Available online at www.sciencedirect.com
SCIENCE DIRECT®
PEPTIDES
Peptides 26 (2005) 1278–1282
www.elsevier.com/locate/peptides

Short communication

# Human sorbin is generated via splicing of an alternative transcript from the ArgBP2 gene locus

Damon Hand, Lee E. Eiden *

*National Institute of Health/National Institute of Mental Health, Section on Molecular Neuroscience, 36 Convent Drive, MSC 4090, 9000 Rockville Pike Bethesda, MD 20892-4090, USA*

Received 4 December 2004; received in revised form 22 January 2005; accepted 24 January 2005
Available online 7 March 2005

**Abstract**

We demonstrate that the human sorbin polypeptide is generated via splicing of an alternative transcript from the ArgBP2 gene locus. Previous studies have demonstrated that the central 139 amino acid region of the porcine sorbin polypeptide exhibits 95% homology to part of the human ArgBP2 protein. Yet neither the sorbin N- nor C-terminus has been identified in ArgBP2 or any other protein to date. Using computational analysis, we locate the sorbin N- and C-termini in the human ArgBP2 gene locus, and demonstrate that they are spliced to the 5′ and 3′ ends of the 95% homologous region. In addition, several sequence anomalies were identified in the putative human sorbin cDNA (AF396457). Thus, a revised human sorbin nucleotide sequence is proposed.
Published by Elsevier Inc.

## 1. Introduction

Sorbin was first identified as a 153 amino acid C-terminally amidated polypeptide (17.5 kDa) from the porcine upper intestine [11]. This protein has been found in a variety of tissues including porcine stomach, pancreas, and the enteric nervous system [4]. The biological activity of sorbin is mimicked by application of the amidated C-terminal heptapeptide (PVTKPQA-$NH_2$) alone, demonstrating increased water and electrolyte absorption in the intestine [3,11], as well as decreased VIP- and cholera toxin-induced secretion [8,9].

The porcine sorbin polypeptide exhibits 86% homology to a portion of the human ArgBP2 adaptor protein (translated mRNA sequence NM_021069, version NM_021069.2), suggesting that it might be generated via proteolytic cleavage or splicing of an alterative transcript from the ArgBP2 gene locus [6]. The homology between sorbin and ArgBP2 is particularly evident in the central 139 amino acid residues of sorbin, which exhibits 95% sequence identity to the translated ArgBP2 mRNA sequence, and is referred to here as the "sorbin/ArgBP2 core" (Fig. 1). However, neither the N- nor the C-terminal peptide sequences of sorbin have been identified in an extant human ArgBP2 protein sequence, or any ArgBP2 mRNA/cDNA deduced open reading frame (ORF) [13].

In this study, we have conducted data mining of nucleic acid sequence databases to examine the ArgBP2 gene locus within chromosomal region 4q35.1 and identified genomic sequences encoding the sorbin N- and C-termini, indicating that human sorbin is spliced from an alternative transcript from the ArgBP2 gene locus. In addition, several anomalies were identified in the putative human sorbin cDNA sequence (AF396457, version AF396457.1), a nucleotide sequence deduced from a PCR amplicon sequence [13]. Subsequently, we propose a revised sequence for human sorbin and confirm its' existence from a previously sequenced, but unidentified, expressed sequence tag.

## 2. Materials and methods

All similarity searches were conducted using the Internet-based Basic Local Alignment Search Tool (BLAST) [1] available through the National Center for Biotechnology In-

* Corresponding author. Tel.: +1 301 496 4110; fax: +1 301 402 1748.
*E-mail address:* eidenl@mail.nih.gov (L.E. Eiden).

0196-9781/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.peptides.2005.01.016

