3.3. Analyse eines Amphioxus Cosmids

3.3.1. Cosmidbank

Die Amphioxus Cosmid-Bank (RZPD-Library-Nummer 117) besteht aus 36.000 Klonen (Burgtorf et al., 1998) mit einer mittleren Insertgröße von 38 kb (Abb. 39). Das entspricht bei einer Genomgröße des Amphioxus von 580 Mb etwa 2,5 Genomäquivalenten und damit nach der unten aufgeführten Formel einer Wahrscheinlichkeit von ca. 92% eine bestimmte Sequenz zu finden. Bei 22 Hybridisierungen, die in verschiedenen Laboratorien weltweit auf diese Bank durchgeführt wurden, wurden durchschnittlich 4,7 Klone pro 'Screen' als potentiell positiv identifiziert. In einem 'Rescreen' stellen sich durchschnittlich 2-3 Klone als 'echte Positive' heraus, womit die Bank sehr genau den theoretisch berechneten Parametern entspricht.

$$p = 1 - (1 - \text{Insertgröße} / \text{Genomgröße})^{n} \qquad n = \textit{Anzahl der Klone}$$

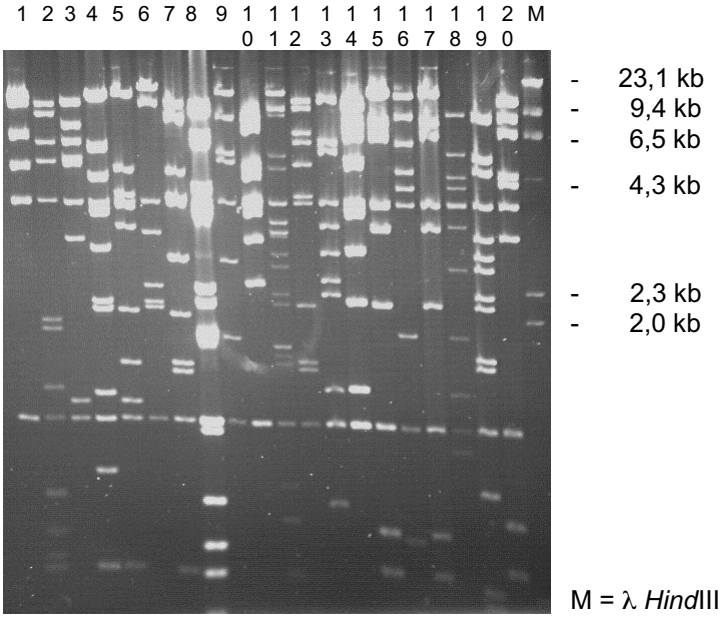


Abb. 39: Zwanzig zufällig ausgewählte, mit EcoRI verdaute Cosmide.

3.3.2. Sequenzanlyse

Die Amphioxus Cosmid-Bank (RZPD Nummer 117) wurde im Rahmen des Xp22-Projektes mit einem PCR-Produkt der Zebrafisch Spermin-Syntase cDNA (GenBank Zugriffsnummer: AJ009864) unter Inter-Spezies-Bedingungen 'gescreent'. Dabei wurden 4 Klone mit überlappenden Restriktionsfragmenten identifiziert. Von diesen wurde der Klon MPMGc117B0533 sequenziert. Dazu wurden 900 Klone aus einer 'Shot gun'-Bank amplifiziert und auf einer ABI 377 Sequenziermaschine mittels 'dye primer'-Technologie (Applied Biosystems)

