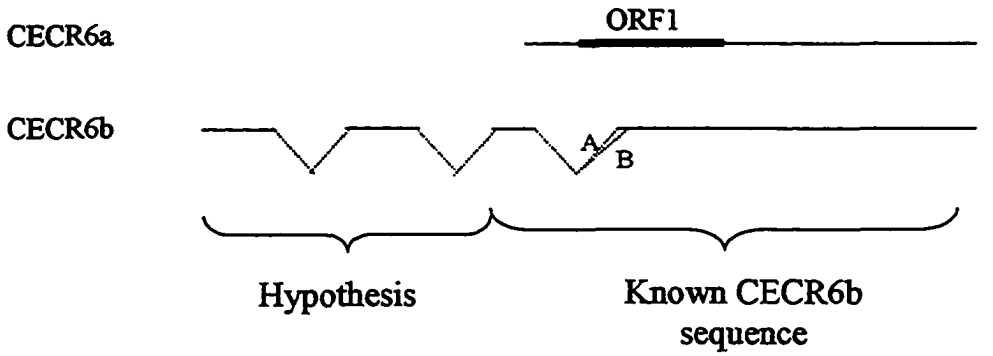


Figure 4.1: The three hypotheses of the translated region and the length of CECR6b. The known CECR6b mRNA could allow the translation of a) a truncated version of ORF1 or b) ORF2. c) Assuming that the full sequence of CECR6b is not known, exons upstream could determine the frame of the translated ORF. This last possibility is not supported by any known ESTs or by the two splice versions A (brain) and B (liver) that alter the reading frame. CECR6b mRNA sequence is compared to the known CECR6a. ORFs are represented by a thicker line. ORFs of a longer version of CECR6b can only be speculated on and the translated sequence would probably not span the last splice due to the 31 bp intron that does not conserve the frame.

REFERENCES

Abramowitz, J., D. Grenet, *et al.* (2004). "XLalphas, the extra-long form of the alpha-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex." Proc Natl Acad Sci U S A **101**(22): 8366-71.

Aparicio, S., J. Chapman, *et al.* (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." Science **297**(5585): 1301-10.

Appel, R. D., A. Bairoch, *et al.* (1994). "A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server." Trends Biochem Sci **19**(6): 258-60.

Ash, W. L., M. R. Zlomislic, *et al.* (2004). "Computer simulations of membrane proteins." Biochim Biophys Acta **1666**(1-2): 158-89.

Baldwin, J. M. (1993). "The probable arrangement of the helices in G protein-coupled receptors." Embo J **12**(4): 1693-703.

Boffelli, D., J. McAuliffe, *et al.* (2003). "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." Science **299**(5611): 1391-4.

Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Chen, C. N., T. Malone, *et al.* (1987). "At least two genes reside within a large intron of the dunce gene of Drosophila." Nature **329**(6141): 721-4.

Dewji, N. N. and S. J. Singer (1997). "The seven-transmembrane spanning topography of the Alzheimer disease-related presenilin proteins in the plasma membranes of cultured cells." Proc Natl Acad Sci U S A **94**(25): 14025-30.

Firth, A. E. and C. M. Brown (2005). "Detecting overlapping coding sequences with pairwise alignments." Bioinformatics **21**(3): 282-92.

Footz, T. K., P. Brinkman-Mills, *et al.* (2001). "Analysis of the Cat Eye Syndrome Critical Region in Humans and the Region of Conserved Synteny in Mice: A Search for Candidate Genes at or near the Human Chromosome 22 Pericentromere." Genome Res. **11**(6): 1053-1070.

Freson, K., J. Jaeken, *et al.* (2003). "Functional polymorphisms in the paternally expressed XLalphas and its cofactor ALEX decrease their mutual interaction and enhance receptor-mediated cAMP formation." Hum Mol Genet **12**(10): 1121-30.

Geer, L. Y., M. Domrachev, *et al.* (2002). "CDART: protein homology by domain architecture." Genome Res **12**(10): 1619-23.

Gilligan, P., S. Brenner, *et al.* (2002). "Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences." Gene **294**(1-2): 35-44.

Grabowski, P. J. and D. L. Black (2001). "Alternative RNA splicing in the nervous system." Prog Neurobiol **65**(3): 289-308.

Gregory-Evans, C. Y., M. J. Williams, *et al.* (2004). "Ocular coloboma: a reassessment in the age of molecular neuroscience." J Med Genet **41**(12): 881-91.

Grzybowska, E. A., A. Wilczynska, *et al.* (2001). "Regulatory functions of 3'UTRs." Biochem Biophys Res Commun **288**(2): 291-5.