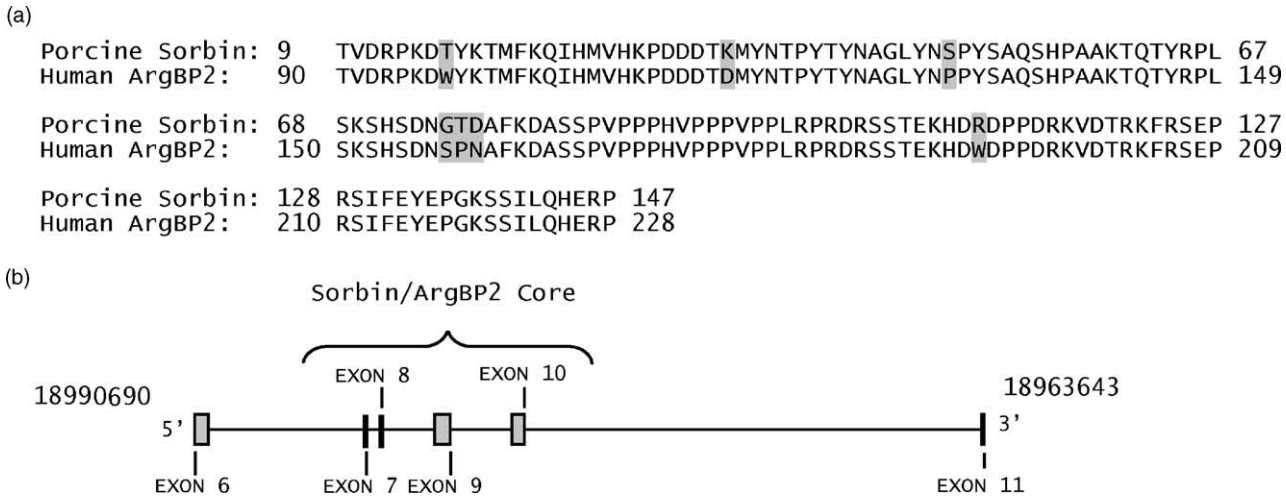


Fig. 1. Porcine sorbin polypeptide homology with the human ArgBP2 adaptor protein: (a) depicts a region exhibiting 95% homology between the porcine sorbin polypeptide sequence and the human ArgBP2 adaptor protein (translated from NM_021069 mRNA), which is referred to as the sorbin/ArgBP2 core; (b) displays the exons of the sorbin/ArgBP2 core within the ArgBP2 gene locus on the contiguous genomic sequence NT_022792 (version NT_022792.17).

formation (NCBI) [15]. Utilized versions of the BLAST program include the nucleotide–nucleotide BLAST search (blastn) and the protein query versus translated database search (tblastn). Unless otherwise noted, nucleotide searches were conducted with a statistical significance threshold (*E*-value) of 10 and a minimum word size of seven characters. Protein query versus translated database searches were completed with an *E*-value of 100000, a minimum word size of three characters, a PAM30 substitution matrix, and a gap insertion and extension penalty costs of nine and one, respectively.

The databases queried with the BLAST program consist of the following: (1) the "nr" database which contains sequence submissions from GenBank, the European Molecular Biology Laboratory, and the DNA Data Bank of Japan [14]; (2) the human chromosome sequence database [5]; (3) the EST database available through GenBank [14]; and (4) human whole-genome-shotgun (WGS) sequences [12].

## 3. Results

A BLAST (tblastn) search was conducted using the porcine sorbin N-terminal query sequence (MRAATPLQ) against the human chromosome sequence database. There were a total of 148 hits across the human genome with an *E*-value of 1000 or less. Six of these hits were on chromosome four with one specifically in the ArgBP2 gene locus. The hit detected in the ArgBP2 gene locus (MKATTPLQ) demonstrated 75% amino acid homology to the porcine sorbin N-terminus sequence. The nucleotide sequence corresponding with the hit (24mer) was located 3888 nucleotides upstream from the sorbin/ArgBP2 core and includes 5′-GU-AG-3′ splicing motifs flanking the ends of the 3888 base region (Fig. 2a). Moreover, the 24mer exhibited 100% identity to the N-terminus of the putative human sorbin cDNA sequence AF396457 (Fig. 2b).

Another BLAST (tblastn) search was conducted using the porcine sorbin C-terminal query sequence (VTKPQA)

