In silico-initiated cloning and molecular characterization of cortexin 3, a novel human gene specifically expressed in the kidney and brain, and well conserved in vertebrates

HAI TAO WANG[1], JI WU CHANG[2], ZHI GUO[1] and BAO GUO LI[1]

[1]Department of Interventional and Minimally Invasive Therapy, Tianjin Cancer Institute and Hospital, Tianjin Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University, Huanhuxi Road, HeXi District, 300060 TianJin; [2]Division of Uropathology and Tumor Immunology, Tianjin Institute of Urologic Surgery, The Second Hospital of Tianjin Medical University, Pingjiang Road 23, HeXi District, 300211 TianJin, P.R. China

Received May 28, 2007; Accepted July 13, 2007

Abstract. We report on the *in silico*-initiated cloning and molecular characterization of CTXN3 (cortexin 3), a new human gene that was specifically expressed in the kidney and brain due to tissue-specific alternative exon 1 usage, on chromosome 5q23.2 using digital gene expression displayer (DGED) and a novel *in silico* cloning approach based on both expressed sequence tags (ESTs) and genomic sequence. The gene CTXN3 included 3 exons and spanned an approximate 9.6-kb region of human chromosome 5q23. Two alternative transcript variants (GenBank accession nos. AB219764 and AB219832) were 1660 and 1458 bp long, respectively, encoding for an 81-amino acid protein with a predicted molecular weight of 8933.4 Da. The predicted human CTXN3 protein had 43% identity with function-unknown protein cortexin, which showed brain-specific expression. Further analysis of the encoded protein using PSORT II, TMpred, and PSIPRED programs demonstrated a putative single membrane-spanning domain in the middle of the CTXN3 amino acid sequence, indicating that it might be an integral membrane protein which may mediate extracellular or intracellular signaling of the kidney or brain. Analysis of the predicted CTXN3 orthologs from different species showed that these proteins are highly conserved in vertebrates. In conclusion, a combination of bioinformatics and molecular approaches is useful in the identification of genes expressed in specific tissues. Selective expression of CTXN3 in the kidney and brain, the amino acid identity to cortexin, and its high conservation among different species indicate that CTXN3 may be involved in a process specifically restricted to kidney and brain tissue function.

Correspondence to: Dr Hai Tao Wang, Department of Interventional and Minimally Invasive Therapy, Tianjin Cancer Institute and Hospital, Tianjin Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University, Huanhuxi Road, HeXi District, 300060 TianJin, P.R. China
E-mail: peterrock2000@126.com

Abbreviations: bp, base pairs; EST, expressed sequence tag; RT-PCR, reverse transcriptase PCR; RACE, rapid amplification of cDNA ends; CDS, coding sequence; NCBI, National Center for Biotechnology Information

Key words: digital gene expression displayer, *in silico* cloning, bioinformatics, tissue-specific gene, alternative splicing, transmembrane protein

Introduction

The human genome is composed of ~100,000 genes. Of these, only 15,000 are thought to be expressed in one particular cell (1). These 15,000 genes include housekeeping genes that are constitutively expressed across all cell types, as well as those that are selectively expressed in one specific tissue or in a small number of tissues. These tissue-specific genes often play an important physiological role for that tissue. Genes such as PSA, PSMA (2,3), leptin (4), MyoD, and MEF2 (5,6) are examples of such tissue-specific genes that contribute to the development and/or function of the prostate, fat and muscle, respectively. Therefore, identification and characterization of tissue-specific genes can play an important role in our understanding of molecular mechanisms of tissue physiology and pathophysiology.

In contrast to other tissues such as muscle, fat and brain, relatively little is known about the molecular characteristics of kidney tissue. In the literature only ~20 genes have been identified that are expressed in a highly renal-specific manner (7-9). In order to further characterize kidney tissue at the molecular level, we initiated a systematic search for genes that are expressed specifically in the kidney. Many different traditional methods are currently being used to identify tissue-specific genes, including representational difference analysis (RDA) (10), differential display (11), and subtractive cDNA library screening (12).

With the development of the human genome sequencing efforts, the unigene database of the National Center for Biotechnology Information provides a comprehensive

collection, not only of expressed sequence tags (ESTs), but also of various data-mining tools to analyze the ESTs. Digital gene expression displayer (DGED) (13,14) of the Cancer Genome Anatomy Project (CGAP) is a quantitative method by which the user is able to determine the fold differences between the ESTs from the different libraries being compared, using a statistical method to quantitate the transcript levels. In this report, we describe the *in silico* cloning and characterization of a new gene using DGED, which is specifically expressed in the kidney and brain. On the basis of significant sequence homology to function-unknown protein cortexin, the HUGO Gene Nomenclature Committee has designated this gene CTXN3 (cortexin 3). By combining *in silico* cloning and 5' RACE PCR, we cloned two CTXN3 full-length cDNAs from human kidney and brain tissue respectively. We then extended our analysis to identify CTXN3 orthologs in several vertebrates, including mouse, rat, cow, dog, chicken, and zebrafish. This study also evaluated experimentally the tissue expression of two CTXN3 transcript variants as well as the bioinformatics analysis of the deduced amino acid, genomic organization and promoter region. This is the first report on the molecular cloning, genomic characterization and expression analysis of the CTXN3 gene.

Materials and methods

EST database mining using DGED on the CGAP website. The CGAP database (http://cgap.nci.nih.gov/Tissues/GXS) was accessed, and the digital gene expression displayer (DGED) (13) tool was used according to the database instructions. DGED takes advantage of the unigene database by comparing gene expression between two pools of kidney and non-kidney libraries. Relative expression was calculated for each unigene cluster as the number of positive libraries out of the total in each tissue type, in addition to the total number of clones detected in each type.

Identification and characterization of putative full-length CTXN3 gene in silico. We applied a novel *in silico* cloning approach, which we referred to as SGC-ESTGS (*in silico* gene cloning based on both ESTs and genomic sequence), which used the combination of ESTs and genomic draft sequence to obtain exact gene prediction, including determination of exon-intron boundaries, potential alternative splicing variants, gene cloning by exon assembling and verified by ESTs and cDNAs.

