ISSN 0006-2979, Biochemistry (Moscow), 2016, Vol. 81, No. 2, pp. 152-162. © Pleiades Publishing, Ltd., 2016.
Published in Russian in Biokhimiya, 2016, Vol. 81, No. 2, pp. 251-264.

Express Sequence Tag Analysis – Identification of Anseriformes Trypsin Genes from Full-Length cDNA Library of the Duck (*Anas platyrhynchos*) and Characterization of Their Structure and Function

Haining Yu[1]*, Shasha Cai[1#], Jiuxiang Gao[1#], Chen Wang[1], Xue Qiao[1], Hui Wang[1], Lan Feng[1], and Yipeng Wang[2]*

[1]*Dalian University of Technology, Institute of Marine Biological Technology, School of Life Science and Biotechnology, 116024 Dalian, Liaoning, China; E-mail: joannyu@live.cn*
[2]*Soochow University, College of Pharmaceutical Sciences, 215123 Suzhou, Jiangsu, China;*
fax: + 86 (411) 847-08850; E-mail: yipengwang@suda.edu.cn

Received March 4, 2015
Revision received May 25, 2015

Abstract—Trypsins are key proteins important in animal protein digestion by breaking down the peptide bonds on the carboxyl side of lysine and arginine residues, hence it has been used widely in various biotechnological processes. In the current study, a full-length cDNA library with capacity of $5\cdot10^5$ CFU/ml from the duck (*Anas platyrhynchos*) was constructed. Using express sequence tag (EST) sequencing, genes coding two trypsins were identified and two full-length trypsin cDNAs were then obtained by rapid-amplification of cDNA end (RACE)-PCR. Using Blast, they were classified into the trypsin I and II subfamilies, but both encoded a signal peptide, an activation peptide, and a 223-a.a. mature protein located in the C-terminus. The two deduced mature proteins were designated as trypsin-IAP and trypsin-IIAP, and their theoretical isoelectric points (p*I*) and molecular weights (MW) were 7.99/23466.4 Da and 4.65/24066.0 Da, respectively. Molecular characterizations of genes were further performed by detailed bioinformatics analysis. Phylogenetic analysis revealed that trypsin-IIAP has an evolution pattern distinct from trypsin-IAP, suggesting its evolutionary advantage. Then the duck trypsin-IIAP was expressed in an *Escherichia coli* system, and its kinetic parameters were measured. The three dimensional structures of trypsin-IAP and trypsin-IIAP were predicted by homology modeling, and the conserved residues required for functionality were identified. Two loops controlling the specificity of the trypsin and the substrate-binding pocket represented in the model are almost identical in primary sequences and backbone tertiary structures of the trypsin families.

DOI: 10.1134/S0006297916020097

Key words: Anseriformes, *Anas platyrhynchos*, trypsin, molecular cloning, expression, structure and function

Trypsin is a serine protease found in the digestive system of many vertebrates, where it hydrolyzes proteins [1]. It is produced by pancreatic acinar cells as a trypsinogen, released into the intestine, and converted into active trypsin through the action of enterokinase or by autoactivation [2]. Trypsin is widely distributed in vertebrates, and many trypsinogen genes and proteins have been isolated and characterized, including those of chicken [3], bovine [4], pig [5], ostrich [2], human [6-8], dogfish [9], mouse [10], rat [11-14], etc. Belonging to the large serine protease family, trypsin is used for numerous biotechnological processes, and the prototypes for investigating many aspects of enzyme action [3, 15]. It is a digestive enzyme having strong specificity, selectively catalyzing the hydrolysis of peptide bonds on the carboxyl side of lysine and arginine residues of proteins in the intestinal lumen, except when either is followed by proline. This specificity comes from the matching of the negative charge of an Asp residue in the primary substrate-binding pocket S1 of trypsin and the negative charge of the P1 side in the substrate peptide chain [16].

Abbreviations: cDNA, complementary DNA; EST, express sequence tags; FPLC, fast protein liquid chromatography; ORF, open reading frame; PCR, polymerase chain reaction; RACE, rapid-amplification of cDNA ends; 5′-UTR, 5′-untranslated region.

[#] These authors contributed equally to this work.

* To whom correspondence should be addressed.

Trypsins can be divided into two major subfamilies: the cationic trypsin I subfamily and the anionic trypsin II subfamily [3]. Recently, another type of trypsin, named mesotrypsin, was found from the human pancreas. It is described as a minor trypsin isoform with the remarkable property of near total resistance to biological trypsin inhibitors [17]. Previous studies have shown that these trypsin isoenzymes are encoded by different genes in the genome and have different tissue expression patterns, indicating their functional differences [3, 18-21]. Thus, it is interesting to explore structure–activity relationships in trypsins from new species of evolutionary significance [22]. Although trypsins have been studied in many animal species, there are no reports concerning those from the Anatidae avian family. In the present study, we identified two full-length trypsin cDNA sequences from the constructed cDNA library of *Anas platyrhynchos* based on EST analysis. The deduced amino acid sequences of the complete Anatidae trypsin precursors were aligned with other representative vertebrate trypsins, and the mature proteins were named trypsin-IAP and trypsin-IIAP, respectively. Trypsin-IIAP was expressed in *E. coli* to study its biochemical properties. Furthermore, the three dimensional structure of trypsin-IAP and trypsin-IIAP were predicted by homology modeling to elucidate the possible functional mechanisms.

MATERIALS AND METHODS

Collection of tissues. One adult female *A. platyrhynchos* (weight 1.5 kg) was collected from local market of Dalian. It was sacrificed and tissues were rapidly dissected, frozen in liquid nitrogen, and stored at –80°C until use. The collection procedure was approved by a recognized Animal Ethics Committee of Dalian University of Technology.

Total RNA extraction and purification. The *Anas platyrhynchos* pancreas (0.1 g) was cut into pieces with a small sterile scissor, frozen immediately in liquid nitrogen, and then ground into powder. Total RNA was extracted using the RNeasy AxyPrep^{TM} Multisource Total RNA Miniprep Kit (Qiagen, USA). All RNA isolation procedures were carried out based on the manufacturer's instructions. The total RNA was purified using PolyATtract® mRNA Isolation Systems (Promega, USA) in accordance with the user's manual.