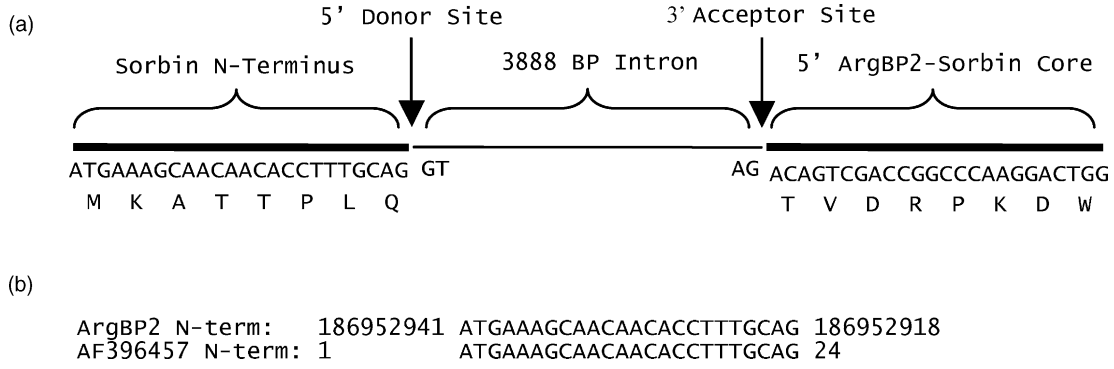


Fig. 2. Human sorbin N-terminal sequence in the ArgBP2 gene locus: (a) illustrates the arrangement of the ArgBP2 gene locus that corresponds to the spliced sections of the sorbin N-terminus and 5′ end of the sorbin/ArgBP2 core; (b) pairwise alignment of the N-terminal coding sequence detected in the ArgBP2 gene locus (from genomic sequence NC_000004, version NC_000004.9) to the putative human sorbin N-terminal coding sequence (AF396457).

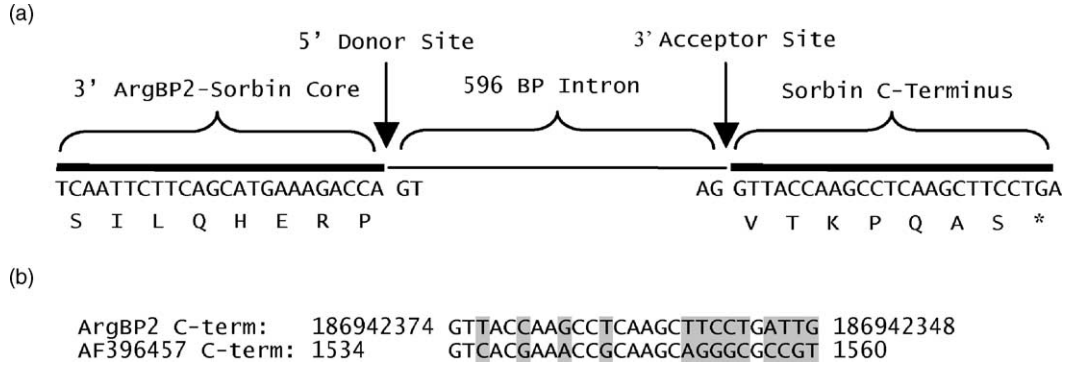


Fig. 3. Human sorbin C-terminal coding sequence in the ArgBP2 gene locus: (a) illustrates the arrangement of the ArgBP2 gene locus that corresponds to the spliced sections of the sorbin C-terminus and 3′ end of the sorbin/ArgBP2 core; (b) pairwise alignment of the C-terminal coding sequence detected in the ArgBP2 gene locus (from genomic sequence NC_000004, version NC_000004.9) to the putative human sorbin C-terminal coding sequence (AF396457).

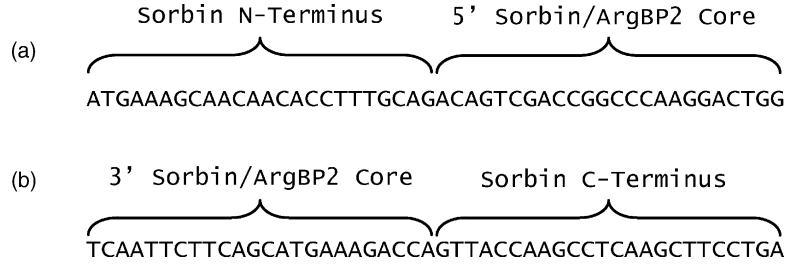


Fig. 4. BLAST query sequence used to probe the human EST database for sorbin-like transcripts: (a) N-terminus nucleotide sequence combined with the 5′ end of the sorbin/ArgBP2 core; (b) C-terminus nucleotide sequence combined with the 3′ end of the sorbin/ArgBP2 core.

against the human chromosome sequence database. This search resulted in 1513 hits overall including 84 hits on chromosome four and one hit specifically in the ArgBP2 gene locus. This hit within the ArgBP2 gene locus demonstrated 100% homology to the query sequence. The nucleotide sequence corresponding with the hit (18mer) was located 596 bases downstream from the 3′ end of the sorbin/ArgBP2 core; and this 596 base region contained 5′-GU-AG-3′ splicing motifs flanking its' ends (Fig. 3a). However, comparison between the 18mer detected in the ArgBP2 gene locus and the AF396457 C-terminal nucleotide sequence demonstrated only 52% sequence similarity (Fig. 3b). In addition, AF396457 codes for a GRR sequence following the C-terminal heptapeptide, whereas the 18mer is followed by a serine codon and a stop codon, respectively (TCCTGA). An attempt was made to identify the gene locus of the AF396457 C-terminal nucleotide sequence in order to reconcile these differences. A series of BLAST searches (blastn) were conducted using varying lengths of the AF396457 nucleotide sequence with an *E*-value threshold of 1000, but no hits were detected with over 78% similarity for any of the input query sequences, other than the AF396457 sequence itself.

