J Hum Genet (1999) 44:337–342 © Jpn Soc Hum Genet and Springer-Verlag 1999

SHORT COMMUNICATION

Shiro Ikegawa · Minoru Isomura · Yu Koshizuka
Yusuke Nakamura

Cloning and characterization of human and mouse *PROSC* (proline synthetase co-transcribed) genes

Received: April 6, 1999 / Accepted: May 11, 1999

Abstract Large-scale DNA sequencing, coupled with *in silico* gene trapping, is a robust approach to identifying unknown genes in selected genomic regions. Using this approach we have isolated a novel human gene, *PROSC* (for proline synthetase co-transcribed [bacterial homolog]), from human chromosome 8p11.2, and its mouse counterpart. The human *PROSC* gene spanned 17 kb of genomic DNA; its cDNA was 2530 bp long, with 8 exons that included an open reading frame of 825 bp (275 amino acids). The mouse cDNA (*Prosc*), 1995 bp long, was predicted to encode 274 amino acids. *PROSC* is ubiquitously expressed in human tissues and has been highly conserved among divergent species from bacteria to mammals, suggesting its important cellular function. The gene product is likely to be a soluble cytoplasmic protein, but its function remains to be determined.

Key words Cloning · Mapping · *PROSC* gene · Large-scale DNA sequencing · *In silico* gene trapping

Introduction

Recent improvements in sequencing technologies and informatics have drastically increased the speed and efficiency of efforts to identify unknown genes. The development of automated fluorescent sequencers, in concert with high-capacity computers and sophisticated assembly programs, have made high-throughput genomic sequencing possible. New data can now be linked easily to a vast body of archived information in public databases, specifically expressed sequences (ESTs) that have been generated by the Human Genome Project. The 800,000 currently available ESTs are considered to represent 40,000–50,000 genes (Rowen et al. 1997), and that number is growing. In addition, computational analyses using gene-finder programs such as GRAIL (Uberbacher and Mural 1991) and FEXH (Solovyev et al. 1994) predict exons from anonymous genomic sequences with reliable sensitivity and accuracy (Claverie 1997; Elkahloun et al. 1997; Ishikawa et al. 1998). Thus, sequencing large genomic regions and "trapping" genes within those sequences using computer software has become a highly efficient and powerful approach for identification of previously unknown genes (McKusick 1997).

We have been determining nucleotide sequences of genomic DNA fragments from human chromosomes 3p22–p21.3 (Ishikawa et al. 1998; Daigo et al., 1999), 8p11.2, and 8p21 (Isomura et al., manuscript in preparation) and identifying genes in those regions (Ikegawa et al. 1999a, b). During this effort we sequenced a 1.8-Mb fragment of 8p11.2 and subsequently trapped a gene, designated *PROSC* (for proline synthetase co-transcribed [bacterial homolog]), that was highly homologous to putative genes of bacterial species associated with proline synthetases. Here we report isolation, characterization, and fine mapping of this evolutionarily conserved and, by implication, biologically interesting gene.

Materials and methods

Construction of sequence-ready contig of a genomic region on human 8p11.2

A cosmid contig representing a 1.8-Mb segment of human chromosome 8p11.2 was constructed from overlapping CEPH YACs (937_b_9 and 854_f_6) according to methods described previously (Murata et al. 1994). The details of the physical map and information about the 1.8-Mb contig and genomic clones were deposited in Japan Science and Tech-

S. Ikegawa (✉) · M. Isomura · Y. Koshizuka · Y. Nakamura
Laboratory of Genome Medicine, Institute of Medical Science, Human Genome Center, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
Tel. +81-3-5449-5233; Fax +81-3-5449-5406
e-mail: sikegawa@ims.u-tokyo.ac.jp

Y. Koshizuka
Department of Orthopaedic Surgery, Faculty of Medicine, University of Tokyo, Tokyo, Japan

nology Corporation Advanced Lifescience Information System (http://www-alis.tokyo.jst.go.jp/HGS/top.html). Several gaps in the cosmid contig were filled by BAC or PAC clones that we isolated using a down-to-well BAC/PAC screening system (Genome Systems, St. Louis, MO, USA), or the Tukuba PAC screening system (Oligoservice, Tukuba, Ibaragi, Japan) according to the manufacturers' protocols.