Full-length cDNA library construction. The full-length cDNA sequence was synthesized using an In-Fusion® SMARTer™ Directional cDNA Library Construction Kit (Clontech, USA) following the manufacturer's protocol. In brief, the first-strand cDNA was synthesized using PowerScript reverse transcriptase with the SMARTer V oligonucleotide primer 5′-AAGCAGTGGTATCAACGCAGAGTACXXXXX-3′ and 3′ In-Fusion SMARTer CDS Primer 5′-CGGGGTACGATGAGACACCATTTTTTTTTTTTTTTTTTTTVN-3′(N = A, C, G, or T; V = A, G, or C). The second strand was amplified using Advantage DNA polymerase (Clontech) with the 5′ PCR primer: 5′-AAGCAGTGGTATCAACGCAGAGT-3′ and the In-Fusion SMARTer CDS/3′ PCR primer. The double-strand cDNAs (ds-cDNAs) were digested with Sfi I restriction enzyme and recovered for longer than 600 bp. The cDNA fragments were ligated into pDNR-LIB vector digested by Sfi I, and ultimately transferred into the competent cells of *E. coli* DH10B. The cDNA library was calculated of clone numbers on plates by titration.

Identification of expressed sequence tag (EST) encoding trypsins. Individual cDNA clones were selected randomly from the *A. platyrhynchos* cDNA library and sequenced. Raw sequences were first trimmed to remove vector sequence and low-quality sequences using the Crossmatch program. ESTs with length less than 100 bp were discarded. The sequences of cDNA clones were compared with sequences in the GenBank database (http://www.ncbi.nlm.nih.gov/blast).

Isolation of full-length cDNAs encoding novel trypsins. The synthesized cDNA library was used as a template to screen the cDNAs encoding trypsin homologs of *A. platyrhynchos*. The obtained EST sequence coupled with the conserved domain from multiple alignments of trypsins was exploited to design the specific primers for cloning the *N*-terminal part of full-length duck trypsin precursor. Thus, the anti-sense primer (5′-AACCTCCACAGAAGTGATACCC-3′) and the 5′-primer supplied by the kit (5′-AAGCAGTGGTATCAACGCAGAGT-3′) were used to amplify the target sequence. The DNA polymerase used was universal rTaq polymerase (Takara, Japan). The reaction conditions were: initial denaturation for 5 min at 94°C, followed by 30 cycles of denaturation at 94°C for 30 s, annealing at 53.9°C for 30 s, extension at 72°C for 30 s, and a final extension at 72°C for 10 min. Finally, the PCR products were verified by 1% agarose gel electrophoresis, purified using Agarose Gel DNA Purification Kit (TIANGEN, China), and cloned into pMD^{TM}19-T vector (Takara). DNA sequencing was performed using an ABI PRISM 377 instrument (Applied Biosystems, USA).

Then, according to the signal domain characterized above, one sense primer trypsin-P2 (5′-CCATGCATTCTCTCTTCCTCCT-3′) was designed and used to amplify the full-length duck trypsin cDNAs, coupled with In-Fusion SMARTer CDS III/3′ PCR primer as described above. The PCR conditions were: 94°C for 4 min, 28 cycles of 94°C for 30 s, 55.9°C for 30 s, 72°C for 1 min, and again followed by a final extension at 72°C for 10 min. The purification of PCR product and the DNA sequencing was performed as described above.

Multiple sequence alignment and phylogenetic analysis. Multiple sequence alignments were performed using the

```
a  atgaaatacatactgtttgtcgcctttgttggtgtggctgttgccttccccatcagtgat 60
   tactttatgtatgacaaacagcggaaacaaccacaccgacaacggaaggggtagtcacta
    M  K  Y  I  L  F  V  A  F  V  G  V  A  V  A  F  P  I  S  D  20
   gatgatgatgacaagattgtgggaggctacacctgtacaaagaactctattccatatcag 120
   ctactactactgttctaacaccctccgatgtggacatgtttcttgagataaggtatagtc
    D  D  D  D  K  I  V  G  G  Y  T  C  T  K  N  S  I  P  Y  Q  40
   gtgtccctgaattctgggtatcacttctgtggaggttctctcatcagcagccagtgggtc 180
   cacagggacttaagacccatagtgaagacacctccaagagagtagtcgtcggtcacccag
    V  S  L  N  S  G  Y  H  F  C  G  G  S  L  I  S  S  Q  W  V  60
   ctgtcagcggctcactgctacaagtcttccatccaagtgaaacttggggaacacaacctg 240
   gacagtcgccgagtgacgatgttcagaaggtaggttcactttgaaccccttgtgttggac
    L  S  A  A  H  C  Y  K  S  S  I  Q  V  K  L  G  E  H  N  L  80
   gcatcccaagaaggcactgagcagaccataagttcatctaaagtcatccgccactctggc 300
   cgtagggttcttccgtgactcgtctggtattcaagtagatttcagtaggcggtgagaccg
    A  S  Q  E  G  T  E  Q  T  I  S  S  S  K  V  I  R  H  S  G  100
   tacaacgcctacacgctggacaacgacattatgctcatcaaactttccaaagcagccaca 360
   atgttgcggatgtgcgacctgttgctgtaatacgagtagtttgaaaggtttcgtcggtgt
    Y  N  A  Y  T  L  D  N  D  I  M  L  I  K  L  S  K  A  A  T  120
   ctcaactcctatgtcaacacagttcctcttcctaccagctgtgtggccactggcaccaca 420
   gagttgaggatacagttgtgtcaaggagaaggatggtcgacacaccggtgaccgtggtgt
    L  N  S  Y  V  N  T  V  P  L  P  T  S  C  V  A  T  G  T  T  140
   tgcctaatctctggatggggcaacacactcagcagtggcagtctgtatccagataacttg 480
   acggattagagacctaccccgttgtgtgagtcgtcaccgtcagacataggtctattgaac
    C  L  I  S  G  W  G  N  T  L  S  S  G  S  L  Y  P  D  N  L  160
   cagtgcctgagagctcctgtactctcctcaagccagtgcagcagtgcctaccccggccaa 540
   gtcacggactctcgaggacatgagaggagttcggtcacgtcgtcacggatggggccggtt
    Q  C  L  R  A  P  V  L  S  S  S  Q  C  S  S  A  Y  P  G  Q  180
   attactagcaacatgatatgtgtaggatacctggaaggagggaaagactcctgccagggg 600
   taatgatcgttgtactatacacatcctatggaccttcctccctttctgaggacggacccc
    I  T  S  N  M  I  C  V  G  Y  L  E  G  G  K  D  S  C  Q  G  200
   gattctggtggtccagtagtctgcaatgggcaactccaaggtattgtttcctggggcatt 660
   ctaagaccaccaggtcatcagacgttacccgttgaggttccataacaaaggaccccgtaa
    D  S  G  G  P  V  V  C  N  G  Q  L  Q  G  I  V  S  W  G  I  220
   ggatgtgcacagaaaggctatcctggagtttacactaaggtttgcaattatgtctcctgg 720
   cctacacgtgtctttccgataggacctcaaatgtgattccaaacgttaatacagaggacc
    G  C  A  Q  K  G  Y  P  G  V  Y  T  K  V  C  N  Y  V  S  W  240
   atccaatcaactattgctgccaactga                                  747
   taggttagttgataacgacggttgact
    I  Q  S  T  I  A  A  N  *                                   248
```