Specifically, the first step consisted of assembling all the EST sequences in the unigene cluster (Hs.66194) by SeqMan program in the DNASTAR sequence analysis package (Madison, WI) to obtain a consensus sequence. The second step dealt with the consensus sequence as a query performing BLAT program on the University of CA-Santa Cruz (UCSC) browser (15). Thus all the ESTs and mRNA sequences overlapping with the contig were aligned to the human genomic sequence on the UCSC browser. Third, according to the resulting graph, exon/intron organization and potential alternatively splicing variants were determined. Potential full-length cDNA sequences were obtained by assembling exons using the reference genomic sequence, and putative 5' and 3' UTRs were assembled according to the longest EST extension. Each base needed to be further verified by all the EST sequences, and the wrong genomic sequence was corrected manually. The tissue expression profile was predicted by EST information.

Experimental verification of the cDNA sequence, tissue expression of the kidney and brain form of CTXN3 mRNA and molecular cloning full-length CTXN3 cDNA sequence by 5' RACE

Primers. Taking advantage of the difference in the 5' end of the sequence of both CTXN3 cDNAs, two pairs of PCR primers were designed that permitted us to amplify each form of CTXN3 mRNA. Another set of PCR primers designed on the common sequence permitted us to amplify either form of CTXN3 mRNA. The nucleotide sequences of the primers used in this study were: primer a (5' AAA AGA TGA ATC GAG ATG CAG TGT G 3'); primer b (5' CAA GAC ACA AAG CAC TTC ATC TCC TC 3'); primer c (5' TGA AGC AGA ATC CTC TGA AGA GTT G 3'); and primer d (5' GAC TCT ATT TCC TGA GCA CCC ACA 3'). All primers were synthesized by Augct (Beijing, P.R. China).

Expression analysis of CTXN3. Total RNA from frozen samples of human kidney and whole brain was extracted using Trizol® reagent (Invitrogen) according to the manufacturer's protocol. Aliquots of total RNA (10 μg) were electrophoretically separated in the presence of formaldehyde and transferred to Hybond N^+ membranes. A DNA fragment, 492 bp in size, was obtained by RT-PCR of human kidney RNA with primer pair c and d and was [a-^{32}P]dCTP radio labeled for filter hybridization at 58˚C. The filter was exposed at -80˚C for 3 days.

The tissue expression of the two *CTXN3* alternative transcripts was determined by RT-PCR using poly-A-purified RNA from a panel of human tissues using AMV Reverse Transcriptase (Promega) as per the manufacturer's instructions. The full-length transcript was amplified with primer d and primer b to give a 492-bp product, and the exons 1a- and 1b-containing transcripts were amplified with primers a, c, and exon 3 antisense primer d (above) to yield a 640- and 816-bp product, respectively. PCR conditions included denaturation at 94˚C for 4 min, followed by 20-35 cycles at 94˚C for 20 sec, 57.2˚C for 20 sec, and 72˚C for 2 min, with final elongation at 72˚C for 4 min using Taq polymerase (Promega). RT-PCR products were sequenced to confirm their identity.

Molecular cloning of the full-length cDNA sequence of the kidney and brain form of CTXN3 mRNA by 5' RACE. 5' RACE using the Marathon cDNA amplification kit (Clontech) was performed according to the manufacturer's instructions. To amplify the 5' region of *CTXN3* cDNA, we used a gene-specific reverse primer b and the adapter primer AP1 supplied in the kit. The cDNA template was synthesized from human kidney and brain mRNA (Clontech) respectively. The PCR products were cloned using a TA cloning kit (Invitrogen, Carlsbad, CA), and nucleotide sequences were determined with an ABI PRISM 3700 DNA sequencer (Applied Biosystems, Foster City, CA).

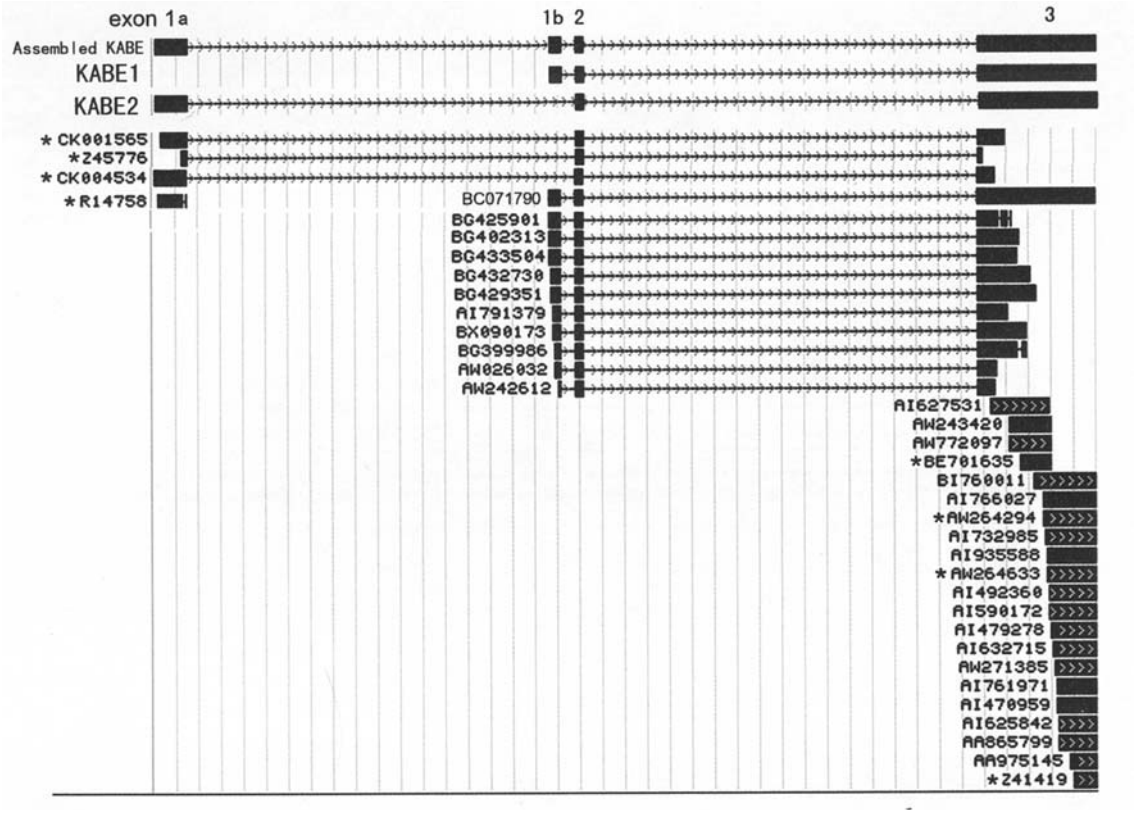

Figure 1. Schematics showing the Hs.66194 unigene cluster and the ESTs aligned to it. There are 36 ESTs, 8 from brain and 28 from kidney libraries. The rectangular black boxes represent exons, and the line connecting the boxes represents introns. The accession nos. for each EST in the cluster are indicated on the left side. The accession nos. of ESTs denoted by * originated from brain libraries of human tissue; others were from kidney libraries. Primers used for RT-PCT analysis are also showed. Primers a, b and c are specific for the exon 1a, 1b, and 2, respectively. Primer d is antisense to exon 3 and was used in all of the reactions.

