UPTEC STS05001
Examensarbete 20 p
Februari 2005

Identification and Analysis of Splice Variants in the Human Beta Subgroup of Rhodopsin G Protein Coupled Receptors

Malena Ingemansson

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Identification and Analysis of Splice Variants in the Human Beta Subgroup of Rhodopsin GPCRs

Malena Ingemansson

G-Protein-Coupled Receptors -GPCRs is a family of integral membrane protein responsible for signaling transduction in physiological processes in virtually every system in the body. This makes them attractive drug targets and today 50 percent of all therapeutics is designed to target these receptors. Alternative RNA splicing makes it possible for a single gene to be expressed as several different proteins -splice variants, and it is a widespread phenomenon in the mammalian genome. To learn more about this mechanism and how it affects the expression of GPCRs is therefore an important research objective in the search for more effective and new drugs. This report presents a bioinformatical method to identify and analyse splice variants for GPCRs and was used on the b subgroup of the Rhodopsin family. The method proved to be efficient and straightforward. Surprisingly, only a fraction of all the discovered splice variants resulted in proper receptors while the rest was classified as non-functional. However, the commonness of these seemingly non-functional receptors and the limited number of splice variants for each receptor indicates that the splicing mechanism is regulated and that these variants in fact serve a purpose. This purpose could be as regulative factors controlling the amount and function of functional receptors. The results of this study hint how the splicing mechanism regulates GPCRs but to be able to draw general conclusions the method should be used on larger datasets.

Handledare: Thóra K. Bjarnadóttir, Robert Fredriksson
Ämnesgranskare: Helgi B. Schiöt
Examinator: Elisabet Andresdóttir
ISSN: 1650-8319, UPTEC STS05001
Tryckt av: Intendenture, Ångströmslaboratoriet

Populärvetenskaplig beskrivning

G-proteinkopplade receptorer (förkortas GPCR på engelska) är en av de största proteinfamiljerna i däggdjursgenomet. Dessa proteiner agerar receptorer för en mängd olika *ligander* (små molekyler som binder till specifika större molekyler) och förekommer i praktiskt taget alla kroppsvävnader. På grund av detta utgör de mål för många olika typer av mediciner, så mycket som 50 procent av dagens alla läkemedel är designade för dessa receptorer. Den *centrala dogmen* visar hur en gen kodar för ett visst protein och att det är *mRNA* (messenger-RNA) som överför genens information till proteinsyntesen. En gen består av *exoner* (proteinkodande regioner) och *introner* (icke-kodande regioner) medan den motsvarande mRNA-sekvensen endast består av exoner, de proteinkodande delarna av genen. Den mekanism som avlägsnar intronen kallas för *RNA-splicing* (eng.) och innebär att intronen klipps bort och att de omgivande exonerna fogas ihop. Under denna process kan det hända att skarvningen inte sker på det stället i mRNA-sekvensen som det brukar göra. En annorlunda skarvning resulterar i en annorlunda mRNA-sekvens –en *splice-variant*, som translateras till ett protein. Denna förekomst av *alternativ RNA-splicing* innebär att en enda gen kan uttryckas som ett flertal olika proteiner och är ett utbrett fenomen i däggdjursgenomet. Att bedriva forskning kring denna mekanism för att bättre förstå hur den påverkar uttrycket av GPCR:er är därför viktigt för att kunna utforma nya och bättre läkemedel. Idag utförs mycket av den biologiska forskningen via datorer och program utformas för att kunna sortera data i form av DNA, RNA och proteinsekvenser. Detta är ett relativt nytt forskningsområde som kallas *bioinformatik* och det finns en ständig efterfrågan på datorer som klarar av att hantera allt större biologiska datamaterial. Den här rapporten presenterar en bioinformatisk metod för identifiering och analys av splice-varianter för GPCR:er och användes på ett litet datamaterial för en av de fem GPCR-familjerna *–Rhodopsin*. Metoden visade sig vara effektiv och användbar. Resultaten, som visade att endast en bråkdel av alla de upptäckta splice-varianterna resulterade i funktionella receptorer, var överraskande. Merparten av splice-varianterna medförde med andra ord receptorer som klassades som icke-funktionella, vilket innebär att dessa proteiner saknade viktiga delar för klassisk känd receptorfunktion. Den höga förekomsten av dessa till synes icke-funktionella receptorer antyder dock att de inte är produkter av en felaktig mekanism utan att de faktiskt tjänar ett syfte. Detta syfte kan vara såsom regulativa faktorer som kontrollerar mängd och funktion av funktionella receptorer. Studiens resultat ger en möjlig bild av hur splicing-mekanismen reglerar GPCR:er men för att kunna dra generella slutsatser bör metoden användas på större datamaterial.

Table of Contents

Introduction

G-Protein-Coupled Receptors -GPCRs constitute a superfamily that is one of the largest protein families in the mammalian genome. This superfamily of integral membrane proteins can be divided into five main families named *Glutamate*, *Rhodopsin*, *Adhesion*, *Frizzled* and *Secretin*. (Fredriksson R. et al., 2003) Because of their central role in many physiological processes such as blood pressure control, insulin secretion and central nervous system functions, these receptors are highly used drug targets.

Characterizing GPCRs

A characteristic feature of the GPCRs are seven transmembrane (TM) alpha helices connected by six loops, three cytoplasmic and three extracellular. This means that the protein winds back and forth through the plasma membrane seven times. The TM region is highly conserved with recurring amino acid motifs whereas the loops can vary in length depending on the receptor. The C-terminal resides on the cytoplasmic side while the N-terminal is located on the extracellular face of the plasma membrane. (Palczewski K. et al., 2000)

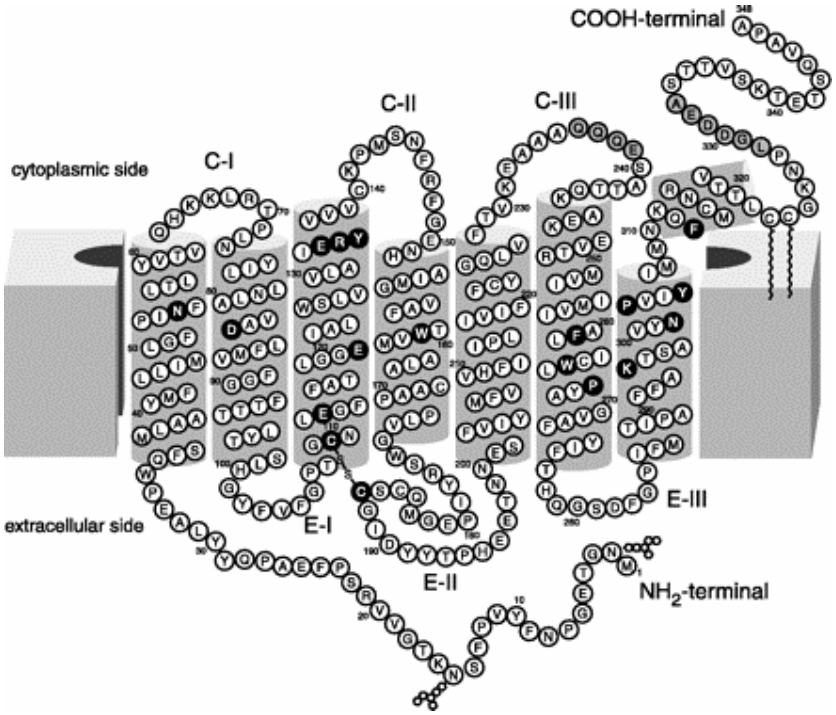


Figure 1. Two-dimensional model of bovine rhodopsin with conserved motifs in black. Picture taken from Palczewski K et al., 2000.

When it comes to function there is a wide variety of GPCRs such as light-activated receptors in the eye, thousands odorant receptors in the nose and receptors for a range of hormones and neurotransmitters (Lodish H. et al., 2000). As a consequence of these different functions, the ligands triggering the signaling pathway of the GPCRs are also of varying kind. A few examples are photons, organic odorants, amines, peptides, proteins, nucleotides etc. (Fredriksson R. et al., 2003). Despite the different functions and ligands,

the signaling pathway that transduces extracellular signals to the cell interior is virtually the same for all GPCRs.

Rhodopsin Family

The *Rhodopsins* represent the largest GPCR family and is divided into the four subgroups α, β, γ and δ (Fredriksson R. et al., 2003). It is the only family with an eighth alpha helix, which is located between the seventh TM region and the C-terminal (see figure 1). The receptors within the *Rhodopsin* family have varying functions but the most famous is the one of *phototransduction*. This means that the receptors convert light signals into nerve signals to the brain and are therefore responsible for our vision. All of these receptors belong to α, the biggest subgroup. With 35 receptors, β is the smallest subgroup and all the known ligands for these receptors are peptides (Fredriksson R. et al., 2003).

Signaling Pathway

The signal transduction involves five main components: a ligand, a receptor, a G-protein, an *effector* protein and a *second messenger*. The heterotrimeric G-protein consists of the three subunits α, β and γ. The GPCR interacts directly with the G-protein via the α-subunit. When the ligand binds to the GPCR and turns it into an active state, the α-subunit releases the GDP that has been bound to the G-protein (hence the name of the protein) in its inactive state. The release of GDP and a subsequent binding of GTP activate the G-protein. (Eckhardt N., 2004) The activation leads to a separation between the GTP-α-subunit and the βγ-dimer, which enables the GTP-α-subunit to bind to an effector protein such as adenylyl cyclase. In this interaction GTP hydrolyses to GDP, which causes the synthesis of a second messenger; for example cAMP. The second messenger in turn initiates a series of intracellular events. The inactive GDP-α-subunit detaches from the effector protein and binds to the βγ-dimer and so the signaling pathway is terminated. (Lodish H. et al., 2000)

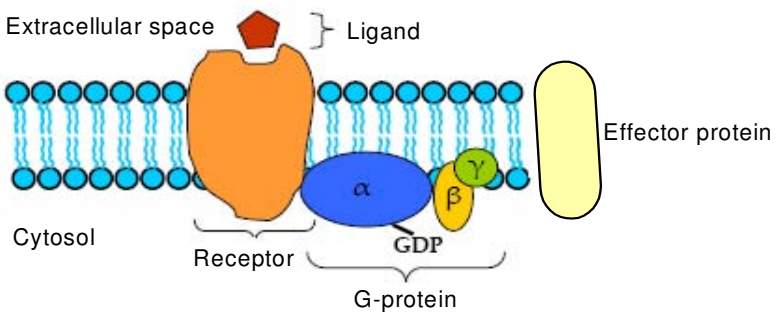


Figure 2. The status of the receptor, the inactive G-protein and the effector protein just before the ligand binds to the receptor. At ligand-receptor binding the GDP-α-subunit is detached from the other subunits and GTP replaces GDP. The second messenger forms when the GTP in the GTP-α-subunit, which is then attached to the effector protein, hydrolyses to GDP. Modified picture from Navigant Consulting, Inc.; PubMed.

Medical Application

GPCRs play a crucial role in physiological cell signaling in almost every tissue in the body and this makes them attractive targets for drug development. More than 50 percent of the therapeutic drugs currently available are targeted towards theses receptors and the complaint or disease that they are meant to cure are of varying kind: asthma, schizophrenia, migraine, nausea, incontinence, anxiety etc. (Nambi P. et al., 2003) The prospects in drug developing for GPCRs are very good since only about 40 percent of these receptors have a known physiological function and has been used as drug targets. The GPCRs with yet unknown function therefore constitute future drug targets with high potential. (www.chairs.gc.ca/web/chairholders, 2005-01-10*)* These receptors are called *orphans* and they have been characterized as GPCRs based on sequence homology with family members of GPCR but their ligand is not known. Many pharmaceutical companies are working on improving their methods to identify orphan GPCRs and finding their ligands to be able to develop appropriate drugs. (Nambi P. et al., 2003)

RNA Splicing Mechanism

The so-called *central dogma* describes how the DNA code for a gene is transcribed into mRNA, which in turn directs the assembly of proteins (Lodish H. et al., 2000). What it does not describe is the multiple steps of mRNA processing that take place before a mature mRNA is produced. RNA splicing is the final step in this process and is of special interest here.

The gene can be divided into two parts called *exons* and *introns*. They are both made up of the same nucleotides: A (adenine), T (thymine), C (cytosine) and G (guanine) but whereas exons code for amino acids introns do not. Being directly translated into amino acids, mature mRNA does not contain introns, only exons. The RNA splicing takes place in the nucleus before the mRNA is transported into the cytoplasm and is translated into proteins by the ribosomes. The procedure of removing internal introns and splicing exons is figuratively a matter of cut and paste, the intron is detached from the sequence and the two exons flanking the intron are joined. It has been established that a "GU" at the 5' splice site and an "AG" at the 3' splice site is invariant which the splicing machinery identifies and uses to correctly splice them together (Lodish H. et al., 2000).

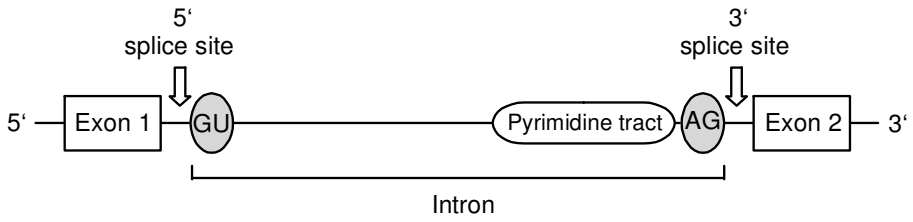


Figure 3. The invariant nucleotides GU at the 5'splice site and AG at the 3' splice site.

The operating tool in this process is called a *spliceosome* and is composed of five splicing s*nRNPs* -small nuclear ribonucleoprotein particles. Two of the snRNPs are responsible for the folding of the intron so that the two flanking exons lay on top of each another. The spliceosome is formed and proceeds with two *transesterifictions*. The first detaches one intron end from one of the exons; the second detaches the other intron end and joins the two exons. The lariat (detached) intron remains linked to the snRNPs while the joined exons are released. Finally, the intron-snRNP complex dissociates and the snRNPs are ready for another splicing. (Lodish H. et al., 2000)

Alternative RNA splicing is a term comprising the event of splicing not taking place at the "usual" site in the mRNA sequence. The resulting mature mRNA sequence is called a *splice variant*. The process renders possible a single gene to be expressed as several different mRNA isoforms, which result in just as many proteins. Hence, it is a mechanism that regulates gene expression and gives multitude to the genome. The possible events are exon skipping, inclusion of alternative exons, use of alternative 5' splice site and 3' splice site and intron retention (an intron is included as an exon). (Cochet O. et al., 2003)

EST Sequences

ESTs –expressed sequence tags, are short cDNA fragments derived from mRNA by using reverse transcriptase (an enzyme transcribing mRNA to complementary DNA - cDNA). ESTs are in various ways practical in identifying genes and analysing gene expression. For instance, the mRNA molecules that are present in a cell or tissue represent a specific portion of the genome that is expressed. Producing ESTs (sequencing bits of cDNA) of that entire mRNA material grants a gene expression profile connected to a specific tissue. (www.pedb.org, 2005-01-11) Hence, these "tags" that represent expressed genes make it possible to find genes by matching base pairs. Being sequenced from cDNA, which only consists of expressed DNA sequence, these sequences are also useful in identifying exons. (www.ncbi.nlm.nih.gov, 2005-01-12)
ESTs were used in the pursuit to complete the nucleotide sequence of the human genome but it was not until later their true value, described above, was discovered. This incomplete but important data has also proven to be much more rapidly and cheaply obtained than full-length sequences. (Boguski M., 1995)

Bioinformatics

Bioinformatics is a fairly new research area that can be called multidisciplinary since it involves the areas of biology, mathematics and computer science. The material of study is biological data processed in various ways. The two main types of databases being created are nucleotide and protein databases for different genomes in the vegetable and animal kingdom. (Andersson S.G.E., 2004) To produce new sequence data is not a problem, it is rather the computer resources to process it and finding efficient algorithms to be able to use it that are the bottlenecks. Sequence data is not of much use if it is not sorted and divided into biologically differing elements such as coding and non-coding regions (Andersson S.G.E., 2004). Bioinformatical work can be divided into two categories: development and analysis. The people working on the development side design algorithms that will sort out the information of interest while the people on the analysis side use these programs to analyse their biological data. From the biological databases users can produce all sorts of information such as protein sequences, predict protein domains and establish evolutionary descendants.