Fig. 1. cDNA sequences encoding duck (*A. platyrhynchos*) trypsin-IAP (a) and trypsin-IIAP (b) and the deduced precursor sequences. The putative signal peptides are in bold and italic. The amino acid sequences of predicted activation peptides are in underlined. The stop codon is indicated by an asterisk (*); the 3′-untranslated region is in lowercase letters.

b

```
atgcattctctcttcctccttctctcctgcctgggagccgctgttgctttccctagagct 60
tacgtaagagagaaggaggaagagaggacggaccctcggcgacaacgaaagggatctcga
 M  H  S  L  F  L  L  L  S  C  L  G  A  A  V  A  F  P  R  A  20
gctgatgatgacaagattgtgggaggctacaactgcccagagcattcagttccctaccag 120
cgactactactgttctaacaccctccgatgttgacgggtctcgtaagtcaagggatggac
 A  D  D  D  K  I  V  G  G  Y  N  C  P  E  H  S  V  P  Y  Q  40
gtgtccctgaatgctggctatcacttttgtggaggatccctcatcaacaaccagtgggtc 180
cacagggacttacgaccgatagtgaaaacacctcctagggagtagttgttggtcacccag
 V  S  L  N  A  G  Y  H  F  C  G  G  S  L  I  N  N  Q  W  V  60
gtgtcagctgctcactgctacaaatcccgtatacaagtgaggctgggagagtacaacatt 240
cacagtcgacgagtgacgatgtttagggcatatgttcactccgaccctctcatgttgtaa
 V  S  A  A  H  C  Y  K  S  R  I  Q  V  R  L  G  E  Y  N  I  80
gatgtgcaggaagacagtgaagtagtcaggagttcttccgtaatcattcgccatcctaac 300
ctacacgtccttctgtcacttcatcagtcctcaagaaggcattagtaagcggtaggattg
 D  V  Q  E  D  S  E  V  V  R  S  S  S  V  I  I  R  H  P  N  100
tacagttcaagaacccttaataatgacattatgttgatcaagctggcatccgccgtggac 360
atgtcaagttcttgggaattattactgtaatacaactagttcgaccgtaggcggcacctg
 Y  S  S  R  T  L  N  N  D  I  M  L  I  K  L  A  S  A  V  D  120
tacagtgccgacgttcaacccatagccctgcccacctcttgtgccaaggcggggacggag 420
atgtcacggctgcaagttgggtatcgggacgggtggagaacacggttccgcccctgcctc
 Y  S  A  D  V  Q  P  I  A  L  P  T  S  C  A  K  A  G  T  E  140
tgcctgatttcgggctggggaaacacactgagcagtggcacctattaccctgaactcctc 480
acggactaaagcccgacccctttgtgtgactcgtcaccgtggataatgggacttgaggag
 C  L  I  S  G  W  G  N  T  L  S  S  G  T  Y  Y  P  E  L  L  160
cagtgcctgcaagcgccaattctgactaaccaagagtgccaagatgcttacccgggtgaa 540
gtcacggacgttcgcggttaagactgattggttctcacggttctacgaatgggcccaccc
 Q  C  L  Q  A  P  I  L  T  N  Q  E  C  Q  D  A  Y  P  G  E  180
atcaccagcaacatgatctgcataggattcctggagggtgggaaagactcatgccagggt 600
tagtggtcgttgtactagacgtatcctaaggacctcccaccctttctgagtacggtccca
 I  T  S  N  M  I  C  I  G  F  L  E  G  G  K  D  S  C  Q  G  200
gactcaggtggaccagttgtgtgcaacggagaactccagggcattgtgtcatggggaatc 660
ctgagtccacctggtcaacacacgttgaatcttgaggtcccgtaacacagtaccccttag
 D  S  G  G  P  V  V  C  N  G  E  L  Q  G  I  V  S  W  G  I  220
gggtgtgctctgcagggttatcctggtgtctacaccaaggtctgcaattatgttgattgg 720
cccacacgagacgtcccaataggaccacagatgtggttccagacgttaatacaactaacc
 G  C  A  L  Q  G  Y  P  G  V  Y  T  K  V  C  N  Y  V  D  W  240
atccaagagaccattgcagcctactgataccttgacaaccacctggctccttggccacta 780
taggttctctggtaacgtcggatgactatggaactgttggtggaccgaggaaccggtgat
 I  Q  E  T  I  A  A  Y  *                                   248
acctcccgccctaatgctttccctaagaagacaacagcacaaataaattcctaacttaaa 840
tggagggcgggattacgaaagggattcttctgttgtcgtgtttagggaaggattgaattt
gagccaaaaaaaaaaaaaaaaaaaaaa                                  867
ctcggtttttttttttttttttttttt
```

```
                |→     Signal     ←|→ Activation ←|←→
                      peptide          peptide                       Mature peptide
Trypsin-IIAP    MHSLFLLLSCLGAAVA FPRAADDD-K IVGGYNCPEHSVPYQVSLNAGYHFCGGSLINNQWVVSAAHCYKSRIQVRLGEYNIDVQEDSEVVRSSSVIIRH 99
                chhhhhhhhhhhhhhc cccccccc-c eecceecccccccceeeeeccceeeeeeeeecceeeecccccccccceeeecccccccccccceeeeeeeeeec
Trypsin-I AP    -MKYI.FVAFV.V... ..ISD...D. .....T.TKN.I.......S..........SS...L........S...K...H.LAS..GT.QTI...KV... 99
                 chhhhhhhhhhhhhc cccccccccc eeceeecccccccceeeeeccceeeeeeeeecceeeecccccccccceeeecccccccccccceeeeeeeeeec


Trypsin-IIAP    PNYSSRTLNNDIMLIKLASAVDYSADVQPIALPTSCAKAGTECLISGWGNTLSSGTYYPELLQCLQAPILTNQECQDAYPGEITSNMICIGFLEGGKDSCQ 200
                ccccccccccceeeecccccccccccccccccccccccccceeeecccccccccccccccceeccccccchhhhhhhcccccccceeeeccccccccccc
Trypsin-I AP    SG.NAY..D........SK.ATLNSY.NTVP.....VAT..T.............SL..DN....R..V.SSSQ.SS....Q.......V.Y......... 200
                ccccccccccceeeeeccccccccccccccccccccccccccceeeecccccccccccccccceeccccccchhhhhhhcccccccceeeeccccccccccc


Trypsin-IIAP    GDSGGPVVCNGELQGIVSWGIGCALQGYPGVYTKVCNYVDWIQETIAAY  249
                cccccceeecceeeeeeeccccccccccceeeeecccchhhhhhhhhhc
Trypsin-I AP    ...........Q............QK.............S...S....N  249
                cccccceeecceeeeeeeccccccccccceeeeecccchhhhhhhhhhc
```