Large-scale sequencing

Clones representing the minimal tiling path were sequenced by shotgun and primer-walking strategies (Ishikawa et al. 1998). Briefly, cosmids, BACs, and PACs were fragmented by means of an ultrasonic disrupter (Tomy, Tokyo, Japan), and the DNA fragments were separated by electrophoresis in 0.8% agarose gels. Fractions 2–5 kb long were excised from the gel and recovered by electrodialysis. The fractionated DNAs were subcloned into plasmid vectors, those from cosmids into pBC and those from BACs and PACs into pBSII-SK(−). The recombinant plasmids were prepared using an automatic DNA extraction machine (PI-100; Kurabo, Osaka, Japan). The shotgun clones were sequenced by means of the ABI377 automated sequencer and the dRhodamine terminator cycle-sequencing FS ready reaction kit (ABI), using T3 and T7 universal primers. Nucleotide sequences were determined by sequencing more than ten subclones per kilobase of the source clones. These shotgun sequences were assembled using the Phred software program (Ewing et al. 1998). Remaining gaps between assembled segments were filled by sequencing linking-plasmid clones obtained by primer-walking.

RT-PCR and direct sequencing of the human and mouse *PROSC* genes

The entire putative coding sequence of a human gene, *PROSC*, which was trapped from the sequenced region by exon-prediction software and by identity with ESTs in the database, was amplified with primers h/F09.ORF/f (GGGGGATGTGGAGAGCTGG) and h/F09.ORF/r (TTTCCCTGGCTCAGTGCTCC) using human testis and fetal liver cDNAs as templates. The putative coding sequence of the mouse homolog was amplified by primers m/F09.ORF/f (GAGCTGGGAGTCGGGTTC) and m/F09.ORF/r (CGGCCATCAGTTTGTTGAC), using mouse testis cDNA as a template. The PCR products were sequenced directly on both strands with an ABI377 autosequencer.

5′-RACE

5′-RACE (rapid amplification of cDNA end) was performed using the Marathon cDNA amplification kit (Clontech, Palo Alto, CA, USA) according to the manufacturer's protocol. Human testis and fetal lung RNAs (Clontech) were used as templates.

Database analysis

The BLAST program was used to search for similarity of the *PROSC* sequence to known DNA and protein sequences. Exon prediction was performed using GRAIL (version 1.3) and FEXH. Comparisons of amino acid sequences among different species were performed with the DNASIS program (Hitachi Software, Tokyo, Japan).

Northern blot analysis

PCR products purified with Suprec II (Takara Shuzo, Ohtsu, Japan) and randomly labeled with [^{32}]P were used as the probes in human and mouse multiple-tissue Northern blot systems (Clontech). The human probe was amplified with primers h/F09.ORF/f and h/F09.ORF/r, and the mouse probe with primers m/F09.ORF/f and m/F09.ORF/r. Prehybridization, hybridization, and washing were done according to the manufacturer's instructions. The membranes were autoradiographed at −80°C for 36 h with intensifying screens.

Results and discussion

Isolation of human *PROSC* cDNA

GRAIL analysis of one of the cosmids (c4545) present in the 1.8-Mb contig predicted five genomic segments, all aligned in the same direction, as exons with "excellent" scores. In addition, ESTs AA852337, AA463379, and AA310517 (Genebank) were found to overlap with the GRAIL-predicted exons. EST-walking revealed that these three ESTs were able to compose a single transcript with a single open reading frame (ORF) of about 800 bp. A database search using BLAST revealed that its predicted amino acid sequence was highly homologous to the product encoded by a *C. elegans* gene, F09.08 (Wilson et al. 1994). RT-PCR experiments designed to cover the entire putative coding sequence yielded a single-band product of the expected size, confirming the existence of the gene.