analysiert. Die Sequenzen wurden mit Hilfe des Staden Pakets (Bonfield, 1995) zu einem 40,8 kb großen *Conitg* assembliert und die gesamte Sequenz unter der Zugriffsnummer Y18367 in GeneBank deponiert. Die Sequenz weist einen G+C–Gehalt von 41,35 % auf. Mit Hilfe der Exonvorhersage-Programme GenScan (Burge und Karlin, 1997, 1998), Grail (Uberbacher, 1991) und Mzef (Zhang, 1997) wurde die Sequenz auf potentielle Exons analysiert. Die putativen Exons wurden translatiert und mittels BLAST (Altschul, 1990) mit den Datenbanken GeneBank, Swissprot, TREMBL und der innerhalb des Amphioxus EST-Projektes erstellten Datenbank (Panopoulou, 1999) verglichen. Außerdem wurde die gesamte Sequenz mittels BlastX, das die Sequenz in allen 6 Leserastern translatiert, mit GeneBank verglichen. Die Ergebnisse wurden in ACEDB-Format (Durbin and Thierry-Mieg, 1991) gespeichert und graphisch dargestellt (Abb. 40). Von den Unterschieden in den Vorhersagen ist es offensichtlich, daß die drei Programme unterschiedliche Kriterien verwenden, um mögliche Exons zu identifizieren. Sowohl auf dem 'forward' als auch auf dem 'reverse' Strang wurden mit mzef die meisten potentiellen Exons identifiziert (92). Grail und Genescan schlugen mit 27 beziehungsweise 24 etwa gleich viele Exons vor.

3.3.3. Gen-Inhalt des Amphioxuscosmids MPMGc117B0533

Innerhalb des 40,8 kb *Conitgs* zeigen einige der identifizierten Exons Homologie zu Genen, die in öffentlichen Datenbanken zugänglich sind. Sieben potentielle Exons des 'forward' Strangs im Bereich von 21 – 24 kb zeigen Homologie zu der Super-Familie der Aldo-Keto-Reduktasen. Ein Exon des reversen Strangs zwischen 2-3 kb zeigt Ähnlichkeit zu α(1a, 2a, 1b, 2b) adrenergen und Serotonin-Rezeptoren und ein weiteres bei etwa 25 kb auf dem reversen Strang ist ähnlich einem 'Low Density Lipoprotein Receptor-related' Protein2-Prekursor. Die weiteren vorhergesagten Exons, die keine Ähnlichkeit zu anderen, bereits in GeneBank deponierten Genen zeigen, könnten entweder Teile von stärker divergierten oder Amphioxus-spezifischen Genen sein, zu Genen gehören, die in anderen Organismen bisher noch nicht identifiziert sind, oder es handelt sich dabei um falsch vorhergesagte Exons. Besonders bemerkenswert sind dabei die Regionen von 11-18 kb und von 28-35 kb des reversen Strangs, die allen verwendeten Exonvorhersageprogrammen nach exonreich zu sein scheinen, aber keine Homologie zu einem Datenbankeintrag zeigen. Man findet innerhalb des Cosmids aber auch umgekehrt den Fall, daß keines der Exonvorhersage-Programme ein Exon vorschlägt, aber ein im Rahmen des Amphioxus EST-Projektes (Panopoulou, 1999) identifizierter Klon einer genomischen Region zugeordnet werden kann. Dies ist am auffälligsten im Bereich von 10,5-13 kb des reversen Strangs; aber auch direkt am Anfang bis ca. 1,2 kb und bei etwa 38,5 kb. Drei von diesen ESTs wurden in GenBank unter den Zugriffsnummern: AI391435, AI391436 und AI391434) deponiert; allerdings entsprechen diese in der Regel keinem der vorhergesagten Exons, zeigen keine Homologie zu irgendwelchen bereits bekannten Protein und weisen teilweise auch keine längeren offenen Leseraster auf. Auffällig ist außerdem, daß teilweise die 'ESTs' mehr Nukleotide enthalten, als die genomische Sequenz, daß die 'Exons' eines ESTs sehr dicht beieinander liegen und in einem Falle (026K17BFLG RRset39) Teile des EST in beiden Orientierungen mit der Cosmidsequenz paaren (Tabelle 5).