Fig. 2. Alignment of precursor sequences of trypsin-IAP and trypsin-IIAP. Dashes represent similar sequences. Gaps are inserted to maximize the similarity. The lowercases indicate the online predicted secondary structures of trypsin-IAP and trypsin-IIAP (http://bioinf.cs.ucl.ac.uk/psipred/), and the letters "h", "e", and "c" represent helix, strand, and coil structure, respectively.

ClustalW program (version 1.8) on basis of tens of known trypsin precursors including the currently deduced *A. platyrhynchos* trypsins (trypsin-IAP and trypsin-IIAP). Multi-sequences were obtained from the database at NCBI. The phylogenetic trees were constructed using the neighbor-joining method (Mega, version 4.0; www.mega-software.net) by calculating the proportion of amino acid differences (p-distance) among all sequences, with sequence from a fan shell used as the outgroup. A total of 1000 bootstrap replicates were used to test the reliability of each branch. The numbers on branches indicate the percentage of 1000 bootstrap samples supporting the branch.

Expression vector construction and trypsin-IIAP expression and purification. Host strain *E. coli* BL21 and pET-32a(+) plasmid (Novagen, Germany) was utilized for trypsin-IIAP expression. The two restriction sites for Kpn I (GGTACC) upstream and Hind III (AAGCTT) downstream of the deduced mature trypsin-IIAP coding sequence were added by PCR. The forward primer was 5′-CGGGGTACCGACCCGGACCCGATTGT-3′ and the reverse primer was 5′-CCCAAGCTTTCAGTAGGCTGCAA-3′, each containing one site for Kpn I and Hind III to amplify target gene. PCR was performed by running 30 cycles with 94°C for 30 s, 55°C for 30 s, 72°C for 30 s, and a final extension 72°C for 10 min. The purified PCR product was digested with Kpn I and Hind III, and then ligated into the pET-32a(+) plasmid at the corresponding restriction sites. The resultant recombination vector was named trypsin-IIAP/pET-32a(+).

The trypsin-IIAP/pET-32a(+) construct was transformed into *E. coli* strain BL21 for protein expression and cultured to OD_{600} of 0.6 in LB broth media (0.1 mg/ml ampicillin) at 37°C. Then fusion protein expression was initiated by adding isopropyl β-D-thiogalactopyranoside (IPTG) to final concentration 1 mM. After an additional 4 h cultivation, the cells were harvested by centrifuging at 5000 rpm for 15 min at 4°C. The bacterial pellet was resuspended by adding 50 ml of buffer A (20 mM NaH_2PO_4/0.5 M NaCl, pH 7.4) and then lysed by sonication (5 s, 5 s, 30 min). The whole cell lysate was centrifuged, and the supernatant and bacterial pellet were collected. The collected supernatant and precipitate were then identified by 15% SDS-PAGE. Finally, the precipitation containing the inclusion body-expressed protein was collected, washed, and redissolved in binding buffer (20 mM NaH_2PO_4, 20 mM imidazole, 500 mM NaCl, 6 M urea, pH 7.4). The supernatant of the lysate was collected by centrifuging at 12,000 rpm for 30 min at 4°C and purified with a His-tag affinity column (HisTrap FF, 1 ml), equilibrated with buffer B (6 M urea, 25 mM Na_2HPO_4, 300 mM NaCl, pH 8.0) on an AKTA FPLC system (Amersham Biosciences, USA). Unwanted protein was eluted with buffer B containing 60 mM imidazole, pH 8.0, and the target protein was eluted with buffer B containing 300 mM imidazole, pH 8.0. The absorbance of the eluate was monitored at 280 nm. Fractions corresponding to the recombinant His-tagged protein were pooled and identified by 15% SDS-PAGE.

The collected trypsin-IIAP-containing eluate was concentrated and diluted to 0.1 mg/ml. After denaturing in urea, gradient dialysis (50 mM NaCl, 100 mM Tris-HCl/1 mM EDTA, 0.02 mM GSSG, 2 mM GSH, 1% (w/v) glycine, 10% (v/v) glycerin, and 6, 4, 2, 1 M urea each 6 h, pH 8.0) was performed, and the fusion protein was cleaved in 50% (v/v) formic acid at 45°C for 48 h. After removing formic acid by lyophilization, the trypsin-IIAP was further purified on a His-tag affinity column by reverse-phase HPLC (Hypersil BDS C18, 30 × 0.46 cm;

Elite, China) equilibrated with 5% acetonitrile and 0.1% trifluoroacetic acid.