Sequence information from the EST-walking, RT-PCR, and 5′-RACE experiments were integrated with the large-scale genomic sequence to determine the cDNA sequence of the trapped gene, designated *PROSC* (DDBJ accession number, AB018566). The cDNA was 2530 bp long with an ORF of 825 bp (Fig. 1). The 1669-bp 3′-untranslated region contained an *Alu*-like sequence at nucleotides 1438–1598, 78.5% identity). The cDNA was considered to be full length because the size of the clone corresponded well to the size of the *PROSC* mRNA indicated by northern blotting. Two possible initiating methionine codons were present, one at nt 37–39, and another at nt 55–57. The latter provided a better alignment with the initiating methionines of *PROSC* sequences from other species (Fig. 2) and was compatible with the Kozak consensus (GGC AGC ATG T), while the former was not (CGG GGG ATG T) (Kozak 1986). Never-

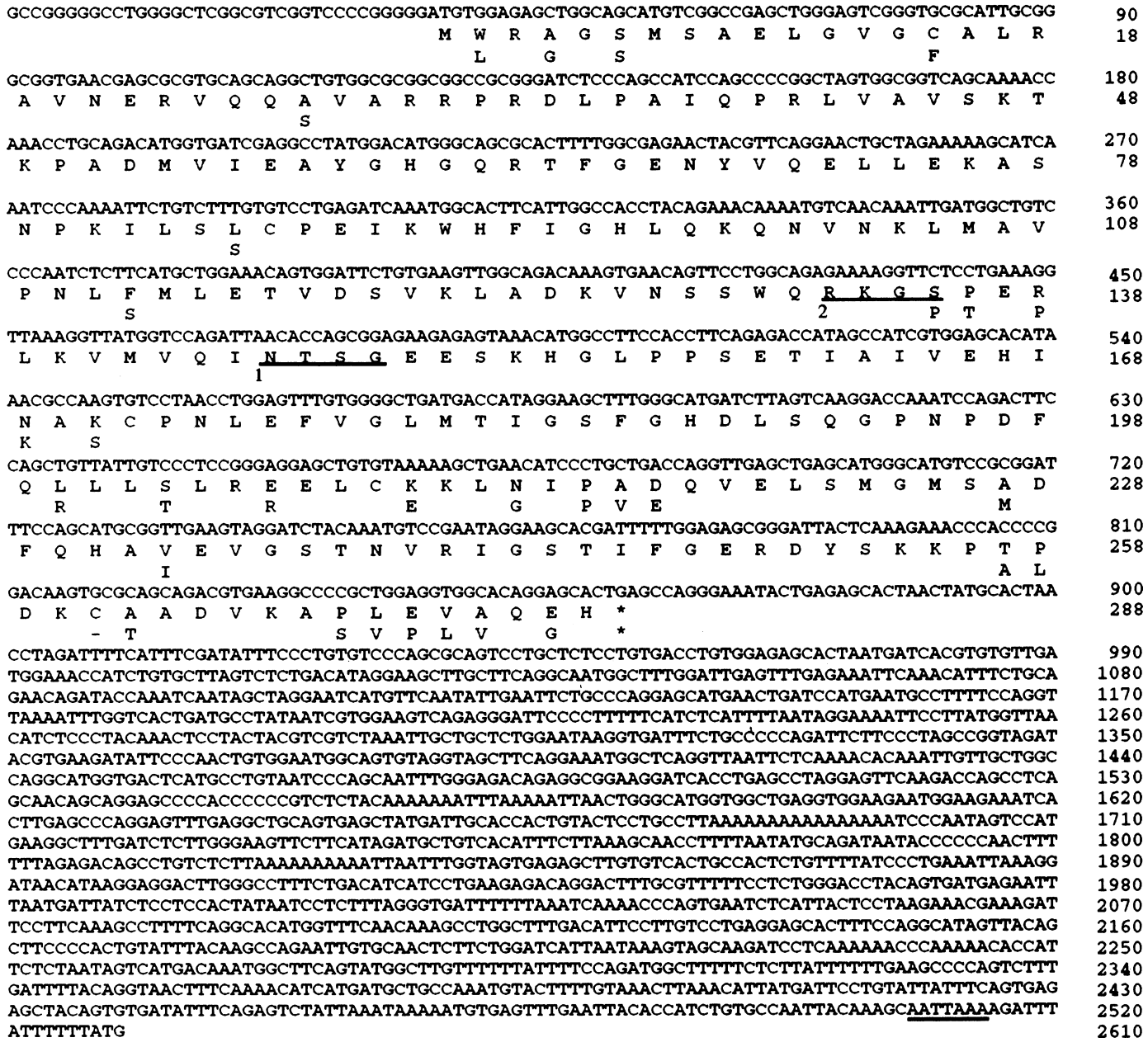

Fig. 1. Nucleotide (*upper rows*) and deduced amino acid (*middle rows*) sequences of human *PROSC* and deduced amino acid sequences of mouse *PROSC* (*lower rows*). Only amino acids different from those of the human protein are indicated for the mouse. Numbering at *right* refers to the human nucleotide and amino acid sequences. The putative polyadenylation signal (AATTAA) is *underlined*; termination codons are indicated by *asterisks*. Protein motifs are also *underlined* and identified with numbers as follows: *1*, putative N-linked oligosaccharide attachment sites; *2*, a cAMP- and cGMP-dependent protein kinase phosphorylation site. The human and mouse *PROSC* cDNA sequences are deposited in DDBJ under accession numbers AB018566 and AB018567, respectively

theless, because the presence of an untranslated methionine upstream is unusual, we assumed that the ATG codon at nt 37–39 was the actual initiating methionine.

Isolation of the mouse *Prosc* gene

A database search revealed that multiple mouse ESTs had significant homology to the human *PROSC* sequence. EST-walking from these mouse ESTs revealed they also could compose a single transcript, with a predicted amino acid sequence highly homologous to the human and *C. elegans* counterparts. The putative coding sequence was amplified by primers m/F09.ORF/f and m/F09.ORF/r, to yield a single-band RT-PCR product of the expected size. Direct sequencing of PCR products and EST-walking determined the mouse cDNA sequence, *Prosc* (DDBJ accession number, AB018567). The 1995-bp cDNA encoded 274 amino acids with 86.5% sequence similarity to its human counterpart (Table 1).