Aim of Study

The aim of this study is to create and use a bioinformatical method to identify and analyse splice variants for GPCRs by using ESTs (and full-length mRNA sequences). Splice variants are expressed as different mRNAs compared to the original, or most common, gene and it is therefore of interest to analyse to what extent they exist for GPCRs and how

they function as receptors. When a splice variant has been identified the analysis should therefore provide information on its exon distribution on the genome, resulting protein and functionality as a receptor. The *Rhodopsin* β subgroup is the chosen dataset due to its availability and limited size.

Material and Methods

Data Retrieval

The greater part of the EST sequences were retrieved from a project earlier conducted at the department (Gloriam D., 2004) while the mRNA sequences and the remaining EST sequences were recovered in the human BLAT search database (www.genome.ucsc.edu/cgi-bin/hgBlat). To simplify, the collection of EST/mRNA sequences will hereafter be referred to as just mRNA sequences. The GNRHRII receptor was excluded from the material because of its status as a psuedo gene (a gene not resulting in a protein).

A first and important step was to determine whether or not there were enough mRNA sequences for each receptor to perform the investigation on the *Rhodopsin* β subgroup. A compilation of the material showed that 26 out of 35 receptors had a selection of 10 mRNA sequences or more (see figure 3 under Results). This was considered to be sufficient.

Identification of Splice Variants

To be able to identify splice variants the "recognized" or most occurring receptor sequence for each receptor was used, this sequence will be referred to as the template (Fredriksson R. et al., June 2003). By running the mRNA sequences, collected for each receptor, together with the template sequence in the human BLAT search, a graphic description could be obtained. The description showed the positions of the exons and introns of the template and mRNA sequences on the genome. This means that if there were a splice variant for the receptor this would appear as an mRNA sequence with exons and introns at other positions on the genome than the template sequence. To set the exact positions of the exons, the deviant sequence and the template sequence were separately aligned to the genome in NCBI BLAST (www.ncbi.nih.gov/BLAST/). The genome sequence was obtained from the human BLAT search database. The next step was to establish if the splice sites of the possible variant agreed with the splicing consensus sequence. To verify the invariant nucleotides (GU and AG) at the splice sites is a good quality test of the splice variant's authenticity. A splice variant not displaying these nucleotides at the appropriate positions is consequently not a real splice variant. Instead, it could be the product of a sequencing error.

Determination of Reading Frame

When it had been established that a deviant mRNA sequence could be a genuine splice variant the next step was to investigate its resulting amino acid sequence. First the template sequence was translated in EditSeq (LaserGene, DNAstar) and to minimize the risk of sequence errors the genome sequence, derived from BLAT, was used. To find the correct open reading frame the deviant mRNA sequence, also derived from BLAT, was aligned to the amino acid sequence of the template with BLAST. This was not possible in cases of splice variants not having a single exon aligning to the template. It occurred only for four

receptors. With the correct reading frame the complete amino acid sequence of the variant could then be displayed by using EditSeq.

Establishing Protein Structure

In a research article written by Krysztof Palczewski et al. (2000) a two-dimensional model of the protein structure of bovine rhodopsin is presented (see figure 1 under Introduction). Most importantly, it displays the positions of conserved motifs, amino acids of which many recur in receptors in the *Rhodopsin* family. The conserved motifs always follow the same pattern regarding their positions in the seven TM regions. To recover these conserved motifs in the amino acid sequence of a receptor is therefore a way of estimating its protein structure. By using MegAlign (LaserGene, DNAstar) to align the amino acid sequence of the template and the possible splice variant to the amino acid sequence of bovine rhodopsin, the conserved motifs could often be located. A useful and complementing source of information in this regard was the BLAST conserved domain database. It contains conserved domains often present in proteins and displays a hit with the query sequence as a pair wise or multiple alignment. With these two tools the recurring amino acids were found in the template sequences and determination of the protein structure, for the templates as well as the splice variants, was made possible. To be able to see what part of the protein structure that was encoded by which exon a simple calculation of the resulting number of amino acids from each exon was performed (see appendix 3).

Questioning the Authenticity

Since sequencing errors and other laboratory related errors are factors that must be considered the library origin of the mRNA sequences supporting the different variants were inspected. A splice variant that is supported by mRNA sequences of different library origin is more likely to be genuine. The information was collected from NCBI Entrez (www.ncbi.nlm.nih.gov/entrez/query.fcgi). If the mRNA sequences came from normalized libraries this was also noted. In a normalized library the representation of abundant genes has been reduced. This facilitates the identification of rare genes but there is also the chance that a rare gene thought to be authentic is just part of the noise, meaning incorrect sequences, present in all libraries. Hence, the normalization process is vital in the search for unusual genes but at the same time it increases the possibility of selecting erroneous sequences. (Reddy A., 2002)

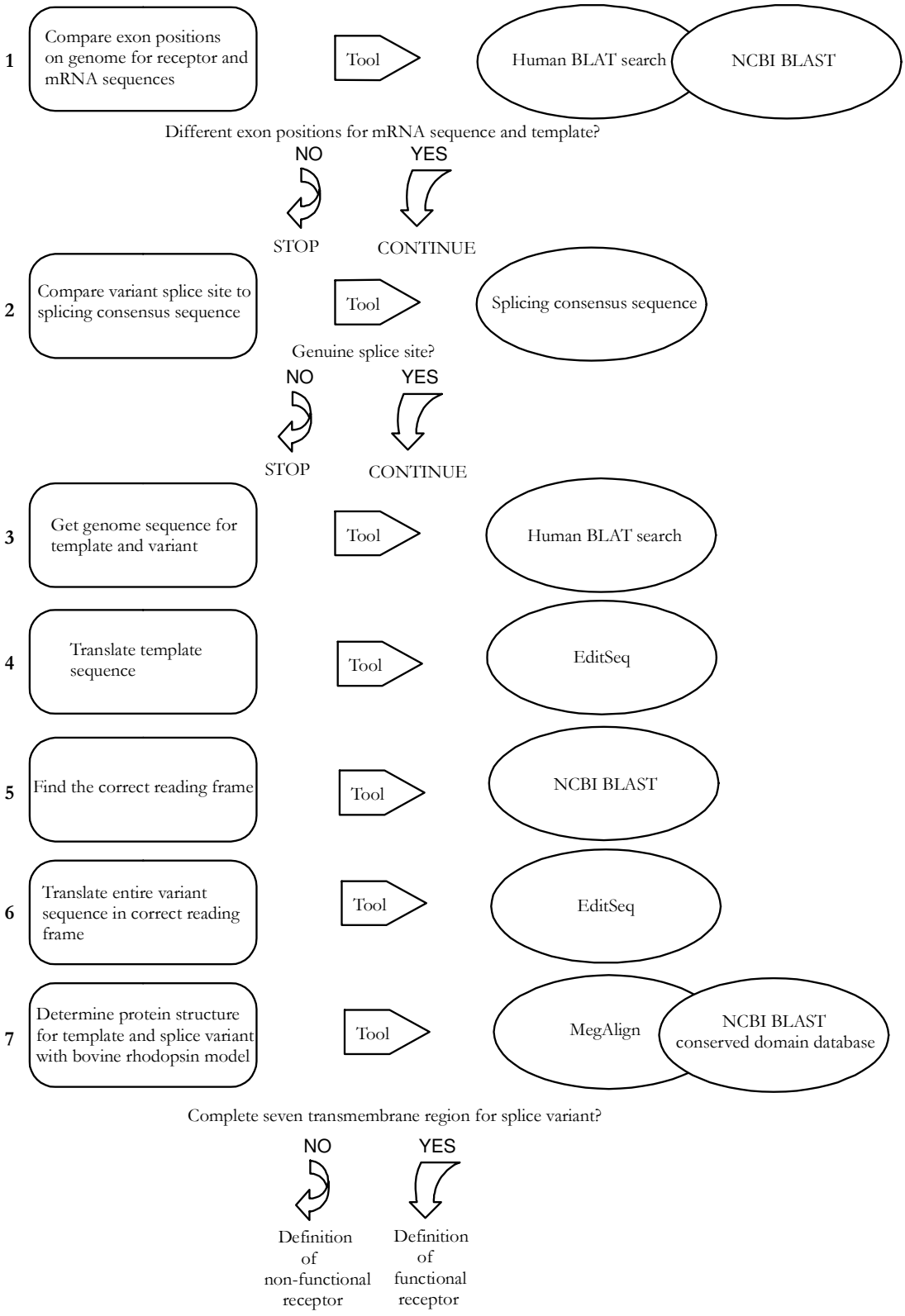


Figure 4. A step-by-step model of the procedure used to identify and analyse splice variants.

Results

Exon Distribution on Protein Structure and on the Genome

The schematic presentation of exon distribution on protein structure and on genome for the templates and the splice variants as well as detailed information on mRNA designations and tissue origin is shown in appendix 1. For each receptor (with a variant) there is first a template showing the exon distribution on the protein structure and beneath it the exon distribution on the genome, then follows the figures showing the deviant structures of the splice variants. The representation of the variant protein structure is presented in reference to the template and so it shows how the variant differs from the template, not the variants actual structure. In appendix 2 and 3 this schematic information is also presented in tables; appendix 2 contains information on protein structures and tissue origins, appendix 3 contains information on genome exon distributions.

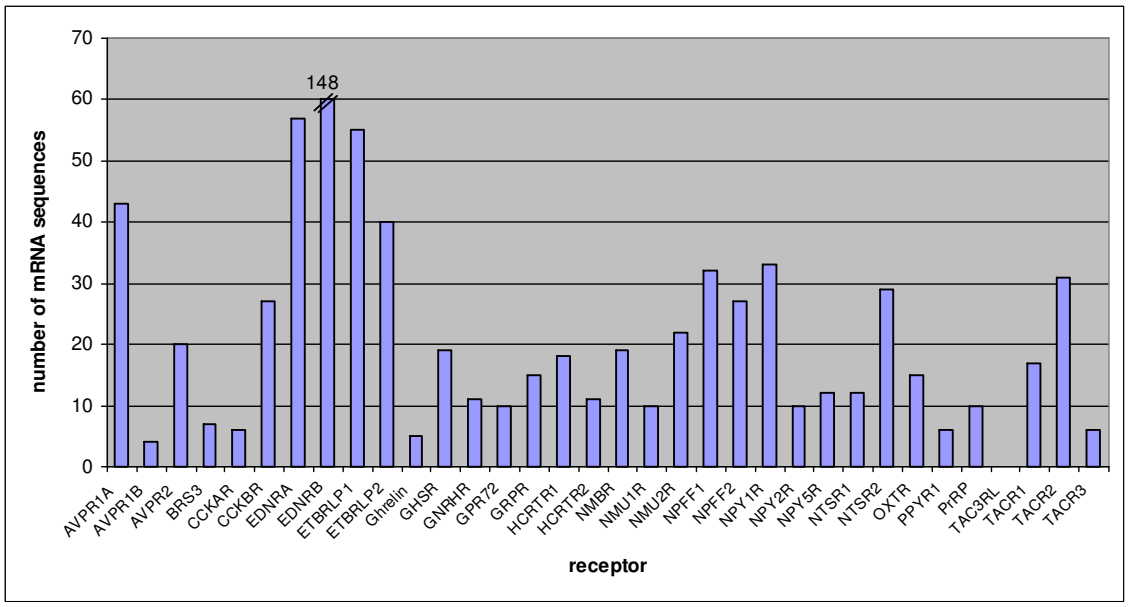


Figure 5. The chart shows the total number of mRNA sequences currently existing for each receptor in the Rhodopsin β subgroup (excluding GNRHRII). The broken column for EDNRB has its correct value written above it, this procedure improved the resolution for the other columns.

Functionality

An assembled mRNA sequence is a variant sequence if its exon positions on the genome differ from the template and the variant splice sites are genuine according to the splicing consensus sequence. For a variant sequence to be functional it must present a complete seven TM region, if not it is considered to be non-functional. An extension or shortening of the N- or C-terminal is within the frame of what is considered to be functional.

The results show that 19 out of 34 receptors have one splice variant or more but that only 2 receptors have a splice variant that is functional. These two receptors also have variants that are non-functional which means that there are non-functional variants for all 19 receptors. In number of mRNA sequences this means that four sequences represent functional variants while 70 sequences represent non-functional variants. Figure 6a

displays the high occurrence of sequences representing non-functional variants. The left most column represents the number of mRNA sequences agreeing with the template, the middle column represents the number of mRNA sequences supporting non-functional variants and the column to the right represents the number of mRNA sequences supporting functional variants. As shown in the chart, this last column only occurs for two receptors: CCKBR and EDNRB.

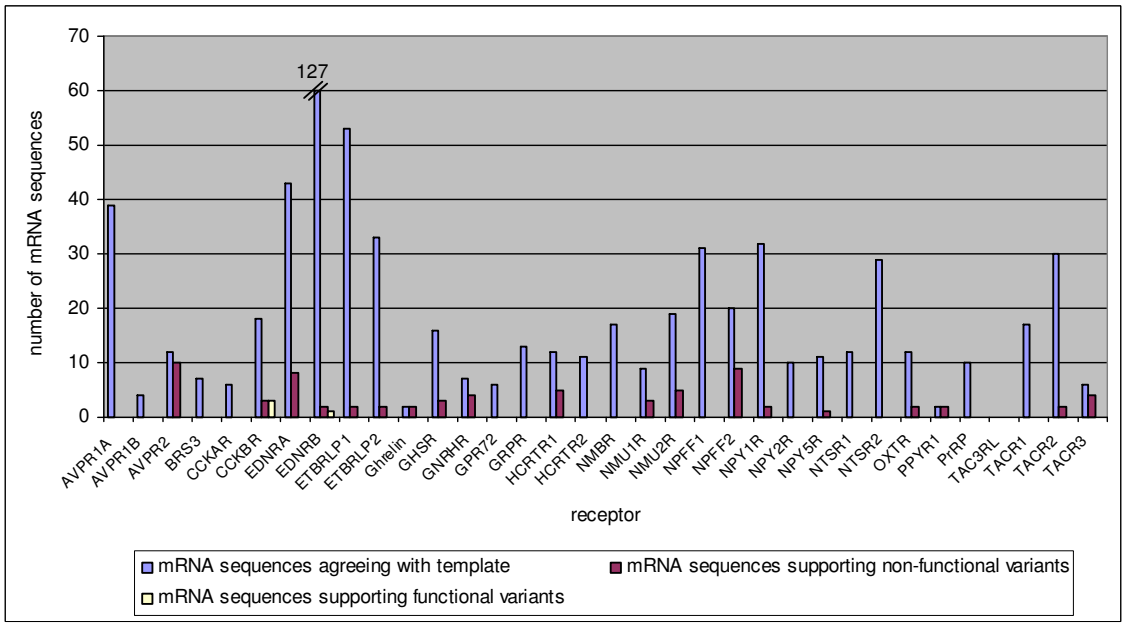


Figure 6a. The chart shows three bars for each receptor. They represent the number of mRNA sequences agreeing with the template, the number of mRNA sequences supporting non-functional variants and the number of mRNA sequences supporting functional variants. mRNA sequences proven not to have genuine splice sites are excluded. The broken column for EDNRB has its correct value written above it, this procedure improved the resolution for the other columns.