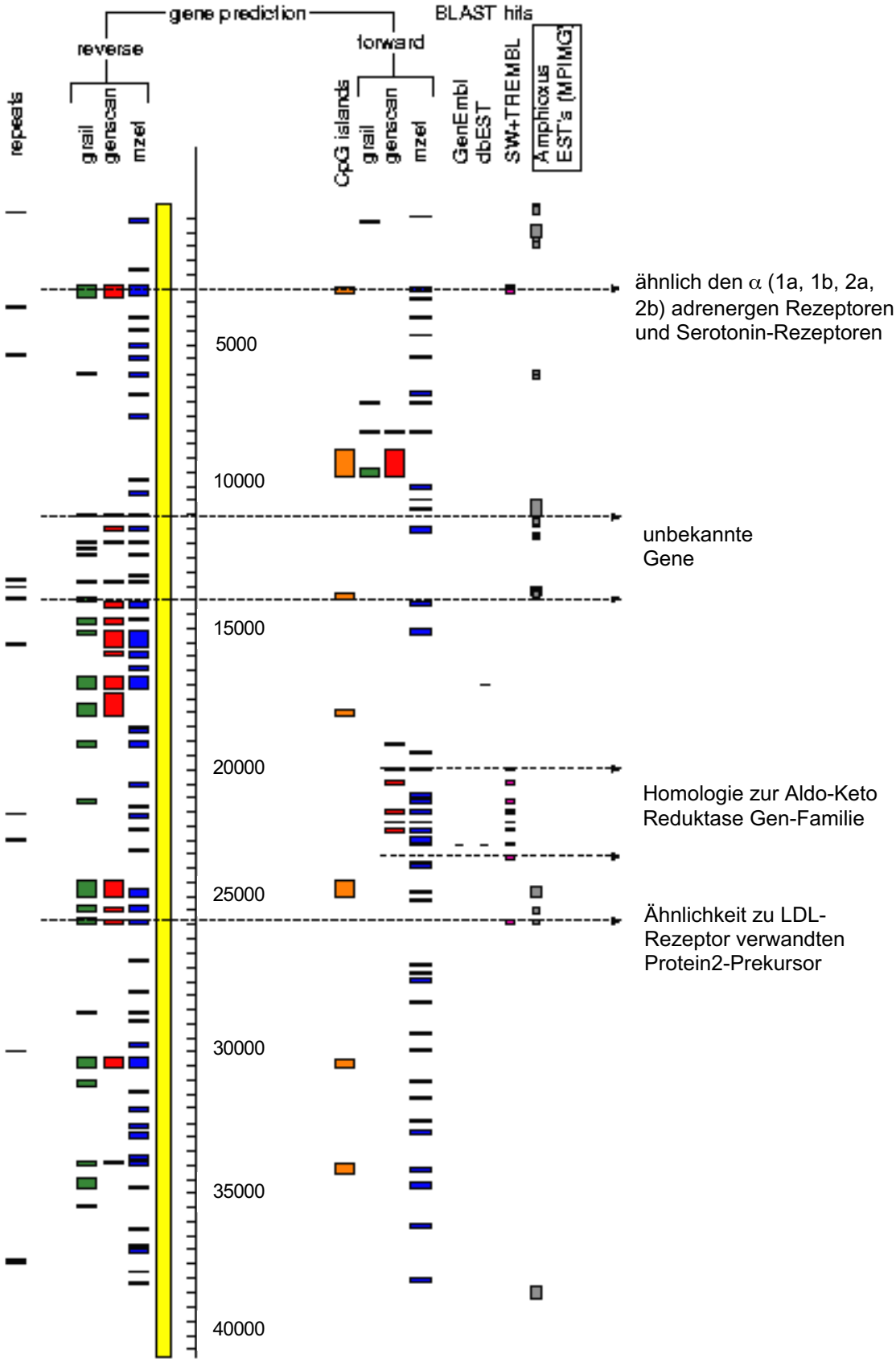


Abb. 40: ACEDB-Darstellung der Amphioxus-Cosmid-Analyse (MPMGc117B0533)