Bioinformatics analysis of the predicted amino acid. Open reading frames (ORF) were identified by the ORF Finder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Derived protein sequences were analyzed using various bioinformatics tools. The signal peptide and transmembrane domain were searched for with the Kyte and Doolittle hydrophobicity analysis (http://occawlonline.pearsoned.com/bookbind/pubbooks/bc_mcampbell_genomics_1/medialib/activities/kd/kyte-doolittle.htm) (16), TMpred (http://www.ch.embnet.org/software/TMPRED_form.html) (17) and TMHMM server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) (18), and the nuclear localization signal with the PSORT II program (http://psort.nibb.ac.jp/). The domain structure of novel protein was first searched for with the RPS-BLAST program (http://www.ncbi.nlm.nih.gov). Protein secondary structure was predicted by PSIPRED Web service (19). The molecular weight was calculated with ProtParam tool (http://ca.expasy.org/tools/protparam.html) (20).

Identification of orthologs and analysis. To obtain possible orthologous sequences of CTXN3 from other species, various sequence databases were searched using the TBLASTN program at NCBI with the human protein sequence as a query. The CTXN3 ortholog sequences were assembled by the above described *in silico* cloning approach. The predicted cDNAs, genomic structures, and splicing events were verified using EST sequences. Multiple sequence alignments were generated using CLUSTALW program (http://www.ebi.ac.uk/clustalw/index.html) (21) with default parameters and alignment results in CLUSTALW ALN format were imported into ClustalX (22) for visualization.

Results

Computer analysis of the unigene cluster Hs.66194. The DGED was performed by digitally comparing all the kidney-derived libraries against all the other libraries in the database. The up-regulated or down-regulated genes were discovered, and statistically significant hits (Fisher's exact test) showing $P<0.05$ were compiled (data not shown). Using this approach, we identified a new unigene cluster (Hs.66194) of the ESTs which were derived from kidney libraries and represented a novel gene. Further *in silico* analysis showed that this gene was also expressed in the brain as well as the kidney. There were a total of 36 ESTs in this cluster, of which 8 were from brain and 28 from kidney libraries (Fig. 1).

Identification and characterization of the putative full-length CTXN3 gene in silico. By using the BLAT program on the UCSC server (Fig. 1 and Table I), human mRNA and EST records in the unigene (Hs.66194) were aligned with human genomic sequences NT_034772.5, from 127012635-127022221 on chromosome 5q23.2. Alignment identified 14 ESTs and mRNAs, 4 from brain and 13 from kidney libraries, which matched perfectly with the genomic draft sequence and possessed consensus intron/exon splicing sites. The numbers

Table I. *In silico* cloning of CTXN3 alternative transcribed variants based on both genomic and EST sequences from the UCSC genome browser.[a]

Accession no.	Tissue	Exon 1a	Exon 1b	Exon 2	Exon 3
Assembled KABE 1			127016639-127016782	127016908-127017014	127021014-127022206
Assembled KABE 2		127012657-127012979		127016908-127017014	127021014-127022206
CK001565	Brain	127012691-127012979		127016908-127017014	127021014-127021282
Z45776	Brain	127012903-127012979		127016908-127017014	127021014-127021058
CK004534	Brain	127012635-127012979		127016908-127017014	127021014-127021193
R14758	Brain	127012682-127012979			
BC071790	Kidney		127016645-127016782	127016908-127017014	127021014-127022221
BG425901	Kidney		127016640-127016782	127016908-127017014	127021014-127021367
BG402313	Kidney		127016640-127016782	127016908-127017014	127021014-127021437
BG433504	Kidney		127016645-127016782	127016908-127017014	127021014-127021416
BG432730	Kidney		127016670-127016782	127016908-127017014	127021014-127021556
BG429351	Kidney		127016670-127016782	127016908-127017014	127021014-127021609
AI791379	Kidney		127016687-127016782	127016908-127017014	127021014-127021333
BX090173	Kidney		127016688-127016782	127016908-127017014	127021014-127021514
BG399986	Kidney		127016699-127016782	127016908-127017014	127021014-127021425
AW026032	Kidney		127016706-127016782	127016908-127017014	127021014-127021214
AW242612	Kidney		127016735-127016782	127016908-127017014	127021014-127021197
AI627531	Kidney				127021133-127021754
AW243420	Kidney				127021336-127021755
AW772097	Kidney				127021336-127021755
BE701653	Brain				127021438-127021765
BI760011	Kidney				127021584-127022219
AI766027	Kidney				127021675-127022216
AW264294	Brain				127021677-127022222
AI732985	Kidney				127021695-127022220
AI935588	Kidney				127021708-127022213
AW264633	Brain				127021713-127022221
AI492360	Kidney				127021728-127022222
AI590172	Kidney				127021732-127022220
AI479278	Kidney				127021738-127022223
AI632715	Kidney				127021773-127022220
AW271385	Kidney				127021775-127022220
AI761971	Kidney				127021795-127022212
AI470959	Kidney				127021804-127022221
AI625842	Kidney				127021818-127022220
AA865799	Kidney				127021830-127022220
AA975145	Kidney				127021929-127022221
Z41419	Brain				127021818-127022220

[a]Human mRNA and EST records in the unigene (Hs.66194) were aligned with human genomic sequences NT_034772.5, from 127012635-127022221 on chromosome 5q23.2. The numbers in the table indicate the position relative to the human genomic contig sequence NT_034772.5 on chromosome 5q23.2.

in Table I indicate the position relative to the human genomic contig sequence NT_034772.5 on chromosome 5q23.2.