In many cases the difference between the number of sequences representing the template and those representing non-functional variants is small. Even though this is true there is no obvious connection between a receptors total number of mRNA sequences and its number of variant sequences (see figure 6b). Consequently, a receptor with a high total number of mRNA sequences does not necessarily have a high number of variant sequences. Furthermore, there is no trend towards a certain number of sequences guaranteeing the existence of variants. The difference in number between the number of sequences supporting the template in figure 6a and the total number of sequences in figure 6b is the exclusion of sequences supporting non-functional variants, sequences supporting functional variants and sequences not having genuine splice sites in figure 6a.

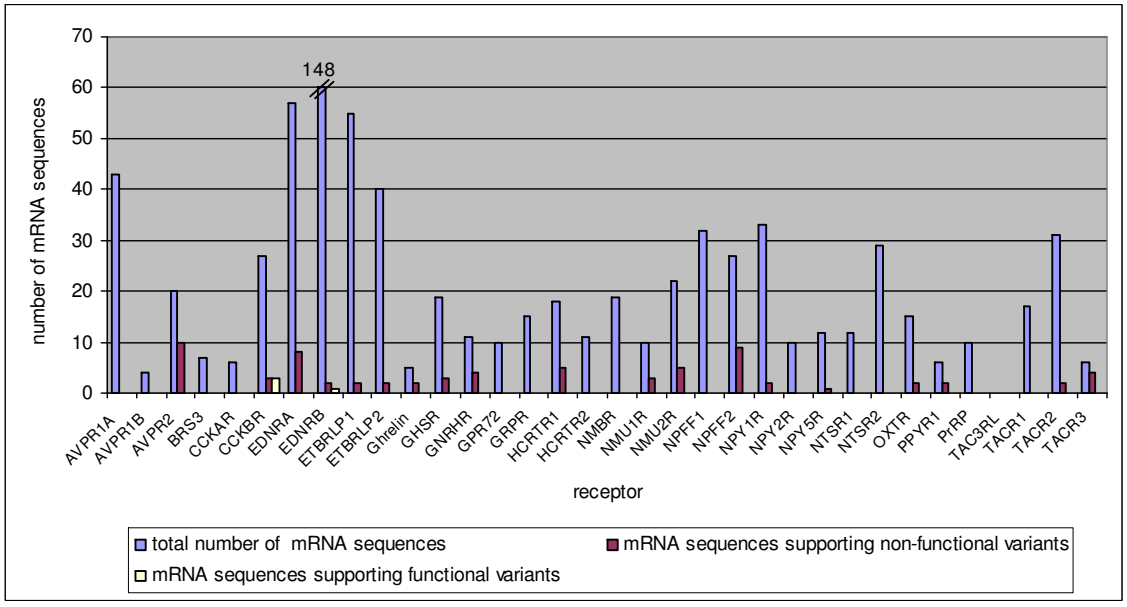


Figure 6b. The chart shows three bars for each receptor. They represent the total number of mRNA sequences, the number of mRNA sequences supporting non-functional variants and the number of mRNA sequences supporting functional variants. The broken column for EDNRB has its correct value written above it, this procedure improved the resolution for the other columns.

Variant Sequences and Authenticity

As mentioned earlier, 19 receptors have one splice variant or more, which implies that 15 receptors are without a variant. This indicates that it is more common to have a splice variant than not (see figure 7a). However, in most cases the receptors each have a small number of variants, like one or two. This means that the incidence of receptors having several different variants is not very common.

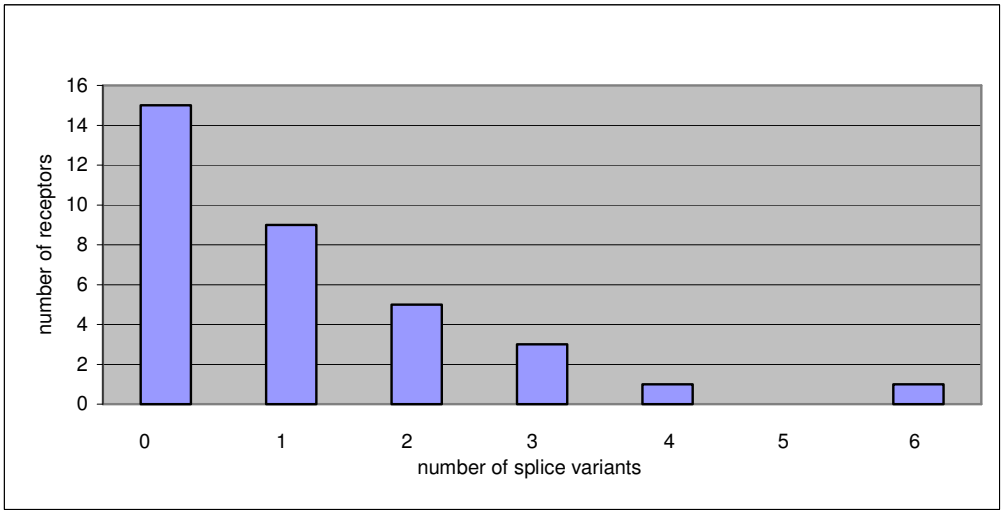


Figure 7a. The chart shows how many receptors that have a certain number of splice variants. For example, 15 genes have no splice variants while nine genes have one splice variant (etc.).

A result that strengthens the authenticity of many variants is shown in figure 7b. Here it is seen that there is a large number of variants supported by two sequences. Since the number of supporting sequences influence the probability of a splice variant being the result of a sequencing error, the fact that more than one sequence supports a splice variant increases the likeliness of that variant being genuine. Such a large proportion of variants being supported by two sequences are therefore an important result.

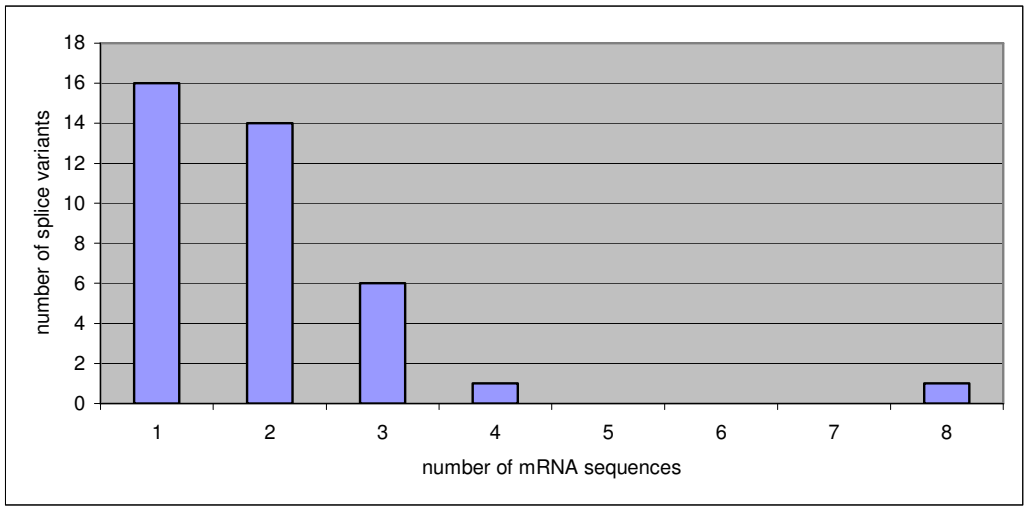


Figure 7b. The chart shows how many splice variants that are supported by a certain number of mRNA sequences each. For example, 16 splice variants are supported by one mRNA sequence each while 14 splice variants are supported by two mRNA sequences each (etc.)

The assessment of library origin of the mRNA sequences from a total of 38 variants shows that from the 22 variants supported by two sequences or more there are (see figure 8):

- Eight variants supported by mRNA sequences from the same library (within each variant) of which four variants have all of their mRNA sequences from normalized libraries.
- Ten variants supported by mRNA sequences from different libraries (within each variant) of which two variants have one of their mRNA sequences from normalized libraries.
- Four variants supported by mRNA sequences of which one or several sequences has unknown library origin.

Further it shows that from the 16 variants supported by one sequence there is only one variant supported by a mRNA sequence from a normalized library.

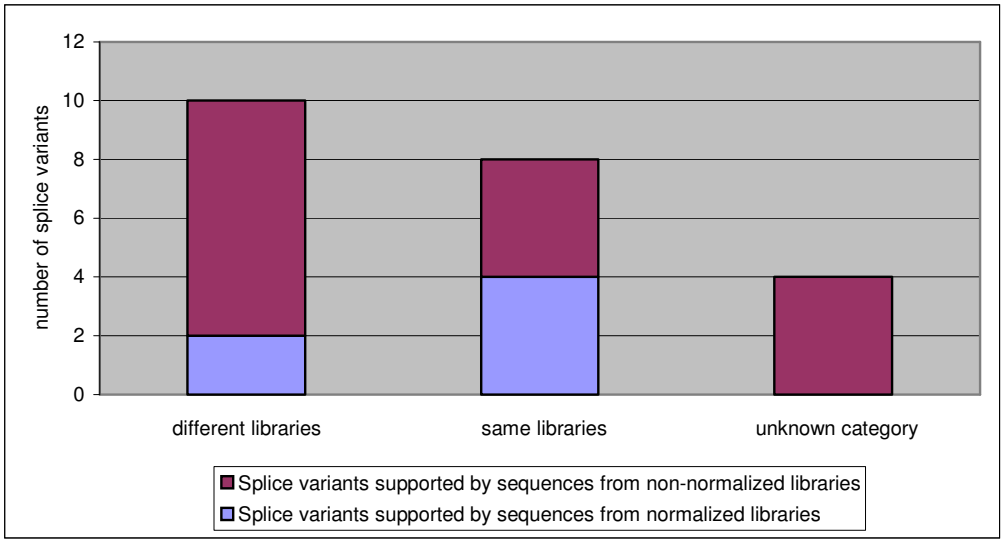


Figure 8. The chart shows three columns that represent the number of variants that are supported by mRNA sequences from different libraries, the same libraries or belong to an unknown category. The latter means that one or several mRNA sequences supporting the same variant are of unknown library origin and therefore these variants cannot be placed in any of the other categories. The starting point was the 22 variants supported by two mRNA sequences or more. Like the legend in the chart explains the columns also show to what extent the splice variants are supported by sequences from normalized libraries.

Discussion

Although the β subgroup in the *Rhodopsin* family represents a small dataset it is still big enough to hint the approximate number of splice variants that can be expected from such a dataset as well as to indicate to what extent these variants are functional. The method used was efficient and straightforward. The splice variants were detected with human BLAT searches and variants with genuine splice sites were selected in accordance with the splicing consensus sequence. The protein structure modelling technique with the procedure of first establishing the correct reading frame and then using the bovine rhodopsin sequence for modelling was also successful except for those variants that did not have a single exon matching the template. This special situation arose for four variants. The technique is based on finding the correct reading frame when aligning the variant to the template. If there is no variant exon aligning to the template there is no way of knowing which of the three possible reading frames is correct. When this problem appeared a translation in all three possible reading frames was performed. The result was that the three possible amino acid sequences contained stop codons in all four cases. The modelling technique could not give the protein structure of these variants but the translation implies that these variants are not functional.

There is no obvious correlation between the total number of mRNA sequences and the number of variant sequences for each receptor. If the alternative splicing mechanism was acting by random it would be logic if the receptors with the largest total number of mRNA sequences had the largest number of variant sequences. The fact that this is not the case is best demonstrated by receptors like AVPR1A, EDNRB, ETBRLP1 and ETBRLP2 in figure 6b. This result suggests that there is an unknown factor influencing the number of

variants for the receptors. With this material as the only source of information it is difficult to determine on what level this factor operates. One possible cause can be the selection of tissues. There is a possibility that the alternative RNA splicing mechanism is connected to certain types of tissues. Then, depending on the selection of tissue types, there would be a large or small amount of variant sequences for every receptor, irrespective of the receptors total number of mRNA sequences. A different explanation could be that the diverse receptors have differing requirements for splice variants, functional as well as non-functional. If splice variants play an important role from a physiological standpoint than it is probable that the different functions of the receptors result in varying needs for splice variants.

The four mRNA sequences displaying functionality represent two variants for separate receptors, CCKBR and EDNRB. Regarding CCKBR the variant presents an extended third cytoplasmic loop. This same variant has also been reported in four published articles and their results are based on "wet" work (Ding W.Q. et al., 2002; Schmitz F. et al., 2001; Zhou J. et al., 2004; Biagini P. et al., 1997). The protein structure in the articles is in accordance with the results in this report, the retained fourth intron results in 69 additional amino acids within the third cytoplasmic loop (Schmitz F. et al., 2001). It is further stated that this misspliced variant is only present in pancreatic cancer cells and "[…] may play an important role in the development and progression of this highly lethal cancer […]" (Ding W.Q. et al., 2002).

Regarding EDNRB the variant presents a shorter C-terminal with 38 diverging amino acids compared to the template. This variant is also found in the literature and the protein structure is identical to the one presented in this report. In one article the commonness in the alteration of the cytoplasmic domain of the receptor is emphasized. Since the cytoplasmic domain is involved in activation of the receptor by ligand binding this could imply that the alternative splicing mechanism is an important component in the production of physiologically differing receptor activity for the same ligand (Elshourbagy N.A. et al., 1996).

It is clear that these functional variants may serve a purpose in the sense that they still function as receptors. But what about all the variants that are considered to be non-functional, are they just the product of an error prone splicing mechanism? As there are articles on the functional variants found in this project there are also articles on many of the seemingly non-functional variants, based on "wet" data. The EDN-A receptor (EDNRA) has several variants with missing receptor domains (see appendix 1) and according to the literature these variants could be the product of a regulative RNA splicing mechanism (Miyamoto Y. et al., 1996). This means that the alternative splicing could contribute to the regulation of EDN-A gene expression. Another example of a non-functional variant referred to in the literature is a splice variant of the GNRH receptor (GNRHR). It has been established that the variant is incapable of ligand binding and signal transduction (Grosse R. et al., 1997). However, in the same article it is shown that this truncated variant has an impairing effect on the signaling function of the template receptor, which is concluded to be a "[…] principle of specific functional inhibition of G protein-coupled receptors […]" (Grosse R. et al., 1997). This negative effect of the splice variant may lie in its presence impairing proper plasma membrane expression of the template receptor, resulting in suppressed signaling capability (Grosse R. et al., 1997). Zhu and Wess (1998) have observed a similar phenomenon for the AVP-2 receptor (AVPR2). This receptor has a variant lacking the seventh TM region resulting in its inability to function as a proper receptor (see appendix 1). Zhu and Wess (1998) show that this truncated receptor (as well as others with at least three TM regions retained) can act as a negative regulator of template receptor function by the formation of a heterodimer. These and several other

articles all come to the same conclusion: seemingly non-functional variant receptors serve a purpose, not as receptors per se but as regulative factors controlling the amount and function of functional receptors.