Kinetic measurements. Serine protease activity of the expressed trypsin-IIAP was assayed based on the method by Xu et al. [23, 24]. The substrates are T6140 (*N*-(*p*-tosyl)-Gly-Pro-Lys 4-nitroanilide acetate salt; Sigma, USA) and B-3133 (*N*-benzoyl-Arg-4-nitroanilide hydrochoride, Bz-L-Arg-pNA; Sigma). The hydrolysis of synthetic chromogenic substrates by trypsin-IIAP in 50 mM Tris-HCl, pH 7.8, was tested at 37°C. The reaction was initiated by the addition of the substrate to final concentration 0.5 mM. The formation of *p*-nitroaniline was monitored continuously at 405 nm for 5 min.

Three-dimensional structure modeling. BLAST search for sequences of trypsin-IAP and trypsin-IIAP were performed to obtain the most suitable templates. Based on maximum similarity, the crystal structure of 3MYW and 1CO7_E (PDB ID) deposited in the Protein Data Bank (PDB) were used as the templates for homology modeling of trypsin-IAP and trypsin-IIAP, respectively. The comparative 3-D structure model was generated and optimized using the EasyModeller homology modeling program. The 3-D structural model generated was visualized by PyMOL software (http://www.pymol.org) without any other refinements.

RESULTS

Construction of a full-length cDNA library from *A. platyrhynchos*. Total RNA was extracted from the pancreas of *A. platyrhynchos* (with concentration 0.23 μg/μl). The OD_{260}/OD_{280} ratio was 1.88, indicating that the RNA isolated was suitable for a cDNA library construction. A 2-μg sample of total RNA was subjected to reverse transcription for synthesis of the first and double-strand cDNAs, which was concentrated on the range of 100 to 4000 bp, suggesting that double-strand cDNAs were successfully synthesized. Ten clones were picked randomly to carry out colony PCR for confirming the size of cDNA fragments. The amplified cDNA fragments were determined as ranging from 600 to 2500 bp, and 90% of the insertion fragments were more than 1 kb in size, suggesting that the insertion fragments harbored most of the mRNAs and reached the requirement for further studies on gene structure and expression. The capacity of the unamplified constructed cDNA library was $5 \cdot 10^5$ CFU/ml after calculation of clone numbers, which should meet almost all requirements to find a cDNA derived from a low-abundance mRNA. Thus, a high-quality full-length pancreas cDNA library from *A. platyrhynchos* was successfully constructed, thus providing a useful resource for functional genomic research.

Screening of the *A. platyrhynchos* cDNA library. Two hundred individual cDNA clones were picked randomly from the constructed *A. platyrhynchos* cDNA library for sequencing. After removal of the vector sequences and blast with NCBI database, we identified an EST sequence with the length of 665 bp, containing part of an open reading frame (ORF) and a polyA tail, whereas missing the 5′-UTR and adjacent part of the ORF. The blast search of this EST indicated that it showed significant homology (87%) to the known *Gallus gallus* trypsin gene (GenBank Accession No. NM_205384) with function annotation. The trypsin gene has never been reported from the avian Anseriformes.

Identification and characterization of Anseriformes trypsin I and II. The EST clone of duck trypsin was used to design the specific primers to clone the complete cDNA, and finally two full-length duck trypsin cDNAs (GenBank accession Nos. KP876029 and KP876030) were obtained. Their ORFs were both 747 bp, encoding the deduced 248 a.a. precursor (Fig. 1). Sequence alignment showed that trypsinogen-II sequence shares 73% identity with that of trypsinogen-I. The two 223-a.a. mature peptides were predicted and designated as trypsin-IAP (trypsin-IIAP), with a 15 (16)-a.a. signal peptide followed by a 10 (9)-a.a. activation peptide located at the N-terminus of their propeptides, respectively (Fig. 2). Analysis using the ProtParam tool (http://au.expasy.org/tools/protparam.html) showed theoretical p*I*/MW for trypsin-IAP and trypsin-IIAP are 7.99/23466.4 Da and 4.65/24066.0 Da, respectively.

Multiple sequences alignment and phylogenetic analysis. Multi-sequence alignment was performed on the basis of the full precursor sequences of trypsin-IAP and -IIAP with other typical trypsins from vertebrates, revealing the presence of conservative structure characteristics of trypsins, including 12 conserved cysteine residues, catalytic residues His40, Asp84, and Ser177 corresponding to the mature peptide indicated by arrowheads, and key residue Tyr151 that determines the substrate specificity (Fig. 3). The amino acids essential for calcium binding, Glu77, Asn79, Val82, Glu84, and Glu87, are also conserved in the short sequence (G76-EYNIDVQEDS-E87) of duck (underlined in Fig. 3) [25]. In addition, the critical residues involved in the maintenance of structure and function of the serine protease family were also identified, such as the short sequence of Gly-Asp-Ser-Gly-Gly-Pro, shown in frames (Fig. 3).

Multi-sequence alignment involved representative vertebrate trypsin precursors, including human, mammals, birds, reptiles, and amphibians, using fish as an outgroup. A condensed multi-furcating tree was constructed emphasizing the reliable portion of pattern branches without considering the exact distance between each peptide (Fig. 4). Thus, the branch lengths of the condensed tree are not proportional to the number of amino acid mutations. Phylogenetic comparisons of vertebrate trypsins showed that the nearest relative of duck trypsin-IIAP is *G. gallus* trypsin-II. The Anseriformes avian trypsin-I had a distinct evolution pattern from trypsin-II