Guittaut, M., S. Charpentier, *et al.* (2001). "Identification of an internal gene to the human Galectin-3 gene with two different overlapping reading frames that do not encode Galectin-3." J Biol Chem **276**(4): 2652-7.
Gustincich, S., G. Manfioletti, *et al.* (1991). "A fast method for high-quality genomic DNA extraction from whole human blood." Biotechniques **11**(3): 298-300, 302.
Hacia, J. G. (2001). "Genome of the apes." Trends Genet **17**(11): 637-45.
Henikoff, S., M. A. Keene, *et al.* (1986). "Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands." Cell **44**(1): 33-42.
Hirokawa, T., S. Boon-Chieng, *et al.* (1998). "SOSUI: classification and secondary structure prediction system for membrane proteins." Bioinformatics **14**(4): 378-9.
Hooper, P. M., H. Zhang, *et al.* (2000). "Prediction of genetic structure in eukaryotic DNA using reference point logistic regression and sequence alignment." Bioinformatics **16**(5): 425-38.
Hughes, M. J. and D. W. Andrews (1997). "A single nucleotide is a sufficient 5' untranslated region for translation in an eukaryotic in vitro system." FEBS Lett **414**(1): 19-22.
Johnson, Z. I. and S. W. Chisholm (2004). "Properties of overlapping genes are conserved across microbial genomes." Genome Res **14**(11): 2268-72.
Jones, D. T., M. Tress, *et al.* (1999). "Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure." Proteins **Suppl 3**: 104-11.
Juretic, D., L. Zoranic, *et al.* (2002). "Basic charge clusters and predictions of membrane protein topology." J Chem Inf Comput Sci **42**(3): 620-32.
Karlin, S., L. Brocchieri, *et al.* (2002). "Amino acid runs in eukaryotic proteomes and disease associations." Proc Natl Acad Sci U S A **99**(1): 333-8.
Karlin, S., C. Chen, *et al.* (2002). "Associations between human disease genes and overlapping gene groups and multiple amino acid runs." PNAS **99**(26): 17008-17013.
Keese, P. K. and A. Gibbs (1992). "Origins of genes: "big bang" or continuous creation?" Proc Natl Acad Sci U S A **89**(20): 9489-93.
Kelley, L. A., R. M. MacCallum, *et al.* (2000). "Enhanced genome annotation using structural profiles in the program 3D-PSSM." J Mol Biol **299**(2): 499-520.
Kitsukawa, T., A. Shimono, *et al.* (1995). "Overexpression of a membrane protein, neuropilin, in chimeric mice causes anomalies in the cardiovascular system, nervous system and limbs." Development **121**(12): 4309-18.
Klemke, M., R. H. Kehlenbach, *et al.* (2001). "Two overlapping reading frames in a single exon encode interacting proteins--a novel way of gene usage." Embo J **20**(14): 3849-60.
Kozak, M. (1987)a. "An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs." Nucleic Acids Res **15**(20): 8125-48.
Kozak, M. (1987)b. "At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells." J Mol Biol **196**(4): 947-50.
Kozak, M. (2001). "Extensively overlapping reading frames in a second mammalian gene." EMBO Rep **2**(9): 768-9.
Kozak, M. (2001). "New ways of initiating translation in eukaryotes?" Mol Cell Biol **21**(6): 1899-907.

Kozak, M. (2002). "Emerging links between initiation of translation and human diseases." Mamm Genome **13**(8): 401-10.
Krakauer, D. C. and J. B. Plotkin (2002). "Redundancy, antiredundancy, and the robustness of genomes." Proc Natl Acad Sci U S A **99**(3): 1405-9.
Kuersten, S. and E. B. Goodwin (2003). "The power of the 3' UTR: translational control and development." Nat Rev Genet **4**(8): 626-37.
Kumar, S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." Proc Natl Acad Sci U S A **99**(2): 803-8.
Lehmann, S., R. Chiesa, *et al.* (1997). "Evidence for a six-transmembrane domain structure of presenilin 1." J Biol Chem **272**(18): 12047-51.
Levinson, B., S. Kenwrick, *et al.* (1990). "A transcribed gene in an intron of the human factor VIII gene." Genomics **7**(1): 1-11.
Ling, K., P. Wang, *et al.* (1999). "Five-transmembrane domains appear sufficient for a G protein-coupled receptor: functional five-transmembrane domain chemokine receptors." Proc Natl Acad Sci U S A **96**(14): 7922-7.
Luukkonen, B. G., W. Tan, *et al.* (1995). "Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance." J Virol **69**(7): 4086-94.
MacArthur, M. W. and J. M. Thornton (1991). "Influence of proline residues on protein conformation." J Mol Biol **218**(2): 397-412.
Mignone, F., C. Gissi, *et al.* (2002). "Untranslated regions of mRNAs." Genome Biol **3**(3): REVIEWS0004.
Miraglia, S., W. Godfrey, *et al.* (1997). "A novel five-transmembrane hematopoietic stem cell antigen: isolation, characterization, and molecular cloning." Blood **90**(12): 5013-21.
Moller, S., M. D. Croning, *et al.* (2001). "Evaluation of methods for the prediction of membrane spanning regions." Bioinformatics **17**(7): 646-53.
Nobrega, M. and L. A. Pennacchio (2003). "Comparative Genomic Analysis as a Tool for Biological Discovery." J Physiol.
Ota, T., Y. Suzuki, *et al.* (2004). "Complete sequencing and characterization of 21,243 full-length human cDNAs." Nat Genet **36**(1): 40-5.
Petruska, J., M. J. Hartenstine, *et al.* (1998). "Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease." J Biol Chem **273**(9): 5204-10.
Pierce, K. L., R. T. Premont, *et al.* (2002). "Seven-transmembrane receptors." Nat Rev Mol Cell Biol **3**(9): 639-50.
Ponting, C. P., J. Schultz, *et al.* (1999). "SMART: identification and annotation of domains from signalling and extracellular protein sequences." Nucleic Acids Res **27**(1): 229-32.
Poulin, F., A. Brueschke, *et al.* (2003). "Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK." J Biol Chem **278**(52): 52290-7.
Quelle, D. E., F. Zindy, *et al.* (1995). "Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest." Cell **83**(6): 993-1000.