EST-Klonename (nt)	**Nukleotide im Cosmid**	**Orientierung / E-Werte**	**Charakterisierung**
MPIMGBFLG498_26C3 (316-615)	5-307	+ $3,6 \cdot e^{-24}$	kein match / kein längeres ORF
MPIMG531_120G23 BFL26 RRset4 (18-245)	119-360	+ $2,7 \cdot e^{-31}$	kein match / kein längeres ORF
026K17BFLG RRset39 (35-473) (35-208) (285-486)	718-1157 1533-1360 1395-1198	+ $4,7 \cdot e^{-40}$ - -	kein match /kein längeres ORF Orientierung + und -
MPIMGBFLG_32D13 BFLG_set4 (347-486) (238-354)	6047-5913 6178-6067	- $2,0 \cdot e^{-14}$ -	kein match / kein längeres ORF; EST z.T. mehr Nukleotide als die genomische Sequenz
MPIMGBFLG498_111H01 (GenBank AI391435) (64-683) (31-71)	11063-10443 11462-11422	- $1,0 \cdot e^{-133}$ -	kein match / kein längeres ORF
037B17 BFLG RRset40 (GenBank AI391436) (55-312)	11348-11097	- $2,2 \cdot e^{-18}$	kein match / kein längeres ORF
089M11 BFLG RRset29 (GenBank AI391434) (47-185) (237-384)	11612-11750 11754-11862	- $9,4 \cdot e^{-14}$ -	kein match / kein längeres ORF
023B11 BFGL RRset 39 (345-397) (151-342) (22-63)	13679-13626 13896-13702 14025-13983	- $1,2 \cdot e^{-29}$ - -	kein match / kein längeres ORF; EST z.T. mehr Nukleotide als die genomische Sequenz
MPIMGBFLG498_001K24 (28-349)	13855-13537	- $4,0 \cdot e^{-48}$	kein match / kein längeres ORF
MPIMGBFLG498_45P2 (GenBank AI391437) (236-610) (43-247) (29-155)	24500-24126 25112-24916 25482-25356	- $5,1 \cdot e^{-92}$ - -	längeres ORF match zu LDL-Rezeptor-Prekursor
MPIMGBFLG498_094D02 (18-454)	38705-38261	- $4,3 \cdot e^{-65}$	kein match / kein längeres ORF

Tabelle 5: Amphioxus ESTs mit Homologie zum Amphioxus-Cosmid MPMGc117B0533.

Aufgrund der hohen Homologie zur Aldo-Keto-Reduktase-Familie konnte einem der potentiellen Gene des Cosmids eine mögliche Funktion zugeordet werden. Aus der Superfamilie der Aldo-Keto-Reduktasen sind mehr als 45 Sequenzen von so verschiedenen Organismen wie *E.coli*, *S. cerevisiae*, *C. elegans* und einer ganzen

Anzahl von Vertebraten veröffentlicht (Seery et al., 1997). Sie alle katalysieren NADPH abhängige Carbonyl-Reduktionen, wie zum Beispiel die Prostaglandin-F-Synthase (EC 1.1.1.188) die Reaktion von Prostaglandin-H_2 zu Prostaglandin-$F_{2\alpha}$, und die Aldose-Reduktase (EC 1.1.1.21) die Umwandlung von Glukose in Sorbitol. Die genomische Struktur innerhalb der Familie variiert stark. Zum Beispiel besteht die humane Carbonyl-Reduktase aus lediglich drei Exons (Watanabe et al., 1998), wohingegen die humane 3α-Hydroxysteroiddehydrogenase 9 Exons umfaßt (Khanna et al., 1995). Dieses Gen, welches eine genomische Region von etwa 20 kb überspannt, hat eine ähnliche Struktur wie das hier identifizierte Amphioxus-Gen, welches ebenfalls aus 9 Exons besteht. Die mittleren 7 konnten durch eine Kombination von Mzef und Grail identifiziert werden; die beiden terminalen Exons konnten jedoch nur aufgrund von Homologie der genomischen Sequenz zu bereits bekannten Mitgliedern dieser Familie deduziert werden. Insgesamt kodiert das Gen für 273 Aminosäuren. Die Intron-Größen variieren zwischen 254 und 542 bp (Abb. 41) und die Sequenzen um die 'Splice'-Stellen entsprechen dem Konsensus (5'-GT und AG-3'). Ein mögliches Polyadenylierungssignal befindet sich 356 bp 'downstream' des Stop-Kodons.