The resulting ratio between mRNA sequences representing seemingly non-functional variants and mRNA sequences representing functional variants in this study is striking. From a total of 74 mRNA sequences representing variants there are only four sequences that display functionality, meaning they encode receptors with a complete seven TM region. The reason for this rare occurrence of functional splice variants could lie in the role that seemingly non-functional variants play. This overwhelming amount of seemingly non-functional variants could therefore be the result of a greater need for these regulative factors controlling the amount of template receptors than the requirement for variants diversifying the receptor function. The latter referring to the discussion made by Elshourbagy N.A. et al. (1996) (see above) based on the principle that the variants are still functional according to the definition but display different qualities in any of the functions in the GPCR signaling pathway.

Another result in accordance with this argument is the commonness in a receptor having one or two splice variants instead of several (see figure 5a). This implies that the purpose of the splice variants is not dependent on there being several different variants; obviously one or two is sufficient. If there is a specific technique to impair the template function, which has been demonstrated by Grosse et al. (1997) and Zhu and Wess (1998), then the structure of the truncated receptor to do this also should be specific to some extent.

An important result that gives support to the authenticity of many of the splice variants presented in this report is the large share of variants represented by two mRNA sequences. This means that a specific variant has been found in two different mRNA sequences, the risk of that variant being the result of a sequencing error is therefore small. However, a factor that must not be forgotten is the library origin of mRNA sequences supporting the same variant. If the sequences have identical library origin then there is a possibility that the variant they display is just a product of sequencing or cloning errors present only in that particular library. Sequences with different library origin displaying the same variant have appeared independently and therefore certifies the authenticity of the splice variant. An additional dimension to the authenticity issue is the occurrence of mRNA sequences coming from normalized libraries. Since abundant genes have been reduced in a normalized library it is easier to discover rare genes but there is a risk that these seemingly genuine gene variants are just products of sequencing errors. When a study is targeted on gene variants, like this one, this factor must be considered. The results show that even though there are many variants represented by sequences from the same libraries there are more variants supported by sequences of different library origins. In addition, most of these sequences do not come from normalized libraries. This means that even when these factors of uncertainty have been included in the study many of the variants can be considered genuine. But this does not mean that all of the variants that are supported by just one mRNA sequence or by sequences coming from normalized libraries should be seen as non-reliable, what it does mean is that they should be further investigated before it can be established that they truly are genuine splice variants. It should not be forgotten that one of the two functional splice variants (EDNRB) in this report, which has been referred to in the literature, is only supported by one sequence in the used dataset of mRNA sequences.

Conclusions

The method used to identify and analyse splice variants can be considered successful. It is shown that for the β subgroup in the *Rhodopsin* family it is more common to have a splice variant than not. In most cases the variant is seemingly non-functional, meaning the variant does not display a complete seven TM region, but this should not be mistaken as the result of an error prone RNA splicing mechanism. On the contrary, the total large number of seemingly non-functional variants and the small number of these variants for each receptor implies that they actually are the products of a controlled splicing mechanism and therefore serve a purpose. This is in accordance with the literature where there are several articles suggesting this purpose to be as regulative factors controlling the amount and function of so-called functional receptors. Since this is a circumstance that broadens the complexity of the GPCR signaling function it should be further investigated for larger datasets.

References

Published Articles

Biagini P., Monges G, Vuaroqueaux V, Parriaux D, Cantaloube JF, De Micco P., (1997), "The human gastrin/cholecystokinin receptors: type B and type C expression in colonic tumors and cell lines", Life Science, volume 61, number 10, 1009-1018. (1)

Bourgeois C., Robert B., Rebourcet R., Mondon F., Mignot T.M., Duc-Goiran P., Ferre F.J., (1997), "Endothelin-1 and ETA receptor expression in vascular smooth muscle cells from human placenta: a new ETA receptor messenger ribonucleic acid is generated by alternative splicing of exon 3". Clinical Endocrinology Metabolism, volume 82, number 9, 3116-3123. (2)

Brandenberger R., Wei H., Zhang S., Lei S., Murage J., Fisk G.J., Li Y., Xu C., Fang R., Guegler K., Rao M.S., Lebkowski J, Stanton L.W., (2004), "Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation", National Biotechnology, volume 22 number 6, 707-716. (3)

Cochet O., Heard D.J, Fehlbaum P., Ducray C., Bracco L., (2003), "Exploiting Human Genomic Diversity Through Alternative RNA Splicing", PharmaGenomics, January, 26-36. (4)

Dias Neto E., Garcia Correa R., Verjovski-Almeida S., Briones M.R., Nagai M.A., da Silva W. Jr., Zago M.A., Bordin S., Costa F.F., Goldman G.H., Carvalho A.F., Matsukuma A., Baia G.S., Simpson D.H., Brunstein A., deOliveira P.S., Bucher P., Jongeneel C.V., O'Hare M.J., Soares F., Brentani R.R., Reis L.F., de Souza S.J., Simpson A.J., (2000), "Shotgun sequencing of the human transcriptome with ORF expressed sequence tags", Proceedings of the National Academy of Science (USA), volume 97, number 7, 3491-3496. (5)

Ding W.Q., Kuntz S.M., Miller L.J., (2002), "A misspliced form of the cholecystokinin-B/gastrin receptor in pancreatic carcinoma: role of reduced cellular U2AF35 and a suboptimal 3'-splicing site leading to retention of the fourth intron", Cancer Research, volume 62, number 3, 947-952. (6)

Eckardt, N., (2004), "Abscisic Acid Signal Transduction: Function of G Protein-Coupled Receptor 1 in Arabidopsis", Plant Cell, volume 16, 1353-1354. (7)

Elshourbagy N.A., Adamou J.E., Gagnon A.W, Wu H.L., Pullen M., Nambi P., (1996), "Molecular characterization of a novel human endothelin receptor splice variant", Journal of Biological Chemistry, volume 271, number 41, 25300-25307. (8)

Fredriksson R., Lagerström M.C, Lundin L.G., Schiöt H.B., (2003), "The G-protein-Coupled Receptors in the Human Genome Form Five Families. Phylogenetic Analysis, Paralogon Groups and Fingerprints", Molecular Pharmacology, volume 63, number 6, 1257-1272 (9)

Grosse R., Schoneberg T., Schultz G., Gudermann T., (1997), "Inhibition of gonadotropin-releasing hormone receptor signaling by expression of a splice variant of the human receptor", Molecular Endocrinology, volume 11, number 9, 1305-1318. (10)

Harrington J.J., Sherf B., Rundlett S., Jackson P.D., Perry R., Cain S., Leventhal C., Thornton M., Ramachandran R., Whittington J., Lerner L., Costanzo D., McElligott K., Boozer S., Mays R., Smith E., Veloso N., Klika A., Hess J., Cothren K., Lo K., Offenbacher J., Danzig J., Ducar M., (2001), "Creation of genome-wide protein expression libraries using random activation of gene expression", National Biotechnology, volume 19 number 5, 440-445. (11)

Herzog H., Baumgartner M., Vivero C., Selbie L.A., Auer B., Shine J., (1993), "Genomic organization, localization, and allelic differences in the gene for the human neuropeptide Y Y1 receptor", Journal of Biological Chemistry, volume 268, number 9, 6703-6707. (12)

Howard A.D., Feighner S.D., Cully D.F., Arena J.P., Liberator P.A., Rosenblum C.I., Hamelin M., Hreniuk D.L., Palyha O.C., Anderson J., Paress P.S., Diaz C., Chou M., Liu K.K., McKee K.K., Pong S.S., Chaung L.Y., Elbrecht A., Dashkevicz M., Heavens R., Rigby M.,Sirinathsinghji D.J.S., Dean D.C., Melillo D.G., Patchett A.A., Nargund R., Griffin P.R., DeMartino J.A., Gupta S.K.,Schaeffer J.M., Smith R.G. and Van der Ploeg L.H.T., (1996),
"A receptor in pituitary and hypothalamus that functions in growth hormone release", Science, volume 273, number 5277, 974-977. (13)

Jeffery P.L., Herington A.C., Chopin L.K., (2002), "Expression and action of the growth hormone releasing peptide ghrelin and its receptor in prostate cancer cell lines", Journal of Endocrinology, volume 172, number 3, R7-11. (14)

Jin P., (2004), "PCR isolation and cloning of novel splice variant mRNAs from known drug target genes", Genomics, volume 83, number 4, 566-571. (15)

Jin P., Fu G.K., Wilson A.D., Yang J., Chien D., Hawkins P.R., Au-Young J., Stuve L.L. (2004), "PCR isolation and cloning of novel splice variant mRNAs from known drug target genes", Genomics, volume 83, number 4, 566-571. (16)

Kawakami B., Sugiyama A., Takemoto, (2004), "Complete sequencing and characterization of 21,243 full-length human cDNAs", National Genetics, volume 36, number 1, 40-45. (17)

Laemmle B., Schindler M., Beilmann M., Hamilton B.S., Doods H.N., Wieland H.A., (2003), "Characterization of the NPGP receptor and identification of a novel short mRNA isoform in human hypothalamus", Regulatory Peptides, volume 111, number 1-3, 21-29. (18)

Miyake A., (1995), "A truncated isoform of human CCK-B/gastrin receptor generated by alternative usage of a novel exon", Biochemical Biophysical Research Community, volume 208, number 1, 230-237. (19)

Miyamoto Y., Yoshimasa T., Arai H., Takaya K., Ogawa Y., Itoh H., Nakao K., (1996), "Alternative RNA splicing of the human endothelin-A receptor generates multiple transcripts", Journal of Biochemistry, volume 313, pt 3, 795-801. (20)

Monstein H.J., Nilsson I. Ellnebo-Svedlund K., Svensson S.P., (1998), "Cloning and characterization of 5'-end alternatively spliced human cholecystokinin-B receptor mRNAs", Receptors Channels, volume 6, number 3, 165-177. (21)

North W.G., Fay M.J., Longo K.A., Du J., (1998), "Expression of all known vasopressin receptor subtypes by small cell tumors implies a multifaceted role for this neuropeptide", Journal Cancer Research, volume 58, number 9, 1866-1871. (22)

North W.G. Fay M.J., Du J., (1999), "MCF-7 breast cancer cells express normal forms of all vasopressin receptors plus an abnormal V2R", Peptides, volume 20, number 7, 837-842. (23)

Palczewski K., Kumasaka T., Hori T., Behnke C.A., Motoshima H., Fox B.A., Le Trong I., Teller D.C., Okada T., Stenkamp R.E., Yamamoto M., Miyano M., (2000), "Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor", Science, volume 289, number 5480, 739-745. (24)

Ponnal N., Nambi A., (2003), "G Protein-Coupled Receptors in Drug Development", Assay and Drug Development Technologies, volume 1, number 2, 305-310. (25)

Reddy A.R., Ramakrishna W., Sekhar C.A., Ithal N., Babu R.P., Bonaldo M.F., Soares M.B., and Bennetzen J.L., (2002), "Novel genes are enriched in normalized cDNA libraries from drought-stressed seedlings of rice (*Oryza sativa* L. subsp. *indica* cv. Nagina 22)", Genome, volume 45, number 1, 204-211. (26)

Schmitz F., Schrader H., Otte J., Schmitz H., Stuber E., Herzig K., Schmidt WE., (2001), "Identification of CCK-B/gastrin receptor splice variants in human peripheral blood mononuclear cells", Regulatory Peptides, volume 101, number 1-3, 25-33. (27)

Seibold A., Brabet P., Rosenthal W., Birnbaumer M., (1992), "Molecular cloning of the receptor for human antidiuretic hormone", Nature, volume 357, number 6376, 333-335. (28)

Strausberg R., (2002), "Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences", Proceedings of the National Academy of Science (USA), volume 99, number 26, 16899-16903. (29)

Suzuki Y., Yamashita R., Shirota M., Sakakibara Y., Chiba J., Mizushima-Sugano J., Nakai K., Sugano S., (2004), "Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions", Genome Research, volume 14, number 9, 1711-1718. (30)

Tashiro H., Yamazaki M., Watanabe K., (2004), "Complete sequencing and haracterization of 21,243 full-length human cDNAs**",** National Genetics, volume 36, number 1, 40-45. (31)

Yamashita J., Yoshimasa T., Arai H., Hiraoka J., Takaya K., Miyamoto Y., Ogawa Y., Itoh H., Nakao K., (1995), “Cis elements for transcriptional regulation of the human endothelin-A receptor gene”, Journal of Cardiovascular Pharmacology, volume 26, suppl 3, 26-8. (32)

Zhou J., Hu J., Chen Z., Wang W., Zhang Q., Chen M., (2004), “Human gastric tissues coexpress two different splicing cholecystokinin-B/gastrin receptors”, Sheng Wu Yi Xue Gong Cheng Xue Za Zhi, volume 21, number 3, 440-443. (33)

Zhu X., Wess J., (1998), “Truncated V2 vasopressin receptors as negative regulators of wild-type V2 receptor function”, Biochemistry, volume 37, number 45,15773-15784. (34)

Technical Reports

Andersson S.G.E., (2004), Bioinformatics –A New Multidisciplinary Tool, Department of Molecular Evolution, Uppsala University.

Boguski M.S., (1995), National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA.

Gloriam D., (2004), unpublished, Department of Neuroscience, Unit of Pharmacology, Uppsala University.

Books

Lodish H., Berk A., Zipursky L.S., Matsudaira P., Baltimore D., Darnell J., (2000), Molecular Cell Biology, 4:th ed., W.H Freeman and Company, New York.