Four possible exons existed according to all of the ESTs. We named these exon 1a, exon 1b, exon 2 and exon 3, respectively (Fig. 1). Exon 2 and 3 were constitutive exons that were included in both of the two alternative transcripts. Most of the ESTs were from kidney libraries and contained exon 1b, exon 2 and exon 3. Representative cDNA clones included BC071790, BG425901, and BG402313. In addition, four other clones (CK001565, Z45776, CK004534, R14758) were similar to the clones from the kidney except that they contained exon 1a rather than exon 1b in the kidney isoforms (Fig. 1), suggesting that there were two alternatively spliced forms in this gene.

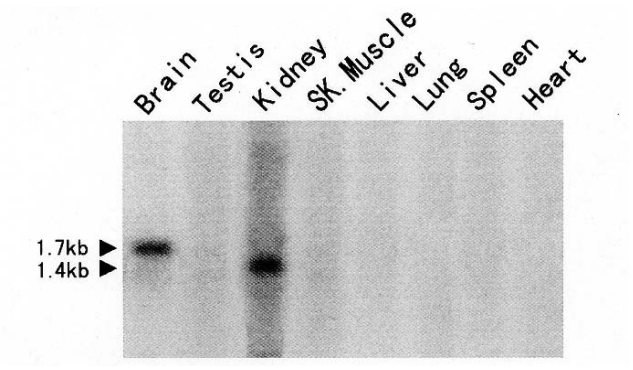

Figure 2. Northern blot analysis of CTXN3 mRNA expression. Two bands of ~1.4 and 1.7 kilo base (kb) in size were detected in kidney and brain tissue, respectively.

By using the above described SGC-ESTGS approach, we determined the possible exon-intron boundaries and the putative full-length cDNA sequences of the two variants. Every base sequence was further confirmed by both genomic draft sequence and the available ESTs, and exon-intron boundary conformed to the GT/AG regulation.

Computational analysis showed that an alternative usage of the first exon produced two types of mRNAs in CTXN3. In addition, according to the EST information, the two alternative transcribed variants showed completely different tissue expression characteristics. This information suggested that alternative splicing regulates the tissue-specific expression of CTXN3.

The ESTs are located on chromosome 5q23.2 of the human genome by comparing EST assembly with the human genome sequence. To experimentally verify the above *in silico* analysis, we designed sense primers specific for each first exon, exon 2 and a common antisense primer specific to the constitutive exon 3 listed in Fig. 1.

Tissue-specific expression of CTXN3. To determine experimentally the tissue expression and the transcript size of CTXN3, Northern blot analysis was performed on blots containing mRNAs from 8 adult tissues. To do this, a PCR-generated DNA fragment from CTXN3 cDNA (492 bp) common to both known isoforms was obtained with CTXN3-specific primer pair c and d, and the digested fragment was used as a probe for CTXN3 detection on a Northern blot. As shown in Fig. 2, a band of ~1.4 kilo base (kb) in size was detected in the kidney, whereas the transcript size in the brain was 1.7 kb. The 1.4- and 1.7-kb transcripts were in agreement with the above bioinformatics-predicted alternative splicing variants. No signal was detected with RNAs from other tissue.

To verify and extend this analysis we used a more sensitive reverse transcriptase-polymerase chain reaction (RT-PCR) method for expression analysis on a panel of cDNAs isolated from many normal tissues including brain, kidney, heart, liver, lung, pancreas, and colon using primer pairs specific for exon 1a, 1b or 2 with a common primer specific to exon 3 designed based on the *in silico* cloned cDNA sequence (Fig. 1). As shown in Fig. 3, three specific bands, ~492, 640 and 816 bp in size, respectively, were detected in adult brain, fetal brain and kidney tissue as expected from the Northern blot. Sequencing of PCR products confirmed that the predicted sequence was accurate. Because of its specific expression in the kidney and brain, we name this gene KABE representing the gene expressed in the kidney and brain; the HUGO Gene Nomenclature Committee has named it CTXN3. Together, through Northern blot and RT-PCR, the brain form variant was detected in the brain only, whereas the kidney form was found in the kidney only. These results indicate a tissue-specific regulation of CTXN3 expression.

Full-length cDNA cloning of human CTXN3. By applying the methods of rapid amplification of cDNA ends (5' RACE) and PCR, the two CTXN3 transcript variant full-length cDNAs, composed of 1660 and 1458 bp, respectively, were obtained (Fig. 4). The RT-PCR and 5' RACE sequencing results further confirmed that the predicted novel gene was accurate. In addition, the *in silico*-assembled cDNAs were extended 66 bp and 12 bp at the 5' end by subjecting them to 5' RACE.

Bioinformatics analysis of the predicted amino acid. BLAST (23) analysis for the complete nucleotide sequence of the cDNA (GenBank accession nos. AB219764 and AB219832) demonstrated that it was a novel gene. The sequenced AB219764 and AB219832 cDNA consisted of 1660 and

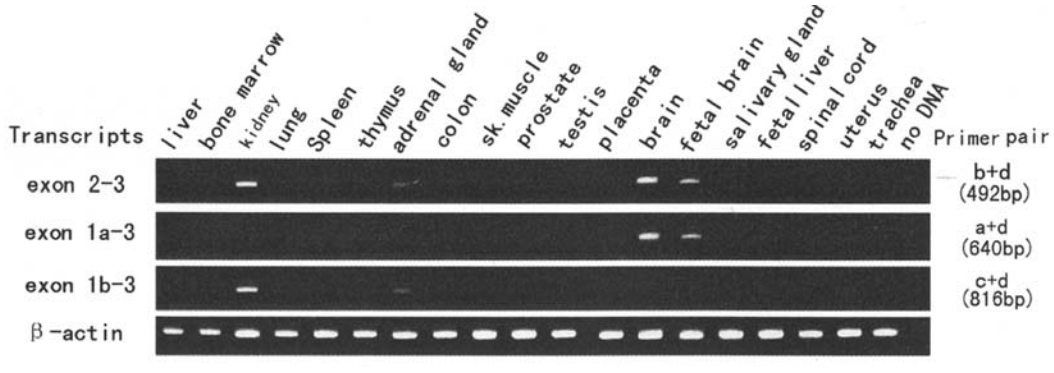

Figure 3. RT-PCR expression analysis of the alternative transcribed transcripts of CTXN3 in adult human tissues (selective usage of first exon 1a or 1b). A 492-bp fragment spanning exon 2 to exon 3 is shown in the top panel and 640- and 816-bp fragments are shown in the second and third panel, representing exon 1a to exon 3 and exon 1b to exon 3, respectively.