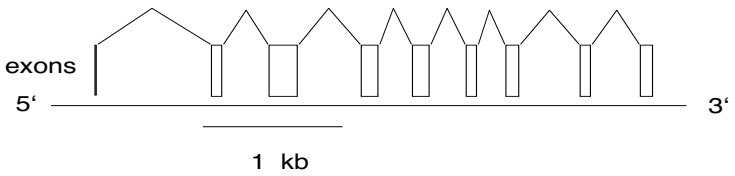


Abb. 41: Genomische Struktur der Amphioxus Aldo-Keto-Reduktase; Exon 1: Position 19107-19124; Exon 2: Position 19956-20037; Exon 3: Position 20386-20564; Exon 4: Position 21045-21157; Exon 5: Position 21412-21534; Exon 6: Position 21826-21878; Exon 7: Position 22093-22187; Exon: 8 Position 22640-22703; Exon 9: Position 23069-23163.

Die höchste Homologie der vorhergesagten Proteinsequenz findet man zu der 3α-Hydroxysteroid Dehydrogenase (EC 1.1.1.53) des Hausschweins (Tanaka et al., 1992) mit 52,6 % identischen Aminosäuren. Des weiteren zeigen die NADPH abhängige Carbonylreduktase von Mensch, Maus, Kaninchen und Ratte, sowie die induzierbare Ratten-Carbonylreduktase große Ähnlichkeit. Im 'Pretty-Box'-Vergleich (Abb. 42) erkennt man mehrere stärker konservierte Bereiche, zum Beispiel in den Bereichen von Aminosäure 218-227 und 238-257, hingegen zwischen Aminsosäuren 130-152 und 174-209 unterscheiden sich die Sequenzen am meisten. Da durch Sequenzvergleich mit der internen, nicht redundanten Amphioxus EST-Datenbank (G. Panopoulou, 1999), die bereits rund 11000 ESTs enthält, keine passende cDNA isoliert werden konnte, mußte die tatsächliche Expression dieses Gens durch RT-PCR bewiesen werden.