Internet

Navigant Consulting Inc.,
www.navigantconsulting.com/lifesciences/SMR/gpcr/gpcrSP.pdf, (2005-01-09)

Canada Research, www.chairs.gc.ca/web/chairholders, (2005-01-10)

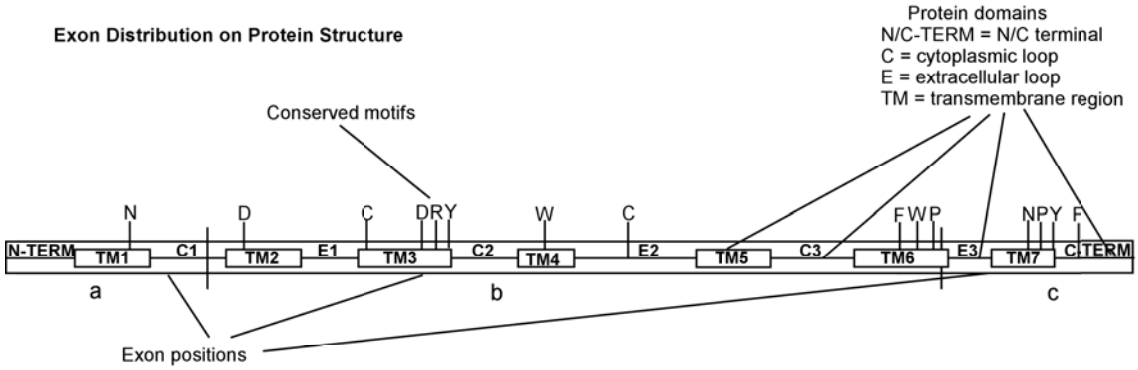
University of Washington,
depts.washington.edu/uweek/archives/2000.10.OCT_19/article16.html, (2005-01-10)

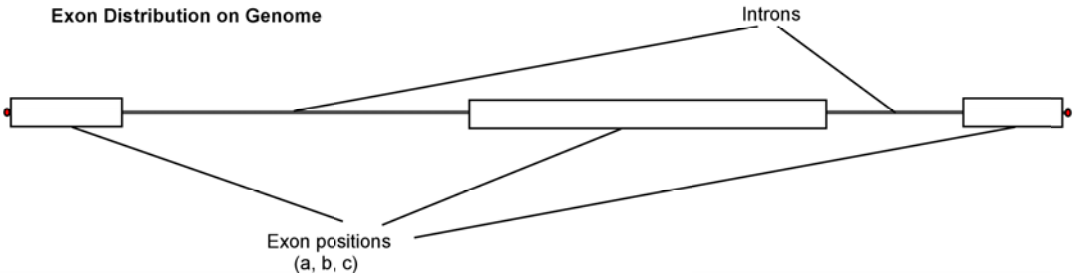
The Prostate Expression Database, www.pedb.org/PEDB/OVERVIEW/introduction.html, (2005-01-11)

NCBI, www.ncbi.nlm.nih.gov/About/primer/est.html, (2005-01-12)

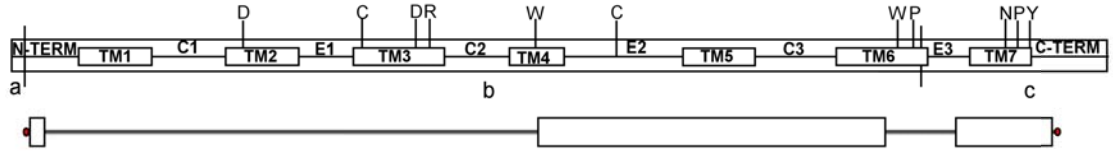
Appendix 1

The imaginary figures below demonstrate how the schematic pictures of the templates and splice variants are organized. First there is a picture of the exon distribution on the protein structure and below on the genome. For every receptor the template comes first and then follows the splice variants, on the bottom of each page there is a list of all the used signs.



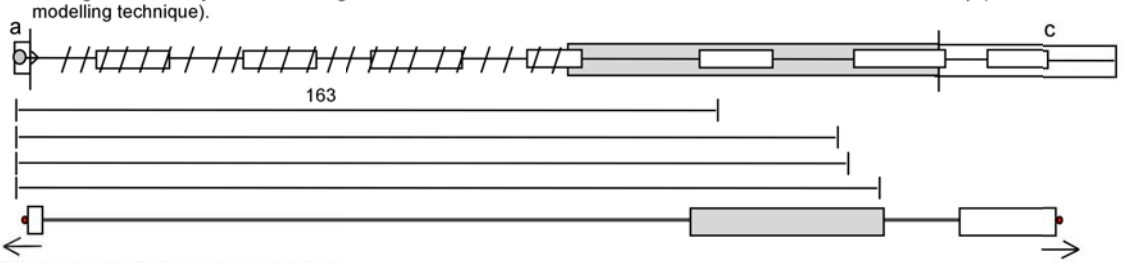


Exon Distribution on Protein Structure and on Genome for AVPR2 and Splice Variants



The recognized wild type. It is presented by 4 full length mRNA sequences (from GenBank) and by 8 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.

A variant that is supported by 2 mRNA sequences (from GenBank): AF032388 from tissue of small-cell carcinoma of the lung and AF101728 from tissue of breast cancer. The number beneath the extended loop represents the number of amino acids making up the loop. The frame shift results in diverging amino acids and stop codons. In this case the extended loop is placed in the middle of a TM-region which may seem confusing but it is done so to illustrate that here is where the extra amino acids end up (with this modelling technique).



A variant that is supported by a total of 8 sequences: 4 full length mRNA sequences (from GenBank) and 4 EST sequences of different lengths. The mRNA sequences are: BC015746, BC033090, BC041642, BC080603 and are all from the tissue of epithelioid carcinoma from pancreas. The EST sequences are: BG830436, BI161438, BI161076, BI160709 and are also from the tissue of epithelioid carcinoma from pancreas. The lines beneath the exon-protein structure represent the coverage of the EST sequences, the line closest to the structure is the first EST mentioned in the text above, the second line is the second EST mentioned and so on. The number in the figure represents the number of amino acids that are missing. The frame shift results in diverging amino acids.

= exon wild type (a,b,c etc.)

= exon splice variant

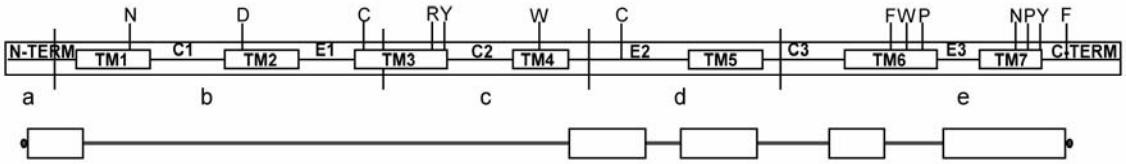
/ = missing amino acids

= frame shift

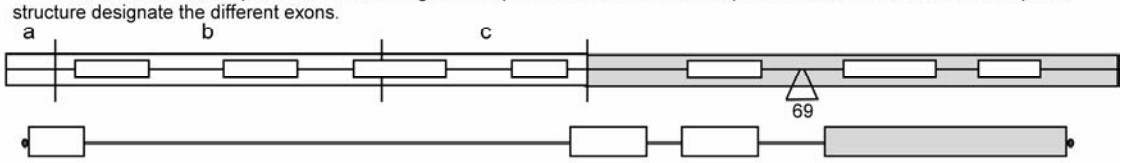
= extended loop with stop codon

= the exon of the variant goes further than the coding region of the wild type

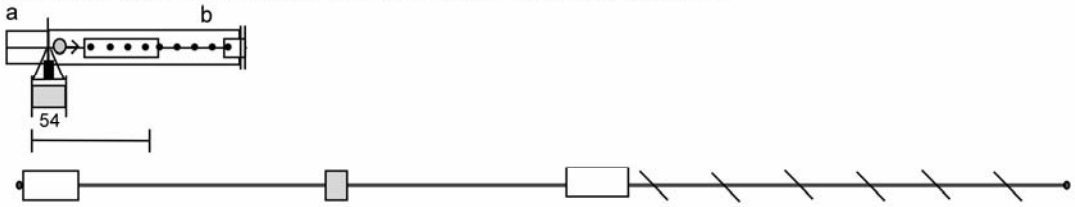
Exon Distribution on Protein Structure and on Genome for CCKBR and Splice Variants



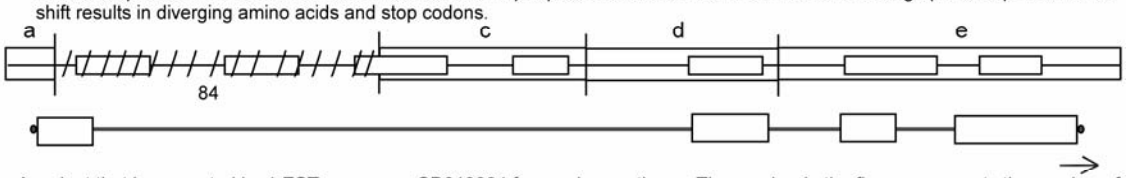
The recognized wild type. It is presented by 8 full length mRNA sequences (from GenBank) and 10 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 3 full length mRNA sequences (from GenBank): AF239668 from tissue of colorectal cancer and premalignant polyps, AF441129 from tissue of pancreatic carcinoma cells and AY029770 from unknown tissue. The number beneath the extended loop represents the number of amino acids making up the loop.



A variant that is supported by 2 mRNA sequences (from GenBank) of different lengths: Y13464 from unknown tissue (the longer one) and S76072 from tissue of the stomach (the shorter one). The line beneath the protein figure represent the coverage of the shorter sequence. The number beneath the extended loop represents the number of amino acids making up the loop. The frame shift results in diverging amino acids and stop codons.



A variant that is supported by 1 EST sequence: CD013884 from unknown tissue. The number in the figure represents the number of amino acids that are missing.

= exon wild type (a,b,c etc.)

= exon splice variant

= extended loop

= extended loop with stop codon

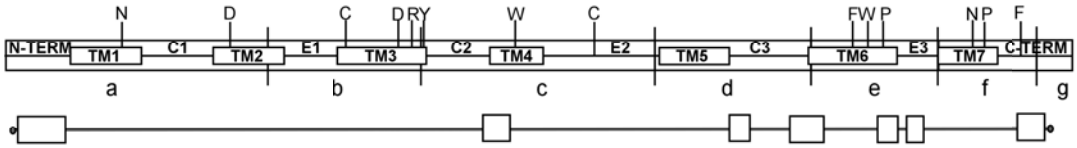
= end of sequence due to sequencing

= frame shift

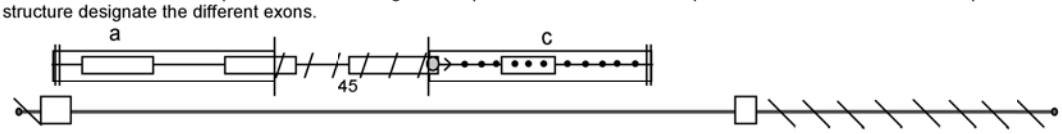
/ = missing amino acids

→ = the exon of the variant goes further than the coding region of the wild type

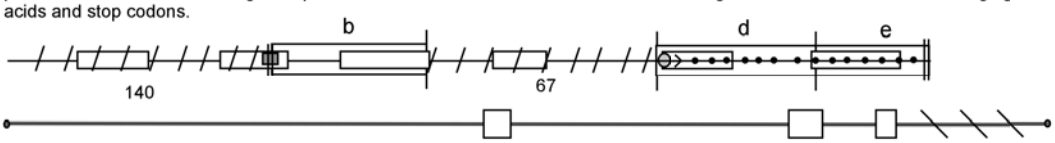
Exon Distribution on Protein Structure and on Genome for EDNRA and Splice Variants



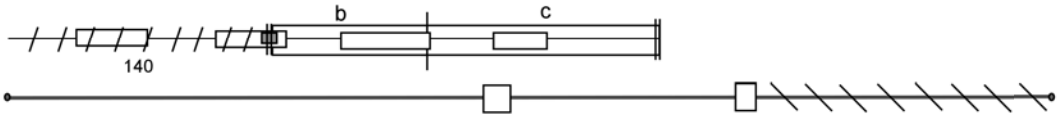
The recognized wild type. It is presented by 10 full length mRNA sequences (from GenBank) and 33 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



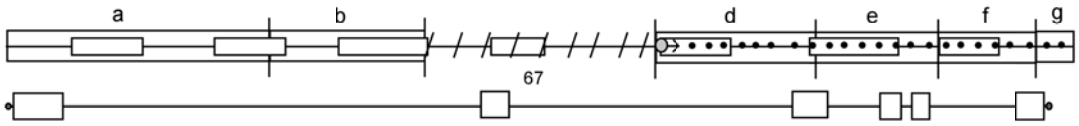
A variant that is supported by 1 mRNA sequence (from GenBank): AF014826 from the tissue of vascular smooth muscle cells from placenta. The number in the figure represents the number of amino acids that are missing. The frame shift results in diverging amino acids and stop codons.



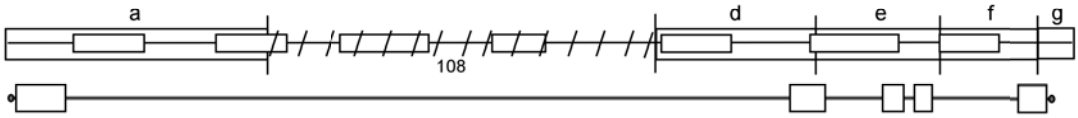
A variant that is supported by 1 EST sequence: BI460633 from tissue of testis. The numbers in the figure represent the number of amino acids that are missing. The frame shift results in diverging amino acids and stop codons.



A variant that is supported by 1 EST sequence: CN313269 from tissue of embryonic stem cell (embryoid bodies derived from H1, H7 and H9 cells). The number in the figure represents the number of amino acids that are missing.



A variant that is supported by 1 mRN sequence (from GenBank): S81542 from unknown tissue. The number in the figure represents the number of amino acids that are missing. The frame shift result in diverging amino acids and stop codons.



A variant that is supported by 1 mRNA sequence (from GenBank): S81545 from tissue of lung. The number in the figure represents the number of amino acids that are missing.

A variant that is supported by 3 mRNA sequences (from Genbank) of different lengths: BX537573 from tissue of cervix, AK123169 from tissue of cerebellum and G36538 from unknown tissue. As noticed only the exon-genome figure is displayed for this variant. The reason for this is that as the figure shows the exon of the variant doesn't align with any of the exons of the wild type. Because of this, the method used can't give the reading frame of interest and consequently not the resulting exon-protein structure. However, translation of the gene sequence of the variant in all three possible reading frames result in diverging amino acids and scattered stop codons. Therefore, one can conclude that this variant doesn't give rise to a functional receptor.

= exon wild type (a,b,c etc.)

= exon splice variant

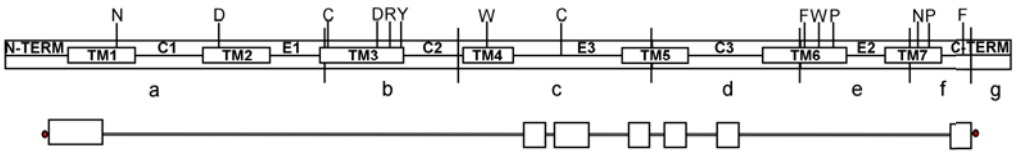
/ = missing amino acids

= end of sequence due to sequencing

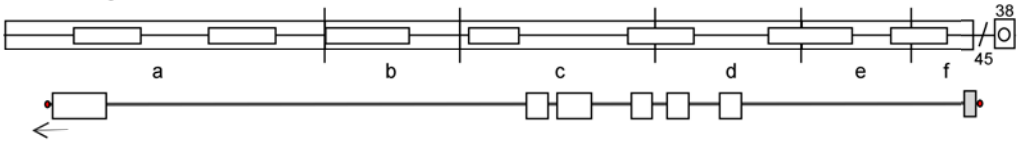
= frame shift

= end of sequence due to alternative splicing

Exon Distribution on Protein Structure and on Genome for EDNRB and Splice Variants



The recognized wild type. It is presented by 11 full length mRNA sequences (from GenBank) and 116 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 1 EST sequence: X99250 from tissue of placenta. The number beneath the exon-protein structure represents the number of amino acids that are missing. The number above the exon-protein structure represents the number of amino acids that are diverging. As the figure shows this variant forms a full 7 TM-region.

A variant that is supported by 1 mRNA secuence (from GenBank): BC031243 from tissue of testis and by 1 shorter EST sequence: BG698082 from tissue of skin. As noticed only the exon-genome figure is displayed for this variant. The reason for this is that as the figure shows the exon of the variant doesn't align with any of the exons of the wild type. Because of this, the method used can't give the reading frame of interest and consequently not the resulting exon-protein structure. However, translation of the gene sequence of the variant in all three possible reading frames result in diverging amino acids and scattered stop codons. Therefore, one can conclude that this variant doesn't give rise to a functional receptor.