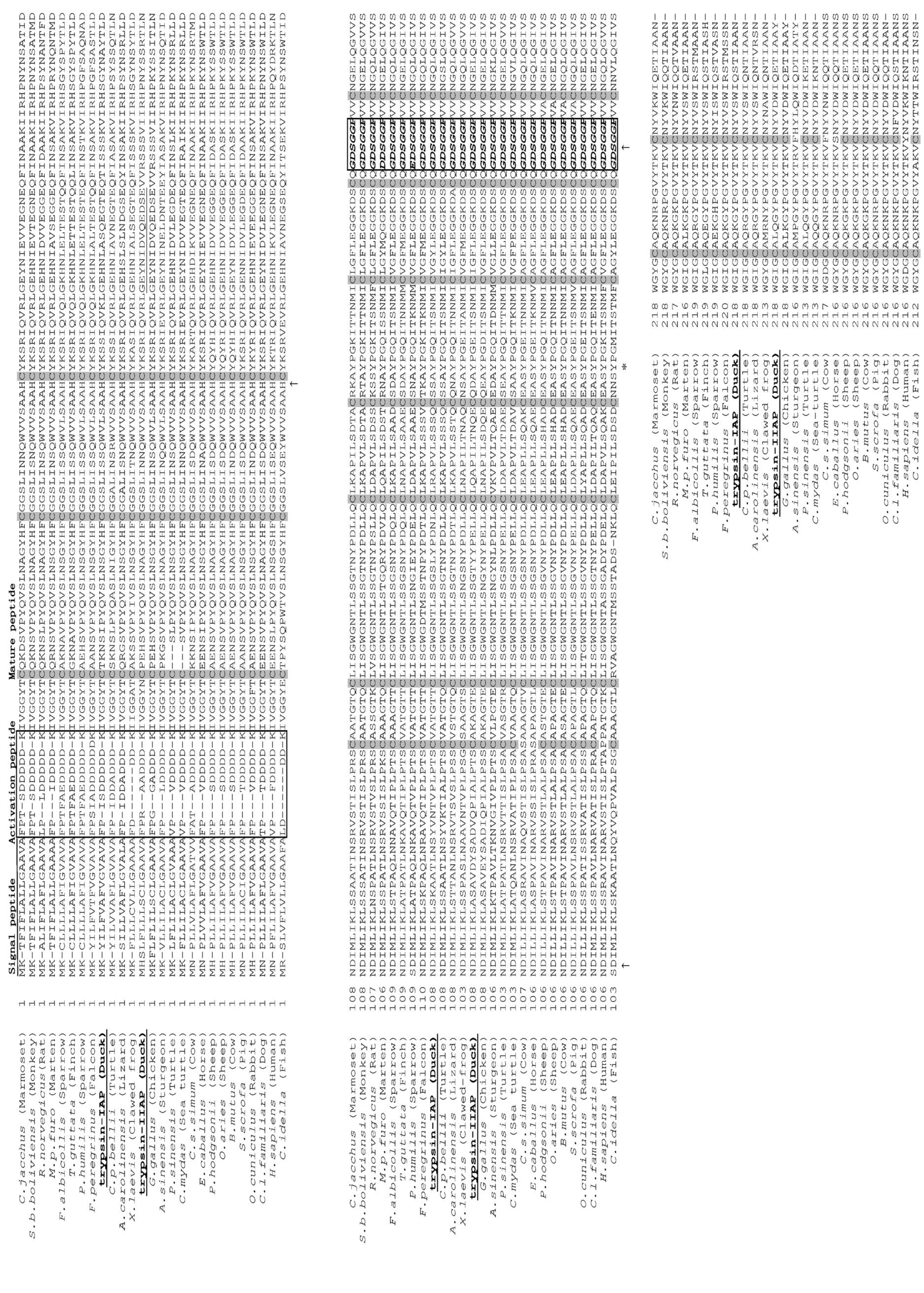

Fig. 3. Multiple sequence alignment of *A. platyrhynchos* trypsins with representative trypsins, with the 12 conserved cysteine residues shaded (*A. platyrhynchos* TPI: trypsin-IAP; *A. platyrhynchos* TPII: trypsin-IIAP). Dashes (- - -) represent gaps that have been inserted to maximize alignment; the key residue Tyr151 (Y) is marked by an asterisk; arrows indicate the catalytic residues of His40 (H), Asp84 (D), and Ser177 (S) in trypsins; the short sequence GDSGGP that is critical for the maintenance of structure and function of the serine protease family is in bold and italic.