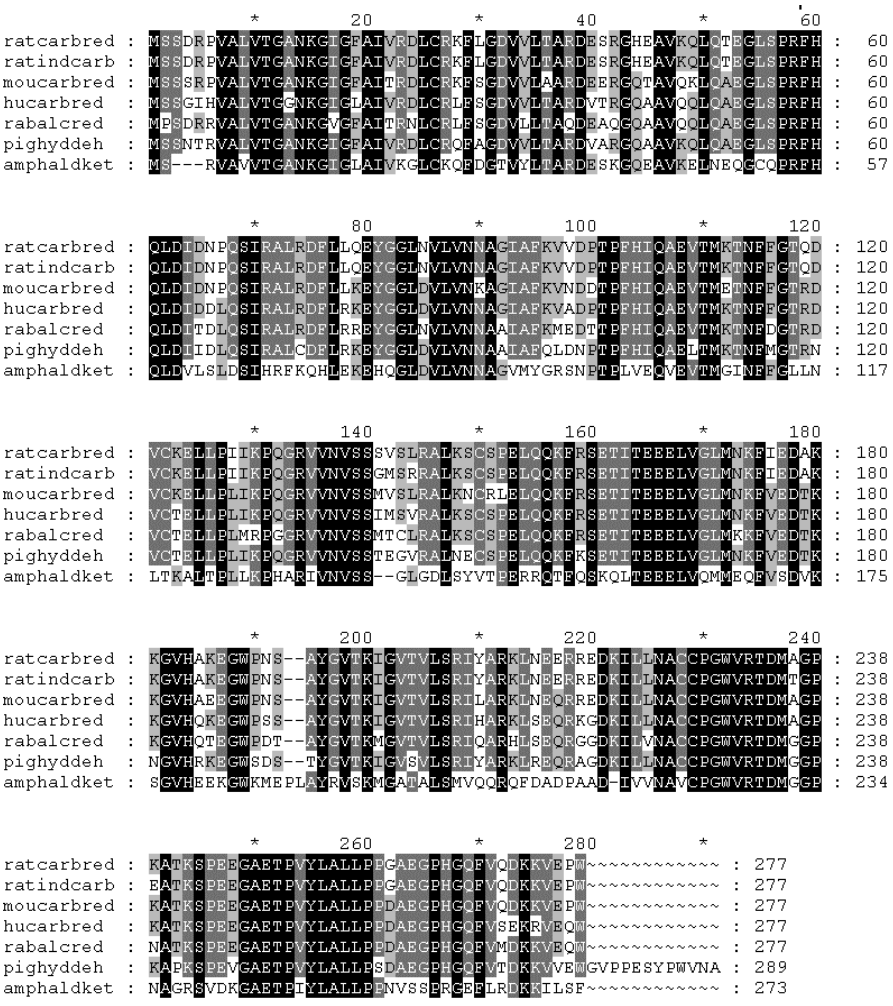


Abb. 42: Homologievergleich der Amphioxus Aldo-Keto-Reduktase mit Ratcarbred: NADPH abhängige Carbonylreduktase der Ratte (Wermuth et al., 1995); ratindcarb: induzierbare Carbonylreduktase der Ratte (Aoki et al., 1997); moucarbred: NADPH abhängige Carbonylreduktase der Maus (Wei et al., 1996); hucarbred: NADPH abhängige Carbonylreduktase des Menschen (Wermuth et al., 1988); rabalcred: Sekundärer Alkohol Oxidoreduktase des Kaninchens (Gonzalez et al., 1995); pighyddeh: 3α-Hydroxysteroiddehydrogenase des Schweins (Tanaka et al., 1992); amphaldket: Amphioxus Aldo-Keto Reduktase Gen.

Innerhalb des Cosmids befindet sich auf dem reversen Strang von Nucleotid 3276-2878 ein weiteres Exon, das eine signifikante Ähnlichkeit zu einem in einer Datenbank deponierten Gen hat. Es zeigt Homologie zu alpha (1a, 1b, 2a, 2b) adrenergen und Serotonin Rezeptoren. Der best E-Wert ($5{,}2e^{-05}$) wird für den *Rattus norvegicus* α–2b adrenergen Rezeptor (Swissprot Zugriffsnummer: P19328) bestimmt (Abb. 43A). Die Ratten cDNA kodiert für 443 Aminosäuren, von denen die 326 carboxyterminalen Reste nicht mehr innerhalb des analysierten Cosmids liegen. In der EST-Datenbank (G. Panopoulou, 1999) wurde keine korrespondierende cDNA gefunden, und damit konnte nicht abschließend geklärt werden, ob es sich dabei um ein tatsächlich exprimiertes Gen handelt.