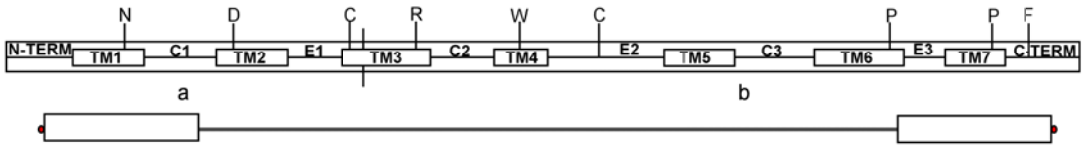
= exon wild type (a,b,c etc.)

= exon splice variant

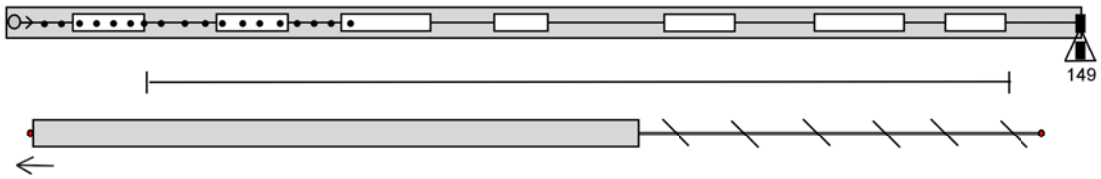
→ = the exon of the variant goes further than the coding region of the wild type

o = diverging amino acids

Exon Distribution on Protein Structure and on Genome for ETBRLP1 and Splice Variant



The recognized wild type. It is presented by 4 full length mRNA sequences (from GenBank), 1 short mRNA sequence (from GenBank) and 48 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 2 sequences of different lengths:1 mRNA sequence (from GenBank): BX649006 from tissue of retina and 1 EST sequence: BI547606 from tissue of hippocampus (shorter sequence). The line beneath the main figure represent the coverage of the shorter EST sequence. The frame shift results in diverging amino acids and stop codons. The number beneath the extended loop represents the number of amino acids and stop codons making up the loop.

= exon wild type (a,b,c etc.)

= exon splice variant

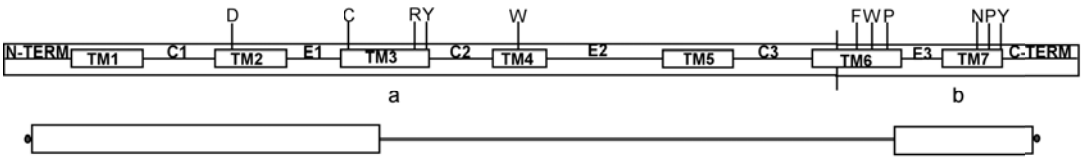
= extended loop with stop codon

= frame shift

= stop codon

= the exon of the variant goes further than the coding region of the wild type

Exon Distribution on Protein Structure and on Genome for GHRELIN and Splice Variant



The recognized wild type. It is presented by 1 full length mRNA sequence (from GenBank) and by 1 shorter EST sequence. The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.

A variant supported by 2 EST sequences of different lengths: BF513101 from unknown tissue and BU553576 from tissue of ovary teratocarcinoma. The exon-genom structure is shown for both sequences (BF513101 on top) since it is a special case. The lines beneath the exon-protein structure display the coverage of the two sequences (same order as exon-genom structure). In this case the extended loop is placed in the middle of a TM-region which may seem confusing but it is done so to illustrate that here is where the extra amino acids end up (with this modelling technique). The number beneath the extended loop represents the number of amino acids and stop codons making up the loop.

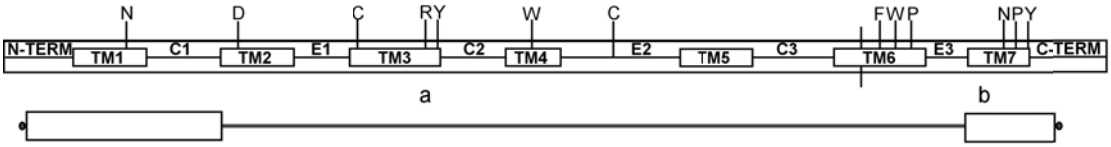
= exon wild type (a,b,c etc.)

= exon splice variant

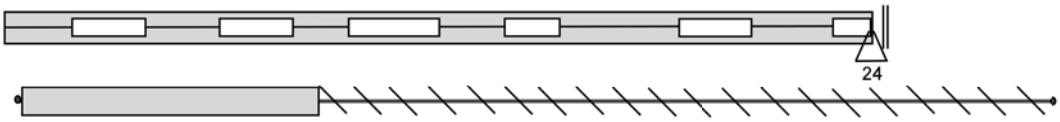
= extended loop with stop codon

= end of sequence due to sequencing

Exon Distribution on Protein Structure and on Genome for GHSR and Splice Variant



The recognized wild type. It is presented by 2 full length mRNA sequences (from GenBank) and 14 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor.
The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons



A variant that is supported by 3 mRNA sequences (from GenBank): BC069068 from unknown tissue, BC069374 from unknown tissue, U60181 from tissue of pituitary. In this case the extended loop is placed in the middle of a TM-region which may seem confusing but it is done so to illustrate that here is where the extra amino acids end up (with this modelling technique). The number beneath the extended loop represents the number of amino acids making up the loop.

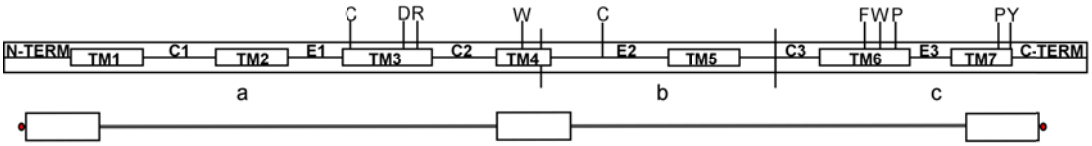
= exon wild type (a,b,c etc.)

= exon splice variant

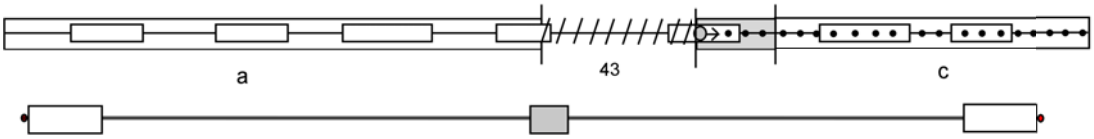
= end of sequence due to sequencing

= extended loop

Exon Distribution on Protein Structure and on Genome for GNRHR and Splice Variant



The recognized wild type. It is presented by 5 full length mRNA sequences and 2 EST sequences of different lengths (not full length). The letters (not the letters representing the protein domains) are the conserved motifs appearing in this particular receptor.



A variant that is supported by 1 mRNA sequence (from GenBank): Z81148 from tissue of pituitary. The number beneath the exon-protein structure represent the number of amino acids that are missing. The frame shift results in diverging amino acids in the variant exon and in diverging amino acids and stop codons in the c-exon.

A variant that is supported by 2 mRNA sequences (from GenBank): BC039430 and BC045560 both from tissue of hypothalamus. As noticed only the exon-genome figure is displayed for this variant. The reason for this is that as the figure shows the exon of the variant doesn't align with any of the exons of the wild type. Because of this, the method used can't give the reading frame of interest and consequently not the resulting exon-protein structure. However, translation of the gene sequence of the variant in all three possible reading frames result in diverging amino acids and scattered stop codons. Therefore, one can conclude that this variant doesn't give rise to a functional receptor.

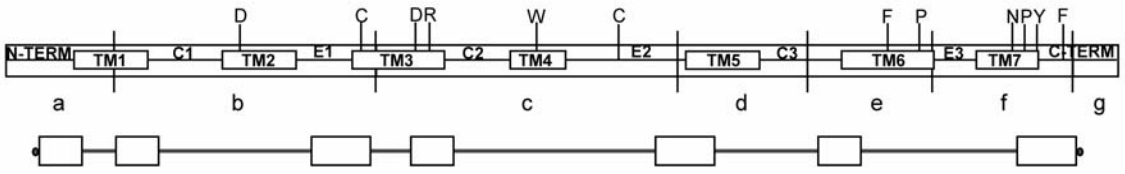
= exon wild type (a,b,c etc.)

= exon splice variant

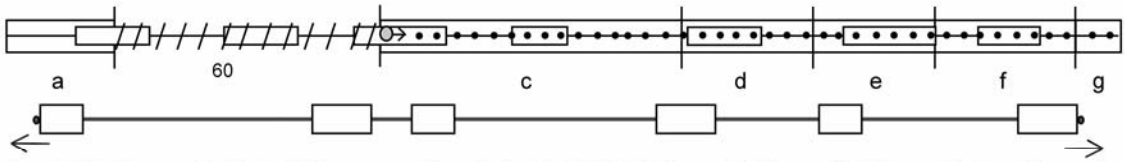
/ = missing amino acids

= frame shift

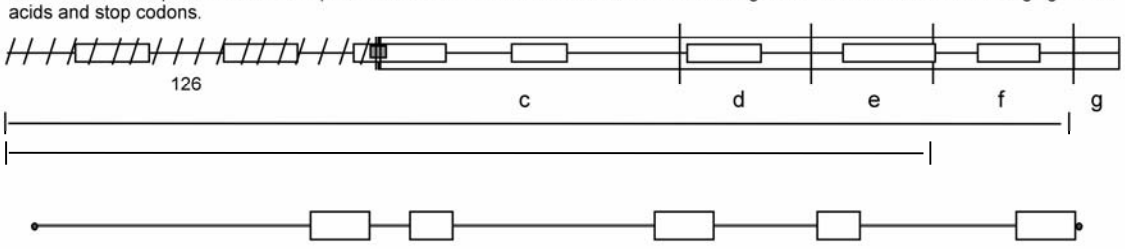
Exon Distribution on Protein Structure and on Genome for HCRTR1 and Splice Variants



The recognized wild type. It is presented by 2 full length mRNA sequences (from GenBank) and by 10 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 1 mRNA sequence (from Genbank): BC035686 from pooled tissue of fetal lung and spleen. The number beneath the exon-protein structure represents the number of amino acids that are missing. The frame shift results in diverging amino acids and stop codons.



A variant that is supported by 1 full length mRNA sequence (from GenBank): CR605321 from tissue of fetal brain and by 3 EST sequences of different lengths: BX433092, BX433093, AL535838 all from tissue of fetal brain. The coverage of the first two EST sequences is shown by the line nearest to the exon-protein structure and the coverage of the third EST by the line below it. The number beneath the exon-protein structure represents the number of amino acids that are missing.

= exon wild type (a,b,c etc.)

= exon splice variant

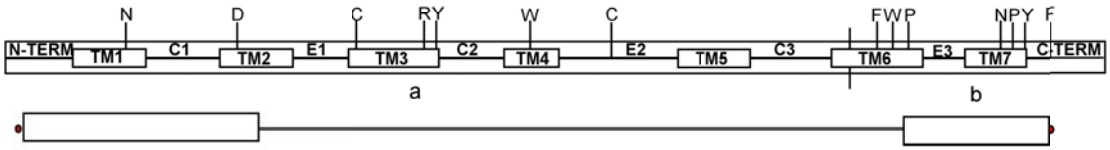
/ = missing amino acids

= end of sequence due to alternative splicing

= frame shift

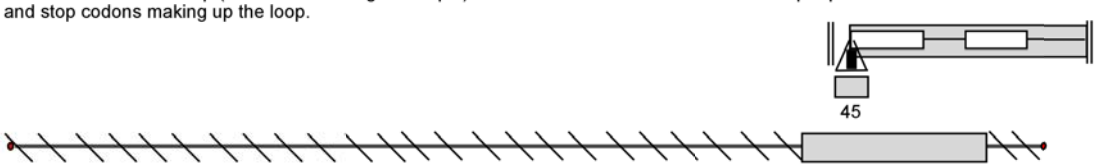
→ = the exon of the variant goes further than the coding region of the wild type

Exon Distribution on Protein Structure and on Genome for NMU1R and Splice Variants



The recognized wild type. It is presented by 3 full length mRNA sequences (from GenBank) and by 6 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.

A variant that is supported by 1 EST sequence: BX491295 from unknown tissue. It is one exon that gives rise to the exon-protein structure above; one part of it aligning with the a-exon and the other part corresponding to intron sequence of the wild type. In this case the extended loop is placed in the middle of a TM-region which may seem confusing but it is done so to illustrate that here is where the extra amino acids end up (with this modeling technique). The number beneath the extended loop represents the number of amino acids and stop codons making up the loop.



A variant supported by 1 EST sequence: CN386259 from tissue of embryonic stem cells (H1, H7 and H9 cells). It is one exon that gives rise to the exon-protein structure above; one part of it aligning with the b-exon and the other part corresponding to intron sequence of the wild type. In this case the extended loop is placed in the middle of a TM-region which may seem confusing but it is done so to illustrate that here is where the extra amino acids end up (with this modelling technique).The number beneath the extended loop represents the number of amino acids and stop codons making up the loop.

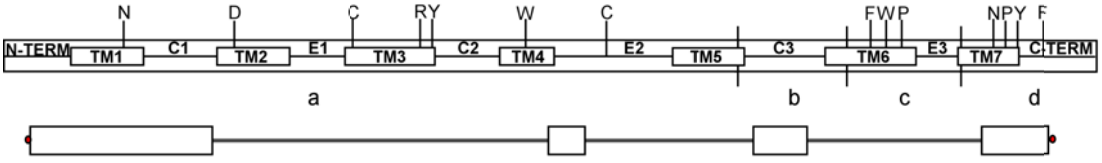
= exon wild type (a,b,c etc.)

= exon splice variant

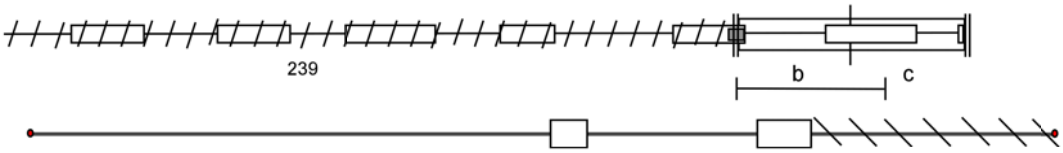
= extended loop with stop codon

= end of EST sequence due to sequencing

Exon Distribution on Protein Structure and on Genome for NMU2R and Splice Variant



The recognized wild type. It is presented by 6 full length mRNA sequences (from GenBank) and by 13 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



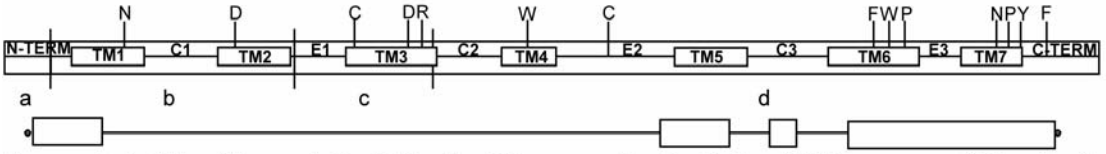
A variant that is supported by 2 EST sequences of different lengths: R13890 and H11359 both from tissue of infant brain. The number beneath the exon-protein structure represents the number of amino acids that are missing. The line beneath the exon-protein structure displays the coverage of the shorter EST sequence (H11359).