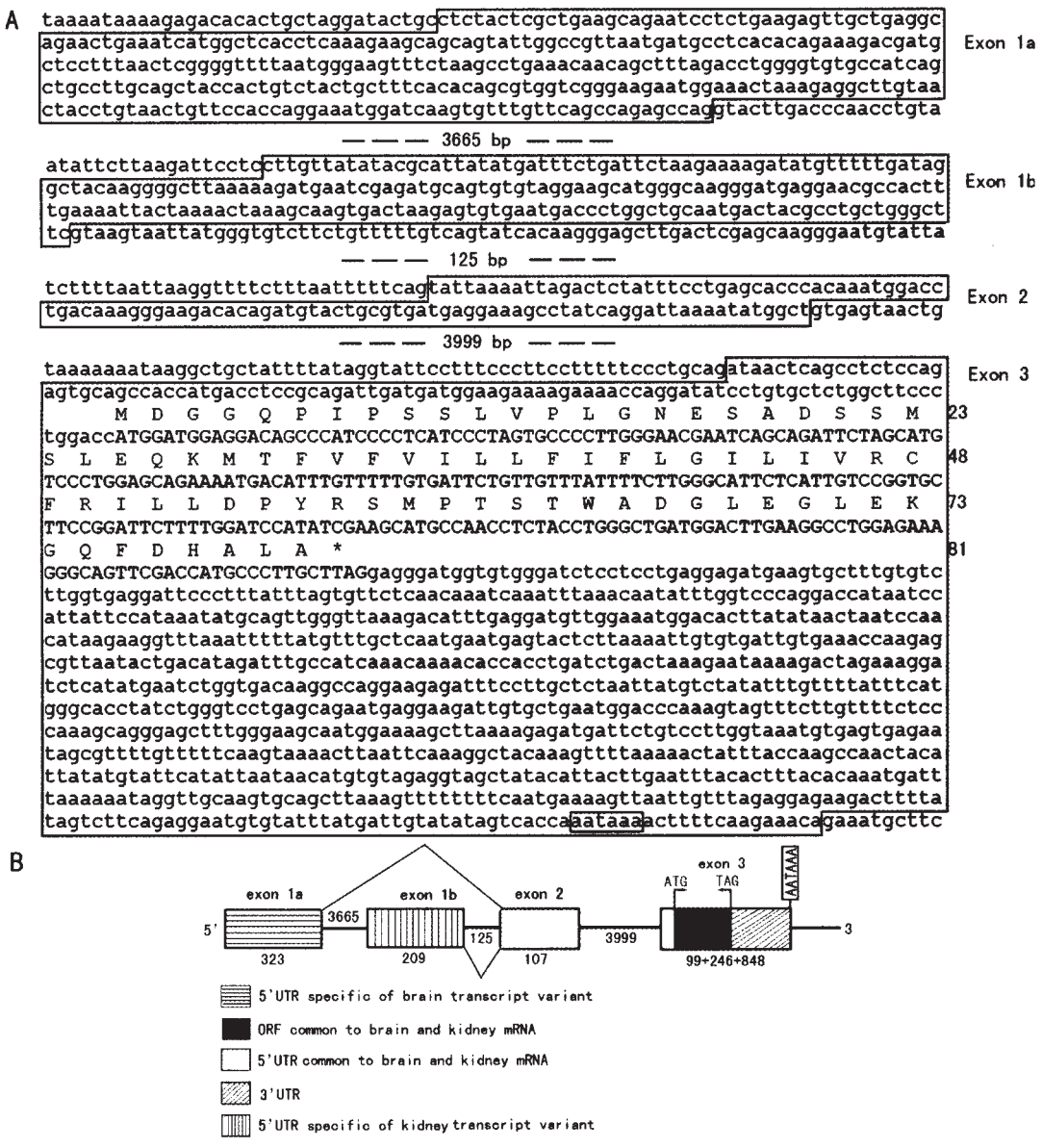

Figure 4. Sequence and gene structure of the human CTXN3 gene. (A) The cDNA sequence of human CTXN3 (GenBank accession nos. AB219832 and AB219764) is shown. Two alternative first exons are numbered 1a and 1b. The sequence was mapped onto the human genome sequence to show exon/intron structure. The predicted protein sequence is shown in single letter code. The polyadenylation site is capitalized. (B) A diagram of CTXN3 gene organization showing the pattern of alternative splicing. Coding sequence (ORF), UTR and intron sizes are also shown.

1458 bp respectively, and contained a longest putative 246-bp ORF capable of encoding a predicted 81-amino acid peptide (Fig. 4). The sequence surrounding the presumed start codon had an adequate context for a translation initiation site, an A in position -3 and a G in position +4, and the in-frame stop codon was present 51 bp upstream of the presumed start codon (24), suggesting that the two alternative transcripts of cDNA had the entire coding potential of the CTXN3 gene. The 3'-untranslated region (UTR) contained a potential polyadenylation signal (AATAAA) at 1094 bp downstream from the start codon. Remarkably, 3'-UTR was 848 bp long and the other 3'-UTR variants were not detected in the Northern blot analysis.