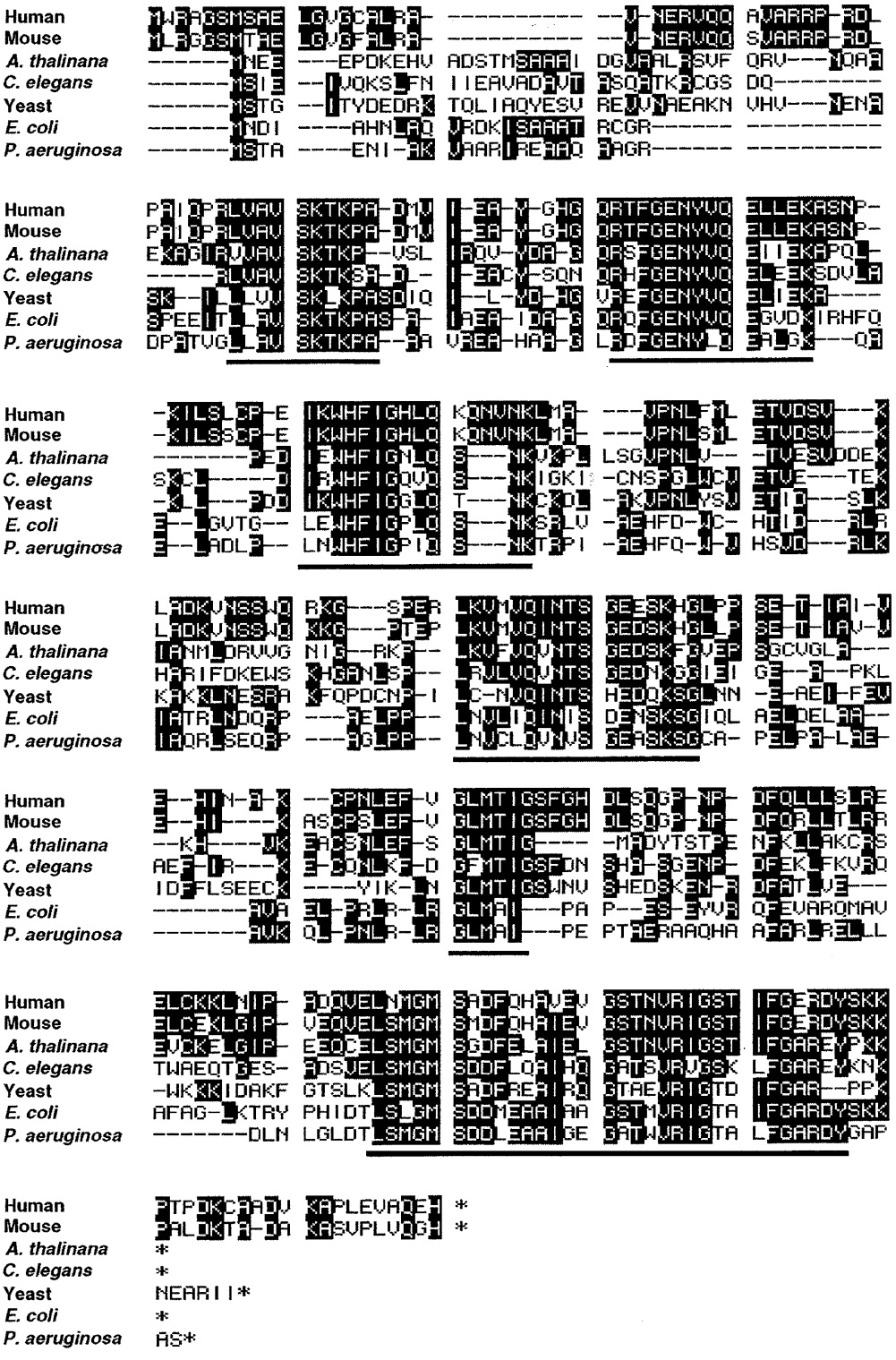

Fig. 2. Alignment of amino acid sequences of *PROSC* genes among different species. *Black background* indicates conserved residues; conserved regions are *underlined*

Primary structure of the human *PROSC* gene product

Human *PROSC* encoded a polypeptide of 275 amino acids. Hydropathy analysis using SOSUI (http://www.tuat.ac.jp//cgi/~mitaku/) predicted it would constitute a soluble protein. PSORT II (http://psort.nibb.ac.jp:8800/) predicted a cytoplasmic molecule with no N-terminal signal peptide. However, the deduced gene product did contain a putative N-linked oligosaccharide attachment site (N-{P}-[ST]-{P}) at amino acids (aa) 146–150 (NTSG), which was a conserved element. It also contained a cAMP- and cGMP-dependent protein kinase phosphorylation site ([RK](2)-x-[ST]) at aa 132–135 (RKGS).

A search of the public database showed that the amino acid sequence of human *PROSC* possessed significant homology to F09E5.8 (YU68_CAEEL), a hypothetical 27.2-

kDa protein encoded in chromosome II of *C. elegans* (Wilson et al. 1994) (Table 1). Significant homology existed also with respect to hypothetical proteins of yeast (YBD6_YEAST; De Wergifosse et al. 1994) and *Arabidopsis thaliana* (F12F1_20). Furthermore, a variety of hypothetical bacterial proteins also showed significant homology (about 30% identity for the entire sequences); these included a 24.5-kDa protein from *Pseudomonous aeruginosa* (YPT5_PSEAE), *Bacillus subtilis* protein ylmE, *E. coli* protein YGGS, *Helicobacter pylori* protein HP0395, *Mycobacterium tuberculosis* protein MtCY270.20, and a protein in the pilT 5′-region of *Vibrio alginolyticus*. All these proteins, which range in size from 24 to 30 kDa, contain a number of conserved regions.