/ = missing amino acids

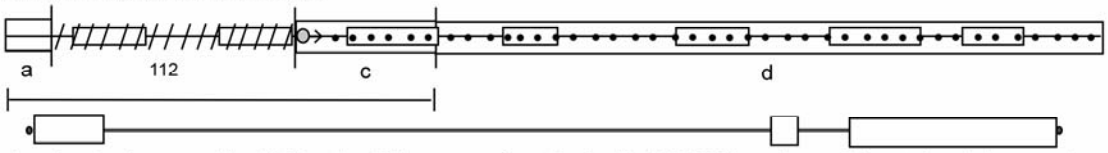
‖ = end of sequence due to sequencing

⫴ = end of sequence due to alternative splicing

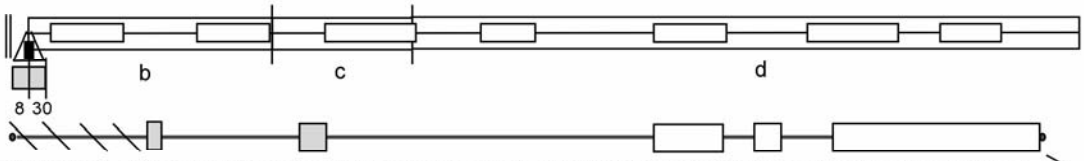
Exon Distribution on Protein Structure and on Genome for NPFF2 and Splice Variants



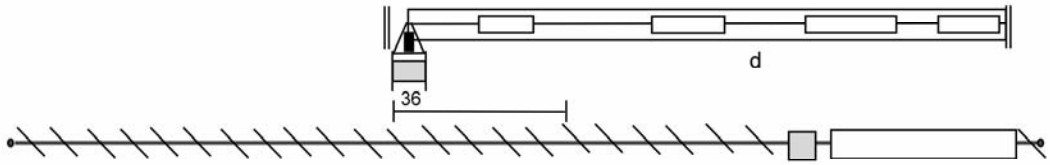
The recognized wild type. It is presented by 1 full length mRNA sequence (from GenBank) and 19 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



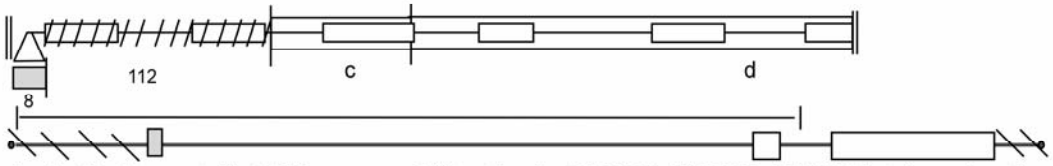
A variant that is supported by 1 full length mRNA sequence (from GenBank): AJ311393 from tissue of placenta/hypothalamus and by 1 EST sequence: CN285251 from tissue of embryonic stem cells (cell lines H1, H7, and H9). The line beneath the exon-protein structure displays the coverage of the EST sequence. The number beneath the exon-protein structure represents the number of amino acids that are missing. The frame shift results in diverging amino acids and stop codons.



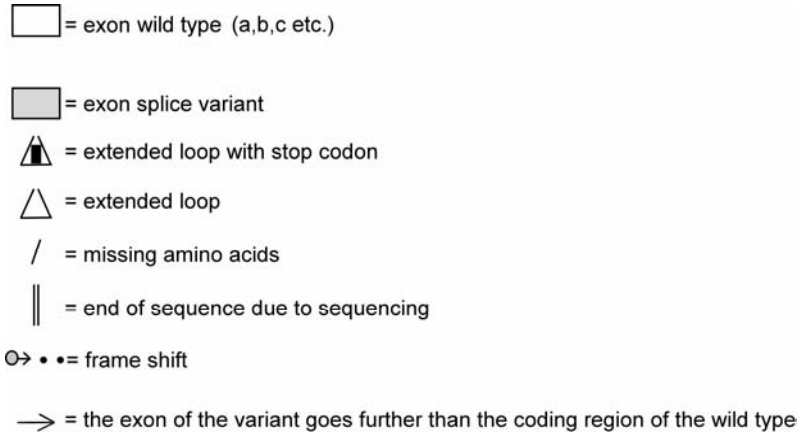
A variant that is supported by 1 mRNA sequence (from GenBank): AF236083 from unknown tissue. The numbers beneath the extended loop represent the number of amino acids (and stop codons) that the two additional exons are coding for, the exon on the right result in 30 amino acids and stop codons while the exon on the left result in 8 amino acids. It should be noticed that the exon on the left most likely is interrupted by premature sequencing termination.



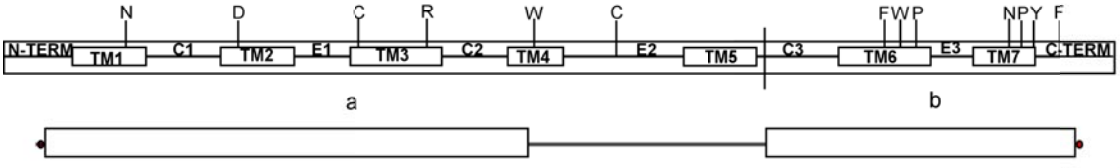
A variant that is supported by 2 EST sequences of different lengths: CB993428 and CB997680 both from tissue of pre-eclamptic placenta. The line beneath the exon-protein structure is displaying the coverage of the shorter EST sequence (CB993428). It should be noticed that the additional exon most likely is interrupted by premature sequencing termination. In this case the extended loop is placed in the middle of a TM-region which may seem confusing but it is done so to illustrate that here is where the extra amino acids end up (with this modelling technique). The number beneath the extended loop represents the number of amino acids and stop codons making up the loop.



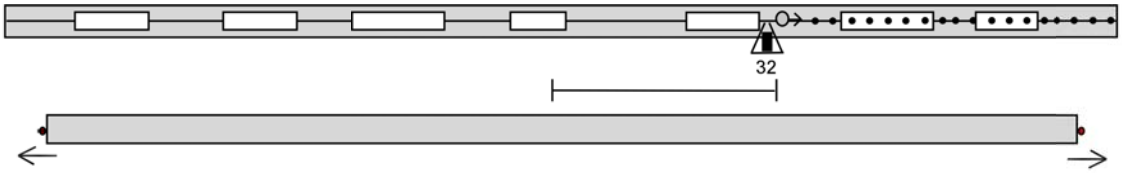
A variant that is supported by 3 EST sequences of different lengths: BX283068, BX280001, AA449919 that all come from tissue of fetus material. The line beneath the exon-protein structure displays the coverage of the two shorter EST sequences (BX283068, BX280001). The number beneath the extended loop represents the number of amino acids making up the loop. The number on the right represents the number of amino acids that are missing. It should be noticed that the variant exon on the left most likely is interrupted by premature sequencing termination.


= exon wild type (a,b,c etc.)
= exon splice variant
= extended loop with stop codon
= extended loop
= missing amino acids
= end of sequence due to sequencing
= frame shift
= the exon of the variant goes further than the coding region of the wild type

Exon Distribution on Protein Structure and on Genome for NPY1R and Splice Variant



The recognized wild type. It is presented by 5 full length mRNA sequences (from GenBank) and by 27 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 1 full length mRNA sequence (from GenBank): L07615 from tissue of peripheral blood and by 1 shorter EST sequence: BG195632 from unknown tissue. The number beneath the extended loop represents the number of amino acids and stop codons making up the loop. The frame shift results in diverging amino acids and stop codons. The line beneath the exon-protein structure displays the coverage of the EST sequence.

= exon wild type (a,b,c etc.)

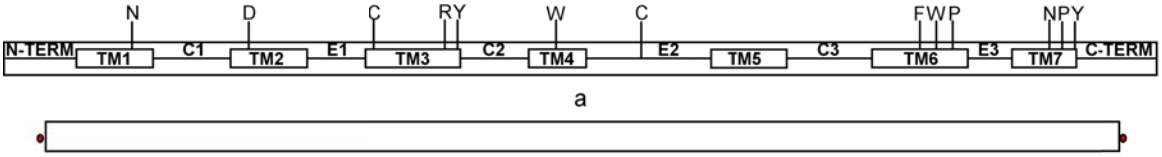
= exon splice variant

= extended loop with stop codon

= frame shift

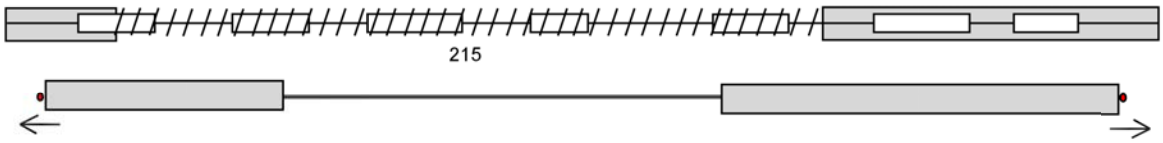
= the exon of the variant goes further than the coding region of the wild type

Exon Distribution on Protein Structure and on Genome for NPY5R and Splice Variant



a

The recognized wild type. It is presented by 5 full length mRNA sequences (from GenBank) and by 6 EST sequences of different lengths. The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 1 EST sequence: BG191732 from unknown tissue. The number beneath the exon-protein structure represents the number of amino acids that are missing.

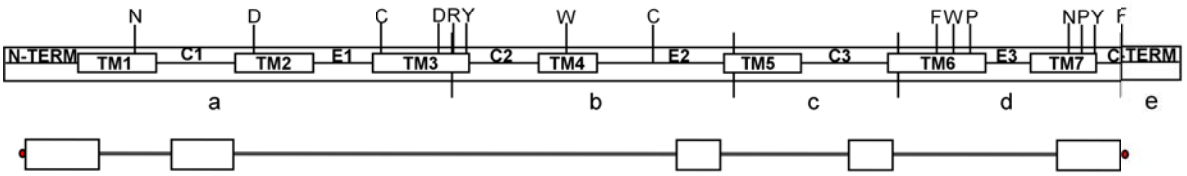
= exon wild type (a,b,c etc.)

= exon splice variant

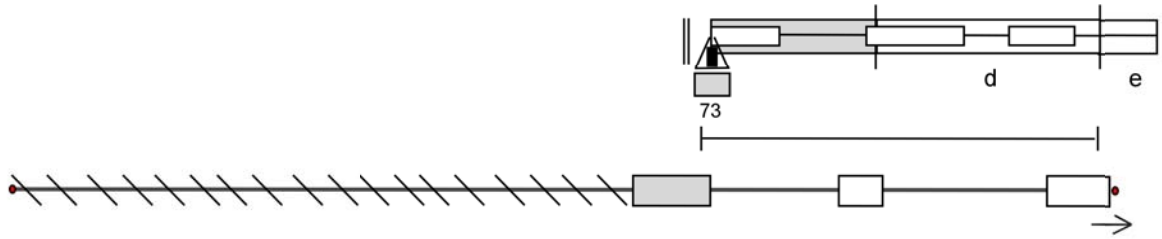
/ = missing amino acids

→ = the exon of the variant goes further than the coding region of the wild type

Exon Distribution on Protein Structure and on Genome for TACR2 and Splice Variant



The recognized wild type. It is presented by 3 full length mRNA sequences (from GenBank) and by 37 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor. The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 1 mRNA sequence (from GenBank): AK096906 from tissue of skeletal muscle and by 1 shorter EST sequence: BP218086 from tissue of caudate nucleus. In this case the extended loop is placed in the middle of a TM-region which may seem confusing but it is done so to illustrate that here is where the extra amino acids end up (with this modelling technique). The number beneath the extended loop represent the number of amino acids (and 1 stop codon) making up the loop. The line beneath the exon-protein structure displays the coverage of the EST sequences. It should be noticed that the extra exon on the left most likely is interrupted by premature sequencing termination.

= exon wild type (a,b,c etc.)

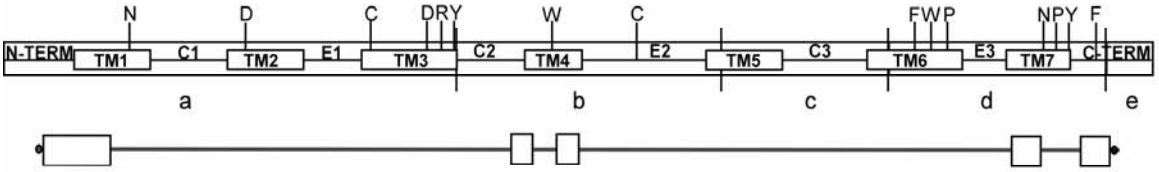
= exon splice variant

= extended loop with stop codon

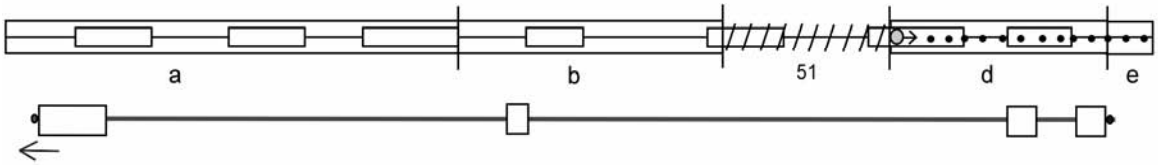
= end of sequence due to sequencing

→ = the exon of the variant goes further than the coding region of the wild type

Exon Distribution on Protein Structure and on Genome for TACR3 and Splice Variants



The recognized wild type. It is presented by 2 full length mRNA sequences (from GenBank) and by 4 EST sequences of different lengths (not full length). The letters above the exon-protein structure are the conserved motifs appearing in this particular receptor.
The letters inside the exon-protein structure designate the protein domains of the receptor. The letters beneath the exon-protein structure designate the different exons.



A variant that is supported by 1 EST sequence: CD013878 from unknown tissue. The number beneath the exon-protein structure represents the number of amino acids that are missing. The frame shift result in diverging amino acids and stop codons.

A variant supported by 3 EST sequences of different lengths: BG217327, BG197413 and BG198983 all from unknown tissue. The latter only covers the first two exons on the right. As noticed only the exon-genome figure is displayed for this variant. The reason for this is that as the figure shows the exons of the variant don't align with any of the exons of the wild type. Because of this the method used can't give the reading frame of interest and consequently not the resulting exon-protein structure. However, translation of the gene sequence of the variant in all three possible reading frames result in diverging amino acids and scattered stop codons. Therefore, one can conclude that this variant doesn't give rise to a functional receptor.

= exon wild type (a,b,c etc.)