BLAST searches of GenBank databases using the predicted protein revealed weak homology with one function-unknown brain-specific protein, cortexin. The overall levels of identity and similarity between CTXN3 and cortexin (human and mouse) are 43% and 68-71%, respectively. To further predict the potential function of this novel gene, multiple prediction programs were used to analyze the amino acid sequence of the encoded protein. Protein structure prediction programs (TMHMM, PSORT II, TMpred, and PSIPRED) proposed a transmembrane region from aa 29-31 to aa 46-51 with the N-terminal outside and no signal peptide; the predicted transmembrane region is shown in Fig 5. SignalP 2.0 HMM revealed a high probability of a signal anchor (signal anchor probability, 0.996; signal peptide probability, 0.001) with maximum cleavage site probability 0.001 between residues 48 and 49. Kyte and Doolittle hydrophobicity plot of the predicted CTXN3 revealed one strong hydrophobic region in correlation with the predicted transmembrane region (data not shown). The expected

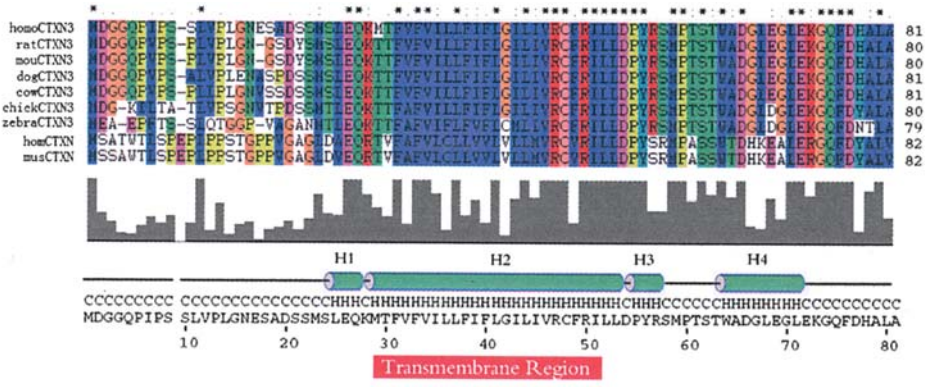

Figure 5. Multiple sequence alignment of human, rat, mouse, dog, cow, chicken and zebrafish CTXN3 protein sequences and human and mouse cortexin proteins. The bar diagram under the alignment indicates sequence conservation. Asterisks, colons, and periods over the alignment correspond to identical, conserved, and partially conserved residues, respectively. Predicted helices (H1-H4) are displayed as green cylinders; the predicted transmembrane regions are displayed as red boxes.

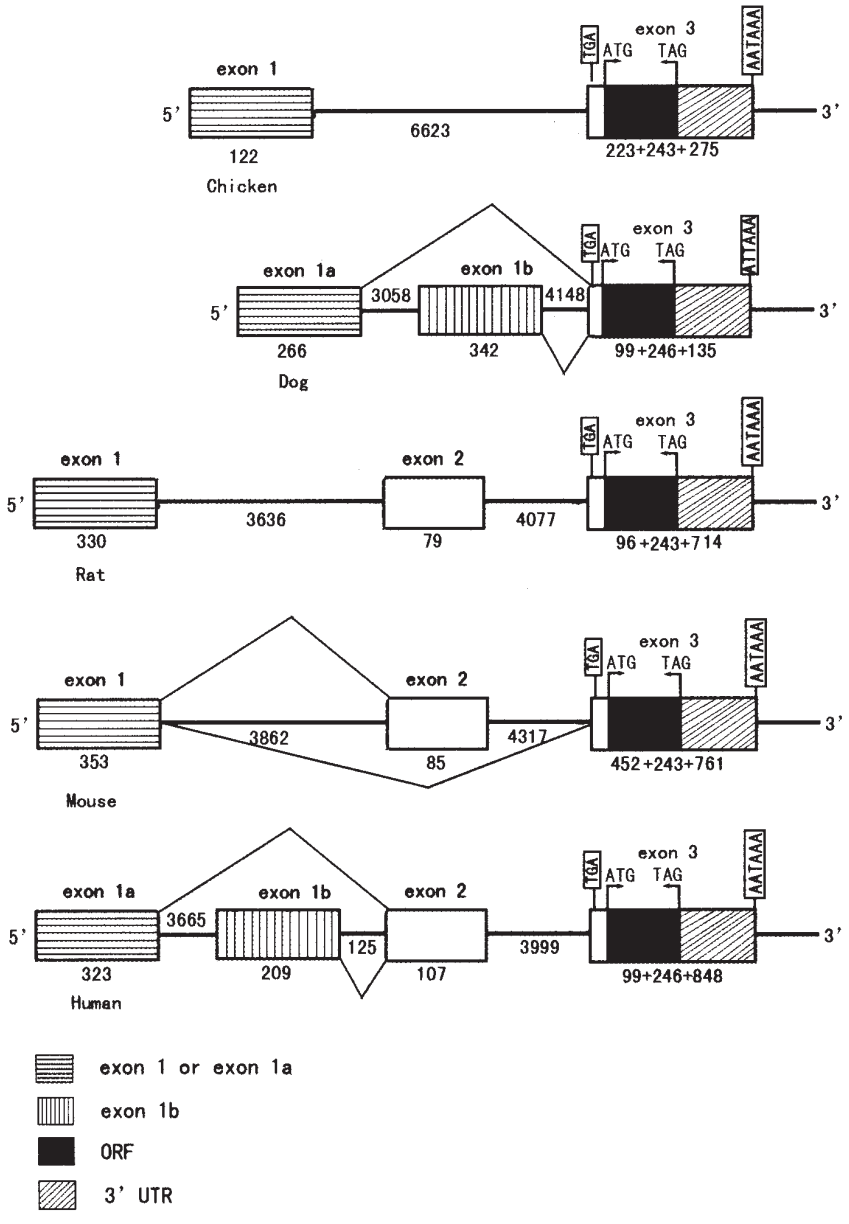

Figure 6. Schematics showing the genomic structure of CTXN3 and orthologs. Genomic structure of the human CTXN3 gene and those of vertebrate orthologs are depicted. Exons are boxed with different shades representing conserved coding sequences, additional coding regions, and UTRs. Alternative splicing events including alternative exon 1 usage and exon 2 skipping are indicated. Positions of start codons (ATG), in-frame stop codons (TGA, TAG, and TAA), and poly(A) signals (AATAAA) are marked, respectively.

	homoKABE	RatKABE	MusKABE	DogKABE	CowKABE	ChiKABE	ZebKABE	homoCort	musCortex
homoKABE		96	96	96	98	90	86	68	71
RatKABE	91		100	96	98	88	86	67	70
MusKABE	91	100		96	98	88	86	67	70
DogKABE	93	93	93		98	90	85	71	73
CowKABE	91	94	94	93		91	86	70	72
ChiKABE	80	80	80	81	80		89	71	72
ZebKABE	69	69	69	67	67	71		68	70
homoCort	43	45	45	44	45	48	46		99
musCortex	43	45	45	44	45	48	45	96	

Figure 7. Pairwise alignment scores of CTXN3 and its orthologs. Pairwise alignment scores among KABE protein sequences are summarized here in percentage values.

molecular weight was 8933.4 Da. The PSORT predicted intracellular localization of the CTXN3 polypeptide to cytoplasmic (0.391), mitochondrial (0.174), nuclear (0.174), and endoplasmic reticulum (0.130). Together, these data predict a type I membrane protein with the N-terminal outside.