Comparison of amino acid sequences among divergent species revealed that the following six elements were conserved during evolution: (i) [LV]-[VL]-[AV]-V-S-K-[TL]-K-[PS]-[A]; (ii) R-x-F-G-E-N-Y-[VL]-Q-E-x(2)-[ED]-K; (iii) [IL]-x-W-H-F-I-G-x(2)-Q-x(1–4)-N-K; (iv) L-x-V-x-[VIL]-Q-[IV]-N-x-S-x-E-x(2)-K-x-G; (v) G-[LF]-M-[TA]-I; and (vi) L-[SN]-[ML]-G-M-S-x-D-x(3)-A-[IV]-x(2)-G-x-[TA]-x-V-R-[IV]-G-[ST]-[IL]-F-G-[AE]-A-R-[DE]-Y. The third of these sequences has been recognized as the signature of an uncharacterized protein family, UPF0001 (consensus pattern: [FW]-H-[FM]-[IV]-G-x-[LIV]-Q-x-[NKR]-K-x(3)-[LIV]; Prosite: http://www.genome.ad.jp/dbget-bin/show_man?prosite). The other five elements revealed no similarities to known sequence motifs in the public databases. Thus, the *PROSC* gene has been highly conserved throughout evolution, and therefore its product is likely to play a vital role in cellular function. De Wergifosse et al. (1994) speculated that bacterial *PROSC* may be involved in proline synthesis because it is located upstream from and may be cotranscribed with proC, a known proline biosynthetic gene (Savioz et al. 1990). Its role in mammals remains to be determined.

Table 1. Amino acid sequence homology of *PROSC* genes to their human counterpart

Species	Homology[a]
Mouse	86.5
A. thaliana	50.2
C. elegans	40.7
Yeast	40.3
E. coli	35.9
P. aeruginosa	28.0

[a] % identity for the entire sequences

Expression of human *PROSC* in various tissues

Northern blot analysis detected a single, ubiquitously expressed human transcript about 2.6 kb long (Fig. 3). Multiple "hits" of human *PROSC* against the EST database also indicated ubiquitous and abundant expression of this gene.

Table 2. Exon–intron boundaries of the human *PROSC* gene

Exon					Intron	
Number	Length (bp)	cDNA position	Splice acceptor	Splice donor	Number	Length (bp)
1	135	1–135		GCGGCCGCGGgtgaggaagg	1	2769
2	108	136–243	ctcttggcagGATCTCCCAG	CGAGAACTACgtaagagccc	2	77
3	36	244–279	gacctttagGTTCAGGAAC	AAATCCCAAAgtaagtagat	3	533
4	76	280–355	ttcccctcagATTCTGTCTT	AAATTGATGGgtaagataaa	4	6418
5	138	356–493	ctcattacagCTGTCCCCAA	GGAGAAGAGAgtaagtaacc	5	2454
6	143	494–636	tttcttgaagGTAAACATGG	AGACTTCCAGgtactggggg	6	436
7	99	637–735	tttctgtcagCTGTTATTGT	CCAGCATGCGgtgagtgtcc	7	1960
8	1799	736–2530	tttcccacagGTTGAAGTAG	(ATTTTTTATGatctggtgta)		