(A)

```
>swiss|P19328|A2AB_RAT RATTUS NORVEGICUS ALPHA-2B ADRENERGIC RECEPTOR
(ALPHA-2B ADRENORECEPTOR).

     Score: 129, Expect = 5.2e-05, 35% identities, length = 90

Query:  3150 LLLAVGGNNPVLLTVLLTEDLRTPGNVLICALAATDILQAVTKIPLAMYFLVIREALVFD 2971
             +L  + GN  V+L VL +  LR P N+ + +LAA DIL A   IP ++   ++     +
Sbjct:    28 ILFTIFGNALVILAVLTSRSLRAPQNLFLVSLAAADILVATLIIPFSLANELLGYWYFWR 87

Query:  2970 AFQVAVSAIYMFFGTFSVCIIAPISLDRYW 2881
             A+     A+ + F T S+  +  ISLDRYW
Sbjct:    88 AWCEVYLALDVLFCTSSIVHLCAISLDRYW 117
```

(B)

```
>swiss|P98158|LRP2_RAT RATTUS NORVEGICUS LOW-DENSITY LIPOPROTEIN RECEPTOR-RELATED
PROTEIN 2 PRECURSOR (MEGALIN) (GLYCOPROTEIN 330).

     Score: 109, Expect = 4.3e-05, 57% identities, length = 33

Query: 25424 QFKCANGRCIHAELECDGINNCGDSSDESNLDH   25328
             QF+C NGRCI     CDG N+CGD SDE    H
Sbjct:  2912 QFQCDNGRCISGNWVCDGDNDCGDMSDEDQRHH    2944
```

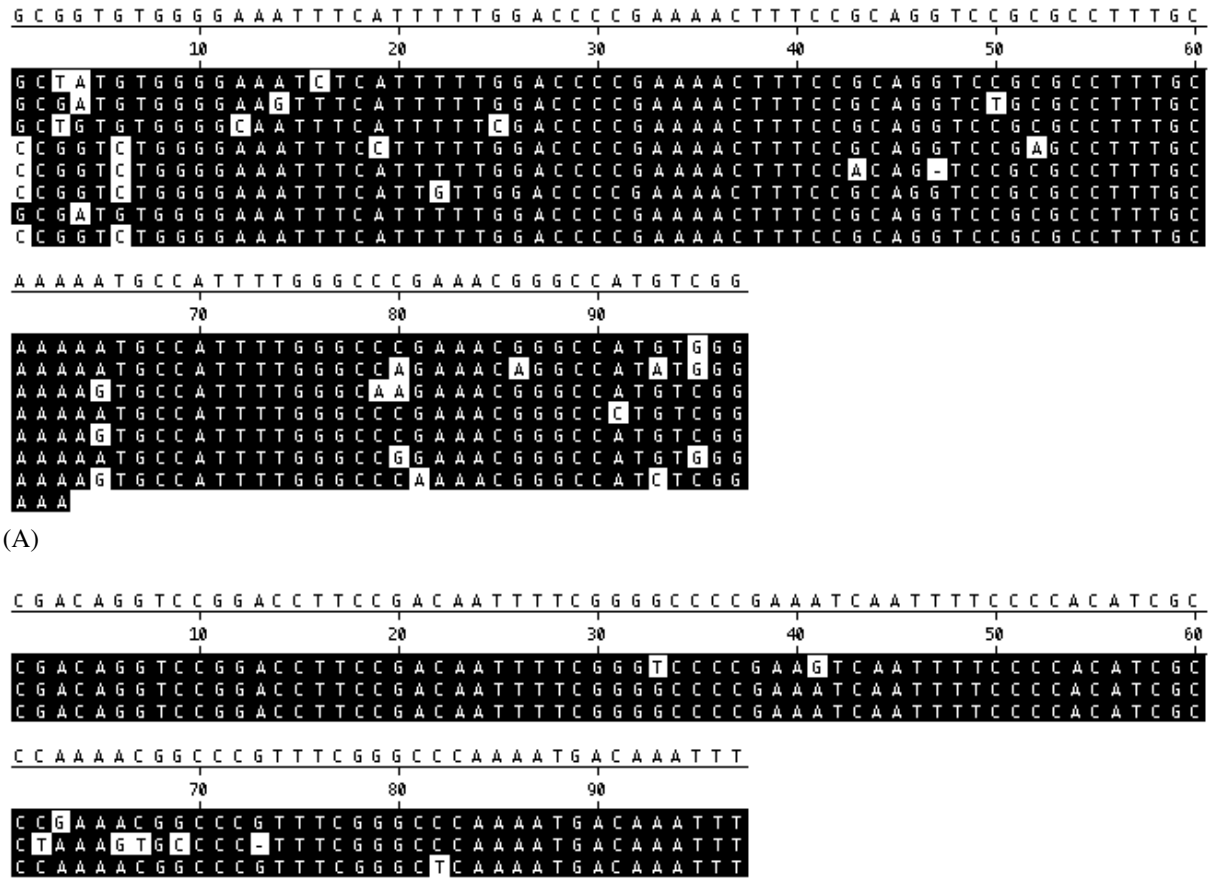
Abb. 43: Sequenzvergleich zwischen der deduzierten Amphioxus Sequenz und dem *Rattus norvegicus* α-2b adrenergen Rezeptor (A) und dem *Rattus norvegicus* 'low-density lipoprotein receptor-related protein 2 Precursor' (B).