In order to demonstrate that both the identified N-terminal 24mer and C-terminal 18mer are actually spliced to the sorbin/ArgBP2 core, we conducted a follow-up BLAST (tblastn) search using a continuous query sequence of the N-terminal coding 24mer and the 5′ end of the sorbin/ArgBP2 core (Fig. 4a), as well as the C-terminus coding 18mer in conjunction with the 3′ end of the sorbin/ArgBP2 core (Fig. 4b). The result was detection of 12 hits with over 97% sequence identity to the N-terminal 24mer and sorbin/ArgBP2 core combination, and six separate hits with over 95% similarity to the C-terminal 18mer and sorbin/ArgBP2 sequence combination. Of these total 18 hits, one EST (BF675132, version BF675132.1) isolated from prostate tissue as part of the mammalian gene collection contained the N-terminal 24mer, the sorbin/ArgBP2 core, and the majority of the C-terminal 18mer. Insertion of two gaps into this sequence to correct for likely sequencing errors provided a revised human sorbin sequence with an ORF that codes for a human polypeptide with 93% homology to the porcine sorbin polypeptide (Fig. 5a and b).

## 4. Discussion

We conclude that human sorbin is spliced from an alternative transcript from the ArgBP2 gene locus. The sorbin N- and C-terminal coding regions were identified upstream and downstream of the sorbin/ArgBP2 core, respectively. In addition, we identified an EST (BF675132) that contains the entire sorbin open reading frame including the N-terminal 24mer, the sorbin/ArgBP2 core, and most of the C-terminal 18mer. This demonstrates that the N- and C-terminal sequences are actually spliced to the sorbin/ArgBP2 core in vivo.