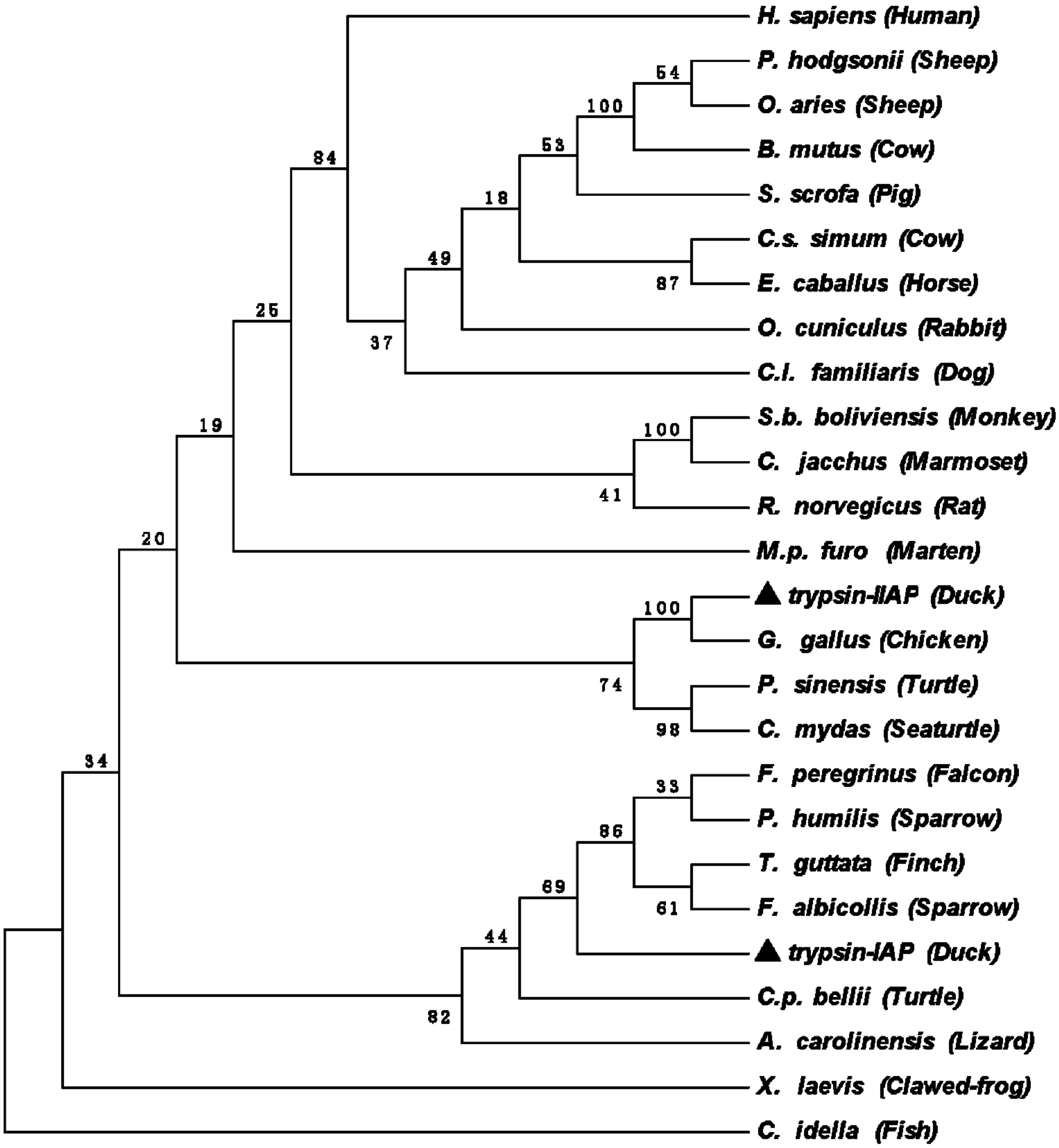

Fig. 4. Phylogenetic analysis of representative vertebrate trypsins. The phylogenetic dendrogram was constructed by the neighbor-joining method based on the proportional difference of aligned amino acid sites of the full sequence of the prepropeptide. *Anas platyrhynchos* trypsins I and II are marked with triangles.

for being juxtaposed with only turtle of reptile and sparrow of avian. Whilst based on trypsin-IIAP sequence in the phylogenetic tree, the duck (representing avian Anseriformes) is confirmed as a transitional taxon between reptiles and mammals. The built phylogenetic tree reveals that vertebrate trypsins are split into five discrete branches. Among the five branches, there are two major clusters: one is represented by various mammalian trypsins, whilst the other clusters involving trypsins from the evolutionarily inferior animals. Anuran (*Xenopus laevis*) trypsin, as a separate clade, is observed located between fish and lizard of reptiles, which bridges the evolutionary land–water gap of trypsins (Fig. 4).

Expression of trypsin-IIAP. *Escherichia coli* strain BL21 harboring the trypsin-IIAP/pET-32a(+) vector was utilized to express a His-tagged fusion protein containing the deduced mature trypsin-IIAP. After induction with 1 mM IPTG for 4 h, the fusion protein was overexpressed (Fig. 5, lanes *3* and *4*). The supernatant and precipitation were separated by centrifugation after ultrasonic treatment and identified by 15% SDS-PAGE. The protein electrophoresis figure shows the fusion protein primarily in the precipitation of whole cell lysate, indicating that the trypsin-IIAP is expressed as inclusion bodies. After His-tag affinity chromatography, the protein fractions eluted by imidazole were detected at the expected molec-

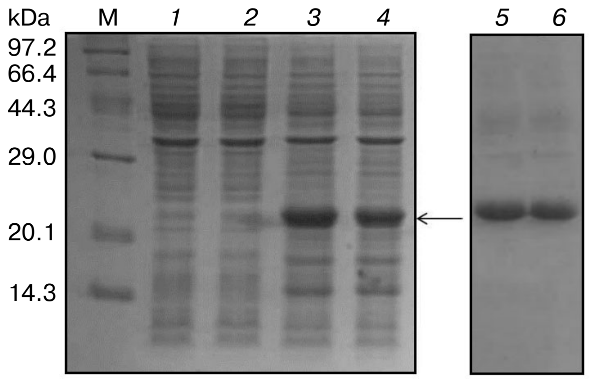

Fig. 5. Expression and purification of trypsin-IIAP protein (indicted by an arrow) as followed by SDS-PAGE (15%). Lanes: *M*, protein markers; *1*) BL21(DE3); *2*) whole cell lysate without IPTG induction; *3*, *4*) whole cell lysate with 1 mM IPTG induction; *5*) fusion protein fractions after purification by HisTrapTM FF affinity chromatography; *6*) trypsin-IIAP protein after renaturation and formic acid hydrolysis.

ular weights by 15% SDS-PAGE (Fig. 5, lane *5*). After renaturation and hydrolysis by formic acid, the supernatant was collected by centrifugation, and a 24-kDa band was detected in SDS-PAGE (Fig. 5, lane *6*).

Kinetic parameters determination. Kinetic parameters of recombinant trypsin-IIAP for different chromogenic substrates are listed in the table. T6140 is a special substrate for plasmin, whilst B3133 is the substrate for trypsin-like serine proteases. Trypsin-IIAP could hydrolyze both, indicating that trypsin-IIAP functions like other trypsin-like serine proteases.

Analyses of advanced structures of trypsin-IAP and trypsin-IIAP. Sequence alignment revealed that crystal structures (PDB ID 3MYW and 1CO7) are the best available templates for homology modeling of trypsin-IAP and trypsin-IIAP, respectively. The best models of trypsin-IAP and trypsin-IIAP were then depicted using the PyMOL Viewer as shown in Fig. 6. Homology modeled structures of trypsin-IAP and trypsin-IIAP reveal some interesting structural features: two barrel-like structures made up of two sets of antiparallel β-strands, flanked by two α-helices on the up-side, and widespread coil construction. Both of the Anseriformes trypsins have 12 cysteines, and they may form six disulfide bonds naturally. His40, Asp84, and Ser177 forming the catalytic triad are conserved as shown in Fig. 6. Residues 171-177, 192-197, and 203-206 form the primary substrate-binding pocket called the S1 binding pocket. Trypsin favors basic residues like lysine and arginine [26]. Residues 165-170 and 198-202 form two loops near the S1 pocket, called Loop 1 and Loop 2, respectively, which play key roles in enzyme specificity (Fig. 6). The S1 binding pocket in trypsin is almost identical in primary structures and backbone tertiary structures [22].