The sequence at the 3′-end of the gene is in parentheses

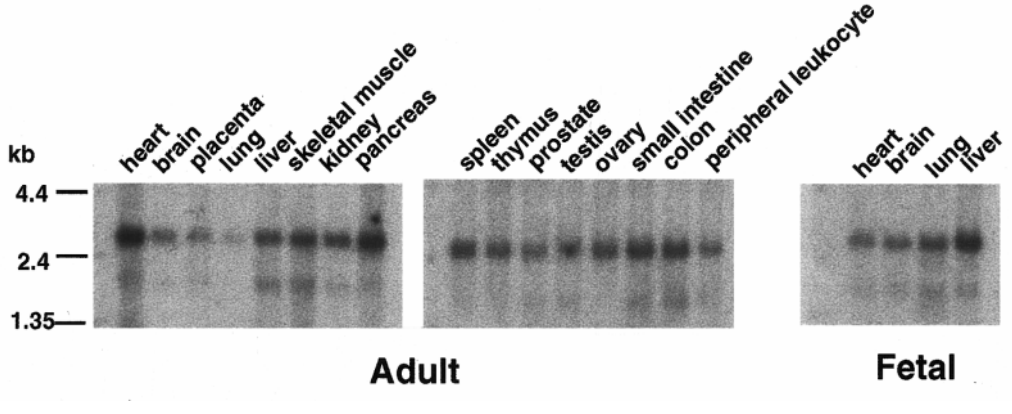

Fig. 3. Expression of the human *PROSC* gene in adult and fetal tissues, showing ubiquitous expression of a single 2.6-kb transcript. Molecular sizes (kb) are indicated at *left*

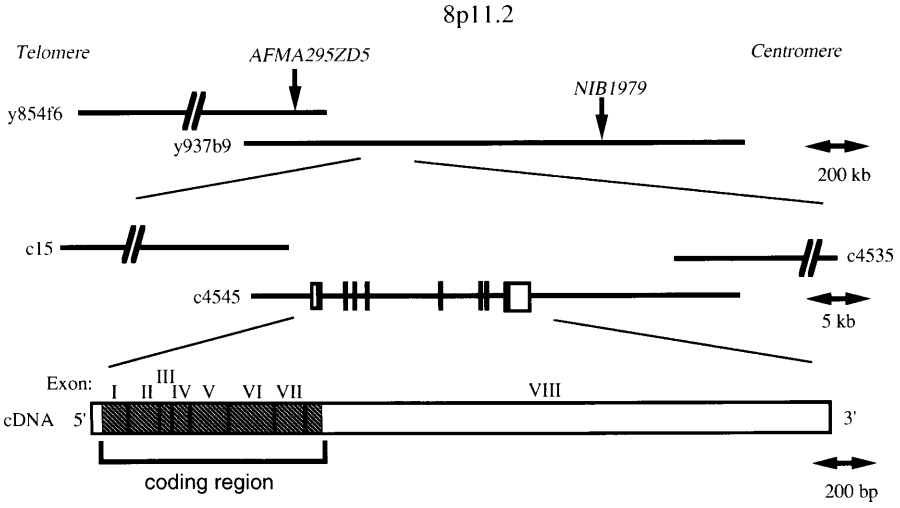

Fig. 4. Local mapping and genomic structure of the human *PROSC* gene. y937b9 and y854f6 indicate CEPH YACs, 937_b_9 and 854_f_6, respectively. *Shaded boxes* indicate coding regions; *open boxes* denote untranslated regions

Chromosomal location and character of the human *PROSC* gene

Comparison of cDNA and genomic sequences revealed that the entire human *PROSC* gene was contained in cosmid clone c4545, one of the sub-clones derived from a YAC, 937_b_9 on 8p11.2. The gene was situated between STS markers *NIB1979* (proximal) and *AFMA295ZD5* (distal), and oriented toward the centromere (Fig. 4). The gene spanned 17.2 kb of 8p11.2 and consisted of eight exons (Table 2); all sequences at exon–intron junctions were consistent with the AG-GT rule. In *C. elegans* (YU68_CAEEL) and *A. thaliana* (F12F1_20), the *PROSC* genes are composed of only seven exons, whose exon–intron junctions were not in good alignment with the human genomic structure.

In summary, we have isolated a novel human gene, *PROSC*, through large-scale sequencing of a genomic region on 8p11.2 coupled with analysis by gene-trapping software. We also identified its mouse counterpart. This gene is ubiquitously expressed in human tissues, and has been highly conserved throughout evolution. The *PROSC* product is likely to be a soluble cytoplasmic protein whose function remains to be determined.