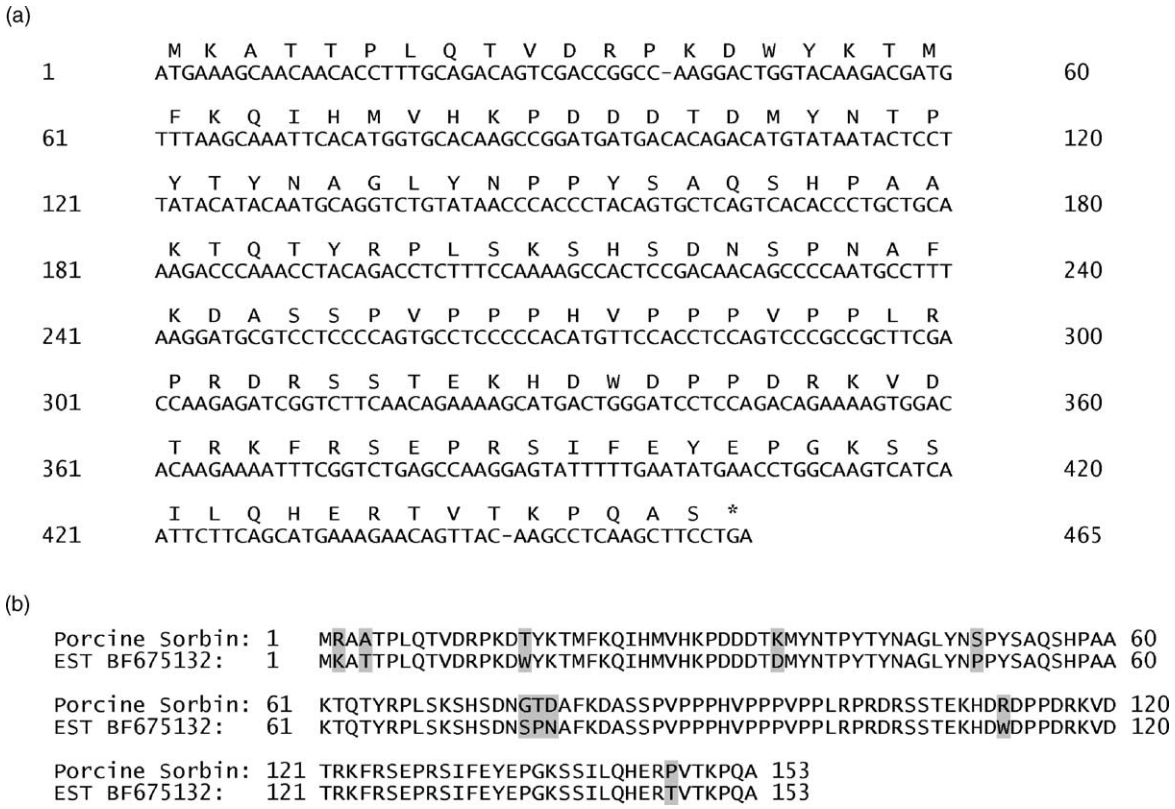


Fig. 5. Previously unidentified human sorbin transcript: (a) the EST BF675132 nucleotide sequence and the corresponding translated amino acid sequence which contains the human sorbin N-terminus, sorbin/ArgBP2 core, and the majority of the C-terminus; (b) the amino acid sequence from the porcine sorbin polypeptide and the translated EST BF675132 exhibit 93% homology.

There were several differences between the sorbin C-terminal 18mer identified in the ArgBP2 gene locus and that of the AF396457 C-terminal sorbin coding sequence deposited in GenBank by Wahbi et al. It is likely that these discrepancies are due to sequence errors in the AF396457 GenBank submission, since no sequences with significant similarity were detected in the human genome, and evaluation of the translated putative human sorbin cDNA sequence published by Wahbi et al. (2001) demonstrated numerous dissimilarities to the corresponding translated AF396457 sequence. This includes several random sequence fragments in the AF396457 submission that were not published in the Wahbi et al. predicted human sorbin peptide sequence, which subsequently caused a shift in the AF396457 ORF of the sorbin C-terminal region.

Location of the human sorbin C-terminus was an especially significant finding of this study, since it is critical for sorbin's biological activity. We found that the C-terminal coding 18mer from the ArgBP2 gene locus is succeeded by a serine codon and a stop codon, respectively, rather than by a glycine residue, the essential amino acid required by peptidylglycine α-hydroxylating monooxygenase (PHM) and peptidyl-α-hydroxyglycine α-amidating lyase (PAL) to complete a peptidyl amidation reaction [2,7,10]. Thus, it seems likely that the sorbin C-terminus in humans is not amidated. It may naturally exist in the serine-extended form in humans. Alternatively, a novel enzyme(s) may exist that utilizes a serine residue in an amidation reaction, although there is currently no precedent for this possibility.

This study emphasizes the importance of utilizing all available genomic and expressed sequences in the search for novel bioactive peptides and their precursors. Clone sequence validation is essential and results in more accurate comparisons of cross-species sequence conservation and improved projections of biological function.

## References

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[2] Bradbury AF, Smyth DG. Biosynthesis of the C-terminal amide in peptide hormones. Biosci Rep 1987;7:907–16.
[3] Charpin G, et al. Effect of sorbin on duodenal absorption of water and electrolytes in the rat. Gastroenterology 1992;103:1568–73.
[4] Fadil FA, Nicol P, Leduque P, Berger F, Descroix-Vagne M, Pansu D. Sorbin in the porcine gastrointestinal tract and pancreas: an immunocytochemical analysis. Endocrinology 1997;138:4989–99.
[5] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[6] Kioka N, Ueda K, Amachi T. Vinexin CAP/ponsin, ArgBP2: a novel adaptor protein family regulating organization and signal transduction. Cell Struct Funct 2002;27:1–7.

[7] Kreil G. Occurrence, detection, and biosynthesis of carboxy-terminal amides. Meth Enzymol 1984;106:218–23.
[8] Marquet F, Botella A, Bueno L, Pansu D, Descroix-Vagne M. Effect of sorbin derivatives on cholera toxin-induced intestinal secretion in rat in vivo. Peptides 1998;19:1417–23.
[9] Marquet F, Grishina O, Pansu D, Descroix-Vagne M. Effect of C-terminal derivatives of sorbin on ileal ion transport stimulated by VIP in rats. Gastroenterol Clin Biol 1994;18:702–7.
[10] Martinez A, Treston AM. At the cutting edge: where does amidation take place? Mol Cell Endocrinol 1996;123:113–7.
[11] Vagne-Descroix, et al. Isolation and characterization of porcine sorbin. Eur J Biochem 1991;201:53–9.
[12] Venter, et al. The sequence of the human genome. Science 2001;29:11304–51.
[13] Wahbi K, Magaud JP, Pansu D, Descroix-Vagne M. Coding region of the sorbin gene in different species. Peptides 2001;22:2045–53.
[14] Wheeler, et al. Database resources of the National Center for Biotechnology Information: update. Nucl Acids Res 2004;32.
[15] http://www.ncbi.nih.gov/BLAST/.