The two alternative variants of CTXN3 cDNA sequences differed in the 5' untranslated region and encode the same protein. Combining the tissue-specific distribution of the two variants suggested that alternative splicing would regulate the tissue-specific expression of CTXN3.

Identification and analysis of CTXN3 orthologs. To determine if there were any CTXN3 orthologs in other species, we utilized bioinformatics to identify CTXN3 orthologs in the mouse, rat, dog, cow and zebrafish as well as in the chicken. Utilizing draft genomic sequences and ESTs deposited in the database, we assembled the CTXN3 orthologs. Except for the chicken ortholog which was predicted by TBLASTN programs, the other orthologs were assembled from the available EST sequences in the GenBank and further verified by the draft genome sequence. All the predicted orthologs had complete CDS (coding sequence). The predicted cDNA sequences of CTXN3 orthologs were submitted to the DDBJ/EMBL/GenBank under the accession nos. TPA: BR000229- BR000235, BR000238. As shown in Fig. 6, the presence of in-frame upstream stop codons is located upstream of the 51 region of the initiation site without any possible ATG codon between them in all CTXN3 vertebrate ortholog cDNA sequences, and the conservation of an ATG in species starting just from this methionine strongly support the above predicted ORF and also suggest that the genomic and the EST sequences for these species in the database are correct.

However, the genomic organization of human and vertebrate CTXN3 genes showed a difference in their exon numbers possibly due to exon duplication during evolution. In addition, comparative analysis of the cDNA sequences identified the presence of variant alternative splicing, such as alternative exon 1 usage in the human and the dog, and even within a species due to a differential splicing event, resulting in species-specific patterns in their transcripts.

The deduced amino acid sequence analysis of CTXN3 and its orthologs show that this protein is well conserved among different species (Fig. 5). As shown in Fig. 7, the amino acid sequence identity of human CTXN3 is 91.3% to rat, mouse and cow, 92.5% to dog, 80.2% to chicken, 69.1% to zebrafish, and 42.6% to human and mouse cortexin. The amino acid sequence similarity of CTXN3 is 98% to cow, 96% to rat, mouse and dog, 90% to chicken, 86% to zebrafish, 68-71% to human and mouse cortexin. In addition, the amino acid sequence of the transmembrane domain of the human CTXN3 gene is highly conserved with those of the human and mouse cortexin genes (Fig. 5). Although no reported functional domain regarding this transmembrane region is yet available, highly conserved amino acid sequences between these CTXN3 and cortexin genes suggest the possibility that two human genes (CTXN3 and cortexin) may have originated from a common ancestor gene.

CTXN3 expression appears to be brain-specific but also expressed in the kidney in humans, while these vertebrate CTXN3 proteins exhibit distinct tissue expression patterns, though all three are enriched in the brain, consistent with the initial observation in human. Interestingly, for six of these genes, the ESTs were represented almost exclusively in brain libraries. These findings imply that the generation of functionally different forms of CTXN3 by alternative splicing of CTXN3 transcripts is a species-specific event, possibly indicating species-specific mechanisms for regulating CTXN3 activities.

Discussion

In this report we described the identification of a new gene, CTXN3, selectively expressed in the kidney and brain. Two alternative transcript variants encoded for an 81-amino acid protein with a predicted molecular weight of 8933.4 Da in size and contained a predicted single membrane-spanning domain. The CTXN3 gene has 3 exons and spreads over an 9.6-kb region of human chromosome 5q23. Orthologs of CTXN3 have been identified in the mouse, rat, cow, dog, zebrafish and chicken.

Molecular cloning *in silico* is becoming a primary procedure to assemble full-length mRNA sequences from ESTs with the coming of bioinformatics. Several studies (14,25,26) have demonstrated that the utility of digital subtraction strategies (DDD or DGED), and experimental

verification are effective approaches to identify tissue-specific genes. Using DGED, we obtained the expected results in this study. The predicted gene sequence and tissue-specific alternative splicing were also demonstrated by experiment. Therefore, it was possible to correctly clone *in silico* novel genes based on both ESTs and genomic sequence. In contrast to conventional EST assembly approaches, SGC-ESTGS will recover all potential putative exon combinations, regardless in which order the data are processed. This eliminates much of the ambiguities in current EST assembly algorithms. For example, given the set of displayed ESTs in Fig. 1, there were two different ways of assembling (partitioning) all inputed ESTs into consensus sequences. Both reconstructions were equally computable from the data and explained all ESTs, but the difference was quite considerable. Depending on the order of the processed ESTs, a conventional approach might have resulted in either reconstruction overlooking the other. In contrast, a splicing graph-based approach does not partition the data but reports exhaustively all two different putative transcripts.

The generation of alternative transcripts is an important regulatory mechanism of gene function now apparent with the high-throughput sequencing of genomes and expression libraries. In this study, we identified two alternative transcript forms of CTXN3 that represent a potential mechanism for tissue-specific target gene regulation. Because the transcribed variants were produced by alternative first exon, belonging to the 5' UTR of the CTXN3 gene (Fig. 4), this indicated that not only coding sequence variants have obvious consequences (27), but variants in noncoding sequences can also be functionally significant for regulating tissue-specific expression. It is interesting to speculate that these alternative UTRs are expressed as a result of alternative promoter use and may function to regulate CTXN3 expression at the post-transcriptional level, by altering translation efficiency and/or transcript stability (28,29). The alternative exon 1 usage of KABE supported the above mechanism. The identity and function of the alternative promoters remain to be determined.

Comparative genomics is very important to identify gene function and structure. Over the past several years the increasing number of complete or nearly complete genome sequences has enabled scientists to examine the process of evolution (30). Among the complete or nearly complete genome sequences of the most important metazoan organisms in the public database are *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* (human), *Fugu rubripes* (pufferfish), *Anopheles gambiae* (mosquito), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *C. intestinalis* (sea squirt). We analyzed these genome sequences and EST databases to identify the orthologs of CTXN3. We successfully identified the orthologs of CTXN3 in the mouse, rat, dog, cow, chicken, and zebrafish. In addition to a very high sequence identity at the amino acid level, the human, the mouse, and the rat orthologs showed very similar genomic organization (Fig. 6), suggesting that the CTXN3 gene is well conserved among these species.