Rogozin, I. B., A. V. Kochetov, *et al.* (2001). "Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon." Bioinformatics **17**(10): 890-900.

Schinzel, A., W. Schmid, *et al.* (1981). "The "cat eye syndrome": dicentric small marker chromosome probably derived from a no.22 (tetrasomy 22pter to q11) associated with a characteristic phenotype. Report of 11 patients and delineation of the clinical picture." Hum Genet **57**(2): 148-58.

Selinsky, B. S. (2003). Membrane Protein Protocol. Totowa, New Jersey, Humana Press.

Shambrook J., R. D. W. (2001). Molecular cloning: a laboratory manual. Cold Spring Harbor, Ney York, Cold Spring Harbor Laboratory Press.

Shendure, J. and G. M. Church (2002). "Computational discovery of sense-antisense transcription in the human and mouse genomes." Genome Biol **3**(9): RESEARCH0044.

Shintani, S., C. O'HUigin, *et al.* (1999). "Origin of gene overlap: the case of TCP1 and ACAT2." Genetics **152**(2): 743-54.

Shukla, A., V. M. Navadgi, *et al.* (2005). "Interaction of hRad51 and hRad52 with MCM complex: A cross-talk between recombination and replication proteins." Biochem Biophys Res Commun **329**(4): 1240-5.

Sigworth, F. J. (2003). "Structural biology: Life's transistors." Nature **423**(6935): 21-2.

Sonnhammer, E. L., G. von Heijne, *et al.* (1998). "A hidden Markov model for predicting transmembrane helices in protein sequences." Proc Int Conf Intell Syst Mol Biol **6**: 175-82.

Stevens, T. J. and I. T. Arkin (2000). "Do more complex organisms have a greater proportion of membrane proteins in their genomes?" Proteins **39**(4): 417-20.

Stormo, G. D. (2000). "Gene-finding approaches for eukaryotes." Genome Res **10**(4): 394-7.

Tatusova, T. A. and T. L. Madden (1999). "BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences." FEMS Microbiol Lett **174**(2): 247-50.

Thomas, J. W., J. W. Touchman, *et al.* (2003). "Comparative analyses of multi-species sequences from targeted genomic regions." Nature **424**(6950): 788-93.

Tusnady, G. E. and I. Simon (2001). "The HMMTOP transmembrane topology prediction server." Bioinformatics **17**(9): 849-50.

Veeramachaneni, V., W. Makalowski, *et al.* (2004). "Mammalian overlapping genes: the comparative perspective." Genome Res **14**(2): 280-6.

Vesa, J., Y. Brown, *et al.* (2005). "Molecular and cellular characterization of the Down syndrome critical region protein 2." Biochem Biophys Res Commun **328**(1): 235-42.

von Heijne, G. (1999). "A Day in the Life of Dr K. or How I Learned to Stop Worrying and Love Lysozyme: a tragedy in six acts." J Mol Biol **293**(2): 367-79.

Wray, G. A., M. W. Hahn, *et al.* (2003). "The evolution of transcriptional regulation in eukaryotes." Mol Biol Evol **20**(9): 1377-419.

Xiong, W., C. C. Hsieh, *et al.* (2001). "Regulation of CCAAT/enhancer-binding protein-beta isoform synthesis by alternative translational initiation at multiple AUG start sites." Nucleic Acids Res **29**(14): 3087-98.

Yohannan, S., S. Faham, *et al.* (2004). "The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors." Proc Natl Acad Sci U S A **101**(4): 959-63.

Yohannan, S., D. Yang, *et al.* (2004). "Proline substitutions are not easily accommodated in a membrane protein." J Mol Biol 341(1): 1-6.