Acknowledgments This work was supported in part by a "Research for the Future" Program Grant (96L00102) of The Japan Society for the Promotion of Science and by Japan Science and Technology Corporation (JST).

References

Claverie JM (1997) Computational methods for the identification of genes in vertebrate genomic sequences. Hum Mol Genet 6:1735–1744

Daigo Y, Isomura M, Nishiwaki T, Tamari M, Ishikawa S, Kai M, Murata Y, Takeuchi K, Yamane Y, Hayashi R, Minami M, Fujino MA, Hojo Y, Uchiyama I, Takagi T, Nakamura Y (1999) Characterization of a 1,200-kb genomic segment of chromosome 3p22-p21.3. DNA Res (6:37-44)

De Wergifosse P, Jacques B, Jonniaux JL, Purnelle B, Skala J, Goffeau A (1994) The sequence of a 22.4 kb DNA fragment from the left arm of yeast chromosome II reveals homologues to bacterial proline synthetase and mouse alpha-adaptin, as well as a new permease and a DNA-binding protein. Yeast 10:1489–1496

Elkahloun AG, Krizman DB, Wang Z, Hofmann TA, Roe B, Meltzer PS (1997) Transcript mapping in a 46-kb sequenced region at the core of 12q13.3 amplification in human cancers. Genomics 42:295–301

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8: 175–185

Ikegawa S, Isomura M, Koshizuka Y, Nakamura Y (1999a) Cloning and characterization of *ASH2L and Ash2I*, human and mouse homologs of the *Drosophila ash2* gene. Cytogenet Cell Genet 84:167-172

Ikegawa S, Isomura M, Koshizuka Y, Nakamura Y (1999b) Cloning and characterization of a novel gene (*c8orf2*), a human representative of a novel gene family with homology to *C. elegans* C42.C1.9. Cytogenet Cell Genet (in press)

Ishikawa S, Kai M, Murata Y, Tamari M, Daigo Y, Murano T, Ogawa M, Nakamura Y (1998) Genomic organization and mapping of the human activin receptor type IIB (hActR-IIB) gene. J Hum Genet 43:132–134

Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283–292

McKusick VA (1997) Genomics: structural and functional studies of genomes. Genomics 45:244–249

Murata Y, Tamari M, Takahashi T, Horio Y, Hibi K, Yokoyama S, Inazawa J, Yamakawa K, Ogawa A, Takahashi T, Nakamura Y (1994) Characterization of an 800-kb region at 3p22-p21.3 that was homozygously deleted in a lung cancer cell line. Hum Mol Genet 3:1341–1344

Rowen L, Mahairas G, Hood L (1997) Sequencing the human genome. Science 278:605–607

Savioz A, Jeenes DJ, Kocher HP, Haas D (1990) Comparison of proC and other housekeeping genes of *Pseudomonas aeruginosa* with their counterparts in *Escherichia coli*. Gene 86:107–111

Solovyev VV, Salamov AA, Lawrence CB (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res 22:5156–5163

Uberbacher EC, Mural RJ (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci USA 88:11261–11265

Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, Connell M, Copsey T, Cooper J, Coulson A, Craxton M, Dear S, Du Z, Durbin R, Favello A, Fulton L, Gardner A, Green P, Hawkins T, Hillier L, Jier M, Johnston L, Jones M, Kershaw J, Kirsten J, Laister N, Latreille P, Lightning J, Lloyd C, McMurray A, Mortimore B, O'Callaghan M, Parsons J, Percy C, Rifken L, Roopra A, Saunders D, Shownkeen R, Smaldon N, Smith A, Sonnhammer E, Staden R, Sulston J, Thierry-Mieg J, Thomas K, Vaudin M, Vaughan K, Waterston R, Watson A, Weinstock L, Wilkinson-Sproat J, Wohldman P (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. Nature (Lond) 368:32–38