Die dritte signifikante Homologie zu einem Datenbank-Eintrag wurde zu einem Teil des *Rattus norvegicus* 'low-density lipoprotein receptor-related protein 2 precursor' (Swissprot Zugriffsnummer: P98158) mit einem E-Wert von $4{,}3e^{-05}$ (Abb. 43B) sowie einer Reihe weiterer LDL-Rezeptoren anderer Spezies gefunden. Das entsprechende Exon wurde von den Exonvorhersageprogrammen von Nukleotid 25476-25360 des reversen Stranges vorhergesagt. Alle verwendeten Programme (Grail, GenScan und Mzef) sagen in dieser Region drei eng benachbarte Exons vorher, die die Nukleotide 25476-25360, 25011- 24916 und 24488-23893 umfassen. Diese drei Exons sind alle in dem gleichen Amphioxus-EST (GenBank Zugriffsnummer.: AI391437), welches innerhalb des Amphioxus-EST-Projektes (Panopoulou et al, 1999) identifiziert wurde, enthalten. Dies läßt den Schluß zu, daß sie Teile eines Gens sind, welches auch tatsächlich exprimiert wird.

3.3.4. Repetitive Elemente innerhalb des Amphioxuscosmids MPMGc117B0533

Das Amphioxus-Cosmid MPMGc117B0533 beinhaltet außer zwei Genen und mindestens einem weiteren potentiellen Exon auch verschiedene repetitive Elemente. Am auffälligsten sind dabei zwei sich tandemartig wiederholende Elemente von jeweils 97 bp Länge. Das erste wird 7,6 mal wiederholt (8641 - 9381), wobei sich jeweils nur einige Nukleotide zwischen den Einheiten unterscheiden (Abb. 44A); das zweite, ebenfalls eine 97 bp Einheit, ist dreimal wiederholt, und schließt sich direkt an (Nukleotid 9382 - 9671; Abb. 44B). Innerhalb der Wiederholungseinheiten liegt der GC-Gehalt bei ca. 54% und in beiden 'Repeats' kommt die

Basenfolge CpG etwas häufiger vor (8 x bzw. 9 x) als man aufgrund ihrer Basenzusammensetzung erwarten würde. Von Grail wurden sie deshalb auch als CpG-Island identifiziert und sind in Abb. 40 entsprechend dargestellt; in einem Fenster von 55 bp zeigen die beiden Elemente 47,5% Übereinstimmung in ihrer Basensequenz. Andere repetitive Elemente in dem Cosmid sind einfachere, kurze, direkte Wiederholungen wie $(A)_n$, $(GAAA)_n$, $(TA)_n$, $(CA)_n$, und $(AAAT)_n$, sowie eine kurze invertierte Wiederholung, deren Positionen in Abb. 40 angegeben sind.



(A)

(B)

Abb 44: Sequenzvergleich der Wiederholungseinheiten zweier tandemartig repetitiven Elemente des Cosmids MPMGc117B0533 (Nukleotid 8641-9671)