In conclusion, identification of CTXN3 validates bio-informatics-based gene discovery to be an effective approach which can identify genes differentially expressed in different tissues. CTXN3 is a novel kidney- and brain-specific gene. CTXN3 shares homology with known brain-specific protein cortexin, and its expression in the kidney and the brain suggests a role in physiological function. We further validated that expression of CTXN3 was highly regulated in a tissue-specific manner producing two alternative splice variants. This together with the EST expression analysis in species, implies that CTXN3 protein may play an important role in kidney and brain development. The existence of the mouse ortholog of CTXN3 will be advantageous in generating knock-out mice to understand the biological functions that CTXN3 plays *in vivo*. Furthermore, it is important to examine differential promoter usage *in vivo* during brain and kidney development, which will likely require the generation of selective gene-targeting approaches to mark the usage of each promoter individually.

Acknowledgements

This study was supported, in part, by Tianjin Medical University Annovation Grant to H.T. Wang (no. 200505) and supported by the Fund of the Health Bureau of Tianjin to HT Wang (no. 06KZ57).

Sequence data: Nucleotide sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Data Libraries under accession nos. AB219832 and AB219764 (NM_001048252.1) for human CTXN3 alternative transcript variants. Nucleotide sequence data reported for mouse, rat, dog, cow, chicken, and zebrafish CTXN3 orthologs are available in the Third Party Annotation Section of the DDBJ/EMBL/GenBank databases under accession nos. TPA: BR000229-BR000235, BR000238.

References

1. Maniatis T, Goodbourn S and Fischer JA: Regulation of inducible and tissue-specific gene expression. Science 236: 1237-1245, 1987.
2. Hakalahti L, Vihko P, Henttu P, Autio-Harmainen H, Soini Y and Ihko R: Evaluation of PAP and PSA gene expression in prostatic hyperplasia and prostatic carcinoma using Northern blot analyses, *in situ* hybridization and immunohistochemical stainings with monoclonal and bispecific antibodies. Int J Cancer 55: 590-597, 1993.
3. Troyer JK, Beckett ML and Wright GL Jr: Detection and characterization of the prostate-specific membrane antigen (PSMA) in tissue extracts and body fluids. Int J Cancer 62: 552-558, 1995.
4. Zhang Y, Proenca R, Maffei M, Barone M, Leopold L and Friedman JM: Positional cloning of the mouse obese gene and its human homologue. Nature 372: 425-432, 1994.
5. Fickett JW: Coordinate positioning of MEF2 and myogenin binding sites. Gene 172: GC19-GC32, 1996.
6. Fickett JW: Quantitative discrimination of MEF2 sites. Mol Cell Biol 16: 437-441, 1996.
7. Maser RL and Calvet JP: Analysis of differential gene expression in the kidney by differential cDNA screening, subtractive cloning, and mRNA differential display. Semin Nephrol 15: 29-42, 1995.
8. Takenaka M, Imai E, Kaneko T, *et al*: Isolation of genes identified in mouse renal proximal tubule by comparing different gene expression profiles. Kidney Int 53: 562-572, 1998.
9. Hu E, Chen Z, Fredrickson TA, *et al*: Rapid isolation of tissue-specific genes from rat kidney. Exp Nephrol 9: 156-164, 2001.
10. Lisitsyn N and Wigler M: Cloning the differences between two complex genomes. Science 259: 946-951, 1993.
11. Liang P and Pardee AB: Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257: 967-971, 1992.

12. Diatchenko L, Lau Y-FC, Campbell AP, *et al*: Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA 93: 6025-6030, 1996.
13. Strausberg RL, Greenhut SF, Grouse LH, Schaefer CF and Buetow KH: *In silico* analysis of cancer through the Cancer Genome Anatomy Project. Trends Cell Biol 11: S66-S71, 2001.
14. Shen D, He J and Chang HR: *In silico* identification of breast cancer genes by combined multiple high throughput analyses. Int J Mol Med 15: 205-212, 2005.
15. Kent WJ: BLAT - the BLAST-like alignment tool. Genome Res 12: 656-664, 2002.
16. Nakai K and Horton P: PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24: 34-36, 1999.
17. Hofmann K and Stoffel W: TMbase - A database of membrane spanning protein segments. Biol Chem Hoppe-Seyler 374: 166, 1993.
18. Moller S, Croning MDR and Apweiler R: Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17: 646-653, 2001.
19. McGuffin LJ, Bryson K and Jones DT: The PSIPRED protein structure prediction server. Bioinformatics 16: 404-405, 2000.
20. Walker JM (ed): The Proteomics Protocols Handbook. Humana Press, pp571-607, 2005.
21. Higgins DG, Thompson JD, Gibson TJ, *et al*: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680, 1994.
22. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG: The clustal x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876-4882, 1997.
23. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: Basic local alignment search tool. J Mol Biol 5: 403-410, 1990.
24. Kozak M: An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15: 8125-8148, 1987.
25. Olesen C, Hansen C, Bendsen E, *et al*: Identification of human candidate genes for male infertility by digital differential display. Mol Hum Reprod 7: 11-20, 2001.
26. Scheurle D, DeYoung MP, Binninger DM, Page H, Jahanzeb M and Narayanan R: Cancer gene discovery using digital differential display. Cancer Res 60: 4037-4043, 2000.
27. Wistow G, Bernstein SL, Wyatt MK, *et al*: Expressed sequence tag analysis of human retina for the NEIBank Project: retbindin, an abundant, novel retinal cDNA and alternative splicing of other retina-preferred gene transcripts. Mol Vis 8: 196-204, 2002.
28. van der Velden AW and Thomas AA: The role of the 5' untranslated region of an mRNA in translation regulation during development. Int J Biochem Cell Biol 31: 87-106, 1999.
29. de Moor CH and Richter JD: Translational control in vertebrate development. Int Rev Cytol 203: 567-608, 2001.
30. Ureta-Vidal A, Ettwiller L and Birney E: Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4: 251-262, 2003.