= exon splice variant

/ = missing amino acids

= frame shift

= the exon of the variant goes further than the coding region of the wild type

Appendix 2

mRNA Information on Splice Variants

GPCR	mRNA gi.nr	mRNA acc.nr	Alteration	Domain	Tissue	Code**	Ref. ***
AVPR2	2654030	AF032388*	extended loop with stop codon and frame shift	H-VI resp. E-III, H-VII, C-term	small-cell carcinoma of the lung	a	28
AVPR2	4323606	AF101728*	extended loop with stop codon and frame shift	H-VI resp. E-III, H-VII, C-term	breast cancer	a	22, 23
AVPR2	18266918	BC015746*	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	Pancreas, epithelioid carcinoma	b	29
AVPR2	23138696	BC033090*	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	Pancreas, epithelioid carcinoma	b	29
AVPR2	27469609	BC041642*	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	Pancreas, epithelioid carcinoma	b	
AVPR2	51873895	BC080603*	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	Pancreas, epithelioid carcinoma	b	
AVPR2	14178023	BG830436	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	pancreas epithelioid carcinoma cell line	b	
AVPR2	14621077	BI161076	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	pancreas epithelioid carcinoma cell line	b	
AVPR2	14620710	BI160709	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	pancreas epithelioid carcinoma cell line	b	
AVPR2	14621427	BI161438	missing TM, missing loop	H-I, H-II, H-III, H-IV, N-term, C-I, E-I, C-II	pancreas epithelioid carcinoma cell line	b	
CCKBR	913752	S76072*	extended loop and frame shift	N-term resp. N-term, H-I, C-I	stomach	c	19
CCKBR	3287189	Y13464*	extended loop and frame shift	N-term resp. N-term, H-I	not reported	c	21
CCKBR	7677459	AF239668*	*functional (extended loop)*	C-III	colorectal cancer and premalignant polyps	d	1,6,27,28
CCKBR	53451714	AF441129*	*functional (extended loop)*	C-III	pancreatic carcinoma cells	d	1,6,27,28
CCKBR	15981132	AY029770*	*functional (extended loop)*	C-III	not reported	d	1,6,27,28
CCKBR	37777414	CD013884	missing TM,missing loop	H-I, H-II, N-term, C-I, E-I	not reported	e	16
EDNRA	2746074	AF014826*	missing TM, missing loop and frame shift	H-II, H-III, E-II resp. C-II, H-IV, E-II	vascular smooth muscle cells from human placenta	f	2
EDNRA	15251289	BI460633	missing TM, missing loop and frame shift	H-I, H-II, H-IV, N-term, C-I, C-II, E-II resp. H-V, C-III, H-VI, E-III	testis	g	
EDNRA	47329683	CN313269	missing TM, missing loop	H-I, H-II, N-term, C-I	embryonic stem cells, cell lines H1, H7 and H9 cells	h	
EDNRA	1478477	S81542*	missing TM, missing loop and frame shift	H-IV, C-II, E-II resp. H-V, C-III, H-VI, E-III, H-VII, C-term	not reported	i	20

EDNRA	1478479	S81545*	missing TM, missing loop	H-II, H-III, H-IV, E-I, C-II, E-II	lung	j	32
EDNRA	31873649	BX537573*	unknown	-	human cervix	k	
EDNRA	34528647	AK123169*	unknown	-	cerebellum	k	17
EDNRA	2734205	G36538*	unknown	-	not reported	k	
EDNRB	2285955	X99250	*functional (missing loop)*	C-term	placenta	l	8
EDNRB	22658394	BC031243*	unknown	-	Testis	m	
EDNRB	13965008	BG698082	unknown	-	skin	m	
ETBRLP1	34368178	BX649006*	extended loop, stop codon and frame shift	C-term, C-term resp N-term, H-I, C-I, H-II, E-I, H-III	retina	n	
ETBRLP1	15434918	BI547606	diverging a.a. sequence	C-I, H-II, E-I, H-III	hippocampus (brain)	n	
ETBRLP2	34526651	AK129946*	no exons	-	Kidney	o	
ETBRLP2	40267005	CD618740	no exons	-	not reported	o	
Ghrelin	22903848	BU553576	extended loop	H-VI	ovary teratocarcinoma cell line	p	
Ghrelin	11598280	BF513101	extended loop	H-VI	not reported	p	
GHSR	46575718	BC069068*	extended loop	H-VI	synthetic constructs	q	14
GHSR	47481092	BC069374*	extended loop	H-VI	PCR rescued clones	q	14
GHSR	1504142	U60181*	extended loop	H-VI	pituitary	q	13
GNRHR	1628389	Z81148*	missing TM, missing loop and frame shift	H-IV, H-V, E-II resp H-V, C-III, H-VI, E-III, HVII, C-term	pituitary (brain)	r	10
GNRHR	24658875	BC039430*	unknown	-	Brain, hypothalamus	s	
GNRHR	28278154	BC045560*	unknown	-	Brain, hypothalamus	s	
GNRHR	39918791	AJ617629*	no exons	-	not reported	t	
HCRTR1	23242909	BC035686*	missing TM, missing loop and frame shift	H-I, H-II, H-III, C-I, E-I resp. H-III, C-II, H-IV, E-II, H-V, C-III, H-VI, E-III. H-VII, C-term	pooled: lung, spleen, fetal	u	
HCRTR1	50486128	CR605321*	missing TM, missing loop	H-I, H-II, H-III, N-term, C-I, E-I	brain, fetal	v	
HCRTR1	47008668	BX433092	missing TM, missing loop	H-I, H-II, H-III, N-term, C-I, E-I	brain, fetal	v	
HCRTR1	30779168	BX433093	missing TM, missing loop	H-I, H-II, H-III, N-term, C-I, E-I	brain, fetal	v	
HCRTR1	45711690	AL535838	missing TM, missing loop	H-I, H-II, H-III, N-term, C-I, E-I	brain, fetal	v	
NMU1R	47373854	CN386259	extended loop with stop codon	H-VI	embryonic stem cells derived from H1, H7 and H9 cells	x	3
NMU1R	32001601	BX491295	extended loop with stop codon	H-VI	not reported	y	
NMU1R	10203987	BE782789	no exons	-	retinoblastoma	z	
NMU2R	766966	R13890	missing TM, missing loop	H-I, H-II, H-III, H-IV, H-V, N-term, C-I, E-I, C-II, E-II	whole brain, infant	aa	
NMU2R	876179	H11359	missing TM, missing loop	H-I, H-II, H-III, H-IV, H-V, N-term, C-I, E-I,	whole brain, infant	aa	

				C-II, E-II			
NMU2R	21167165	BQ428089	no exons	-	melanotic melanoma skin	bb	
NMU2R	12107080	BF759180	no exons	-	brain	bb	5
NMU2R	19895986	BQ066940	no exons	-	not reported	bb	
NPFF2	24370914	AJ311393*	missing TM, missing loop and frame shift	H-I, H-II, N-term, C-I resp. H-III, C-II, H-IV, E-II, H-V, C-III, H-VI, E-III, H-VII, C-term	placenta, hypothalamus	cc	18
NPFF2	47301665	CN285251	missing TM, missing loop and frame shift	H-I, H-II, N-term, C-I resp. H-III, E-I	embryonic stem cells, cell lines H1, H7, and H9	cc	3
NPFF2	14279164	AF236083*	extended loop with stop codon	N-term	not reported	dd	
NPFF2	30287948	CB993428	extended loop with stop codon	H-III	pre-eclamptic placenta	ee	
NPFF2	30292200	CB997680	extended loop with stop codon	H-III	pre-eclamptic placenta	ee	
NPFF2	28847522	BX283068	extended loop and missing TM, missing loop	N-term resp H-I, H-II, C-I	fetus material (8-9 weeks)	ff	
NPFF2	28615449	BX280001	extended loop and missing TM, missing loop	N-term resp H-I, H-II, C-I	fetus material (8-9 weeks)	ff	
NPFF2	2163669	AA449919	extended loop and missing TM, missing loop	N-term resp H-I, H-II, C-I	fetus material (8-9 weeks)	ff	
NPY1R	189284	L07615*	extended loop and frame shift	C-III resp. C-III, H-VI, E-III, H-VII, C-term	peripheral blood	gg	12
NPY1R	13717049	BG195362	extended loop	C-III	not reported	gg	
NPY5R	13713419	BG191732	missing TM, missing loop	H-I, H-II, H-III, H-IV, H-V, C-I, E-I, C-II, E-II, C-III	not reported	hh	11
OXTR1	2166140	AA452471	no exons	-	fetus material	ii	
OXTR1	27825922	BX089158	no exons	-	fetus material	ii	
PPYR1	21328764	BC021910*	no exons	-	Kidney, renal cell adenocarcinoma	jj	29
PPYR1	13133992	BG327554	no exons	-	Kidney, renal cell adenocarcinoma	jj	
TACR2	52090989	BP218086	extended loop with stop codon	H-V	caudate nucleus	kk	30
TACR2	21756503	AK096906*	extended loop with stop codon	H-V	skeletal muscle	kk	31
TACR3	37777408	CD013878	missing TM, missing loop	H-V, H-VI, C-III	not reported	ll	15
TACR3	13743348	BG217327	unknown	unknown	not reported	mm	11
TACR3	13719100	BG197413	unknown	unknown	not reported	mm	11
TACR3	13720670	BG198983	unknown	unknown	not reported	mm	11

* = mRNA from GenBank

** = the letter code shows which mRNA sequences that represent the same splice variant

*** = referring to corresponding number in reference list

Appendix 3

Exon Positions on Genome (1 = exon, 0 = no exon, - = end of sequence due to sequencing)

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1584-1608	1969-2855	1969-3219	2456-2855	2959-3166
AVPR2				1	1	0	0	1
	2654030	AF032388*	non func	1	0	1	0	0
	4323606	AF101728*	non func	1	0	1	0	0
	18266918	BC015746*	non func	1	0	0	1	1
	23138696	BC033090*	non func	1	0	0	1	1
	27469609	BC041642*	non func	1	0	0	1	1
	51873895	BC080603*	non func	1	0	0	1	1
	14178023	BG830436	non func	1	0	0	1	-
	14621077	BI161076	non func	1	0	0	1	-
	14620710	BI160709	non func	1	0	0	1	-
	14621427	BI161438	non func	1	0	0	1	-
Positions for chrX : 152689864-152694612								

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1937-2087	6480-6641	11675-11929	11675-12263	12095-12347	12653-12812	12653-13551	13016-13551
CCKBR				1	0	1	0	1	1	0	1
	7677459	AF239668*	functional	1	0	1	0	1	0	1	0
	53451714	AF441129*	functional	1	0	1	0	1	0	1	0
	15981132	AY029770*	functional	1	0	1	0	1	0	1	0
	3287189	Y13464*	non func	1	1	1	0	-	-	-	-
	913752	S76072*	non func	-	1	1	0	-	-	-	-
	37777414	CD013884	non func	1	0	0	0	1	1	0	1
Positions for chr 11: 6235799-6251285											

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	56935-57354	58082-63907	91098-91234	103756-103957	107128-107282	111068-111203	111627-111738	113727-113868
EDNRA				1	0	1	1	1	1	1	1
	1478479	S81545*	non func	1	0	0	0	1	1	1	1
	2746074	AF014826*	non func	1	0	0	1	-	-	-	-
	15251289	BI460633	non func	0	0	1	0	1	1	-	-
	47329683	CN313269	non func	0	0	1	1	-	-	-	-
	1478477	S81542*	non func	1	0	1	0	1	1	1	1
	31873649	BX537573*	non func	-	1	-	-	-	-	-	-
	34528647	AK123169*	non func	-	1	-	-	-	-	-	-
	2734205	G36538*	non func	-	1	-	-	-	-	-	-
Positions for chr4: 148707505-148878306											

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	20375-20511	22034-22144	22696-22832	23233-23384	25331-25538	25670-25784	29867-30601	40266-40784
EDNRB (rev.strand)				1	1	1	1	1	1	0	1
	2285955	X99250	functional	(18583-18735)	1	1	1	1	1	0	1
	22658394	BC031243*	non func	0	0	0	0	0	0	1	-
	13965008	BG698082	non func	0	0	0	0	0	0	1	-
Positions for chr13: 77349962-77411083											

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1-818	1-7645	173-973	17425-18449
ETBRLP1 (rev.strand)				1	0	0	1
	15434918	BI547606	non func	-	-	1	-
	34368178	BX649006*	non func	0	1	0	-
Positions for chr 7: 123980533-123998981							

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1-901	835-1424	1176-1983	1703-2040
GHRELIN				1	0	0	1
	22903848	BU553576	non func	-	0	1	-
	11598280	BF513101	non func	-	1	0	-
Positions for chr13: 48692475-48694514							

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1-306	2458-3253	2383-3253
GHSR (rev.strand)				1	1	0
	46575718	BC069068*	non func	-	-	1
	47481092	BC069374*	non func	-	-	1
	1504142	U60181*	non func	-	-	1
Positions for chr3: 173645653-173648905						

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	13873-13628	15673-16742	17939-17716	17810-17716	27483-26962
GNRHR (rev.strand)				1	0	1	0	1
	1628389	Z81148*	non func	1	0	0	1	1
	24658875	BC039430*	non func	-	1	-	-	-
	28278154	BC045560*	non func	-	1	-	-	-
	39918791	AJ617629*	non func	no exons				
Positions for chr4: 68421337-68462163								

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	3895-4093	4233-4413	5542-5788	6175-6294	8222-8452	9697-9821	11491-11682
HCRTR1				1	1	1	1	1	1	1
	23242909	BC035686*	non func	1	0	1	1	1	1	1
	50486128	CR605321*	non func	0	0	1	1	1	1	1
	45711690	AL535838	non func	0	0	1	1	1	-	-
	47008668	BX433092	non func	0	0	1	1	1	1	-
	30779168	BX433093	non func	0	0	1	1	1	1	-
Positions for chr1: 31749993-31765568										

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1-381	3079-3906	38-517	2814-3313
NMU1R (rev.strand)				1	1	0	0
	47373854	CN386259	non func	-	-	1	-
	32001601	BX491295	non func	-	-	-	1
	10203987	BE782789	non func	no exons			
Positions for chr2: 232215262-232219167							

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1-314	3267-3394	5870-5954	12198-12914
NMU2R (rev.strand)				1	1	1	1
	766966	R13890	non func	-	1	1	0
	876179	H11359	non func	-	1	1	0
Positions for chr5: 151751945-151764858							

GPCR	mRNA gi.nr	mRNA acc.nr	Receptor	115909 - 116209	123396- 123421	152942- 153032	212590- 212927	222044- 222148	230520- 230628	230983- 231816
NPFF2				1	0	0	1	1	0	1
	47301665	CN285251	non-func	1	0	0	0	1	-	-
	24370914	AJ311393*	non-func	1	0	0	0	1	0	1
	14279164	AF236083*	non-func	-	1	1	1	1	0	1
	30287948	CB993428	non-func	-	-	-	-	-	1	1
	30292200	CB997680	non-func	-	-	-	-	-	1	1
	28847522	BX283068	non-func	-	1	0	0	1	0	1
	28615449	BX280001	non-func	-	1	0	0	1	0	1
	2163669	AA449919	non-func	-	1	0	0	1	0	1
Positions for chr4: 73146746- 73494469										

GPCR	mRNA gi.nr	mRNA acc.nr	Receptor	1-2744	1253- 1710	1708- 2002	1806- 2504
NPY1R (rev.strand)				0	1	0	1
	189284	L07615*	non-func	1	0	0	0
	13717049	BG195362	non-func	-	-	1	-
Positions for chr 4: 164602808- 164606563							

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	1336- 2670	1325- 1485	2128- 2753
NPY5R				1	0	0
	13713419	BG191732	non func	0	1	1
Positions for chr4: 164627696- 164631700						

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	11499- 11760	13755- 13955	15595- 15748	15595- 15967	21615- 21816	22604- 22996
TACR2 (rev.strand)				1	1	1	0	1	1
	52090989	BP218086	non func	-	1	0	1	-	-
	21756503	AK096906*	non func	1	1	0	1	-	-
Positions for chr10: 70823090- 70857583									

GCPR	mRNA gi.nr	mRNA acc.nr	Receptor	129995 - 130309	131518- 131756	133599- 133745	131800- 131998	164593- 164849	196505- 196657	198528- 198718	259439- 259988
TACR3 (rev.strand)				1	0	0	1	0	1	1	1
	37777408	CD013878	non func	1	0	0	1	0	0	1	1
	13743348	BG217327	non func	0	1	1	0	1	-	-	-
	13719100	BG197413	non func	0	1	1	0	1	-	-	-
	13720670	BG198983	non func	0	1	1	-	-	-	-	-
Positions for chr4: 104738449- 105128430											

* = mRNA from GenBank