DISCUSSION

Trypsin plays a central role in pancreatic exocrine physiology because it acts as the trigger enzyme for the

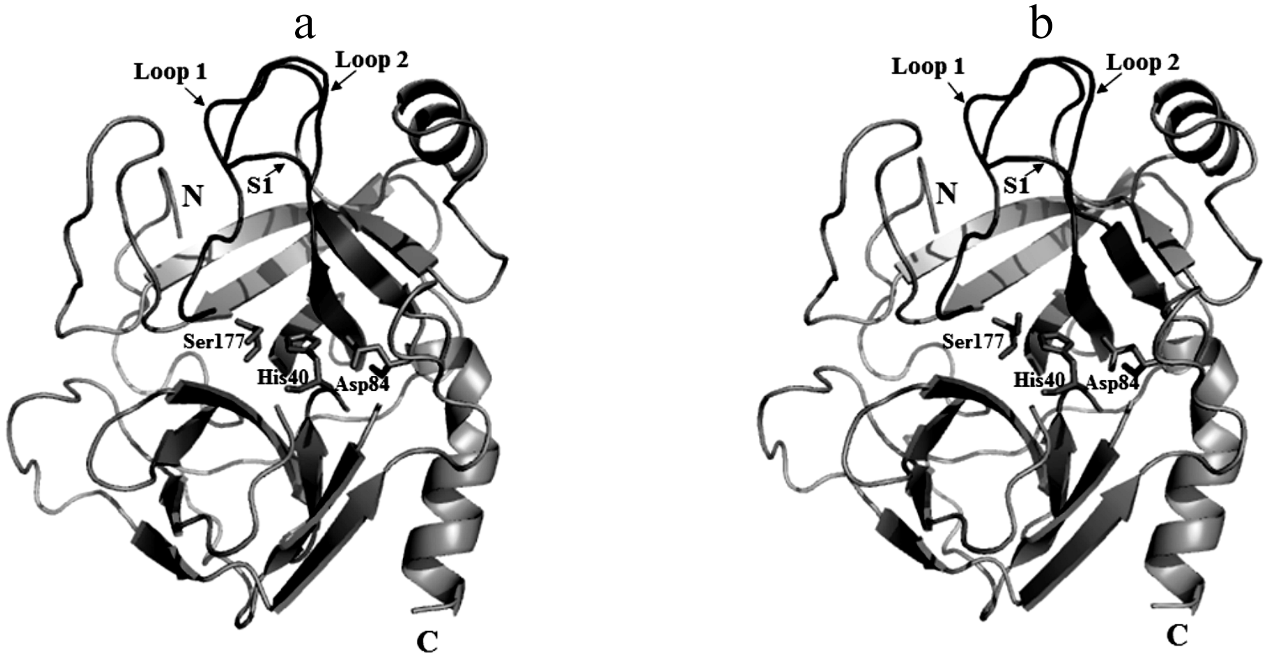

Fig. 6. Homology models of trypsin-IAP (a) and trypsin-IIAP (b). The model produced by the Mod6v2 version of MODELLER. The structure was visualized by PyMOL and is represented in the form of a cartoon. The active site residues of His40, Asp84, and Ser177 are displayed in stick structures. The substrate-binding pocket-S1 and the two loops of trypsin-IAP and trypsin-IIAP are indicated by arrows.

Kinetic parameters of trypsin-IIAP for different chromogenic substrates

Parameter	T6140	B3133
K_m (mM)	9.3	6.25
V_{max} (mM·s^{-1})	0.87	1.05
k_{cat} (10^5 s^{-1})	0.93	0.51
k_{cat}/K_m (10^5 mM^{-1}·s^{-1})	0.10	0.08

activation of all other pancreatic digestive zymogens, as well as its own trypsinogen. Regarding industrial and medical applications, trypsin is often used in proteomics, i.e. the study of proteins. As trypsin cuts proteins specifically between at amino acids arginine and lysine, it is particularly useful in the analysis of amino acid sequences within proteins. Trypsin dissociates individual cells from dissected tissue samples, allowing them to be studied more closely and/or preserved in cell cultures. More importantly, purified trypsin can be used pharmaceutically to help break down blood clots, which could result in serious complications such as strokes or emboli [27]. Trypsin is also useful in treating acute inflammation. Topical preparations of trypsin are commercially available to promote wound healing [28]. Also, trypsin is used in the treatment of pancreatic disease, and more recently it has been examined as a potential therapy in the treatment of cancerous tumors [29]. The existence of various trypsin isozymes may explain the multiple functions of trypsins and for the maintenance of multiple trypsinogen genes in genomes [22].

Trypsin is a structure-conservative serine protease, especially in some activity related sites, like the conserved 12 cysteine residues, the catalytic triplets (Ser-His-Asp), the residues for the charge relay system and the substrate-binding site residues, etc. (Fig. 3). Though a number of trypsins or trypsinogens have been well studied in mammals like bovine [4], pig [5], and human [6-8], there are few detailed reports about trypsins in avians, except one about the chicken trypsinogen gene family without any protein level findings [3], and one about ostrich (representing the Ratite superorder) trypsinogen [2]. In the current study, we successfully cloned two full-length cDNA sequences of the Anseriformes trypsin-IAP and trypsin-IIAP from a constructed cDNA library using the RACE-PCR method on basis of EST analysis. This is the first report about characterization of Anseriformes trypsin genes on the basis of EST analysis with denoted protein structure and function. Trypsinogen-II shares 73% identity with trypsinogen-I in amino acid sequence (Fig. 2). Observed from the trypsin primary and tertiary structures modeled by homology, those highly variable positions of trypsin-IAP and trypsin-IIAP are mainly located on the exterior of the molecule.

As an ancient gene family, trypsin is widely distributed in organisms ranging from invertebrates to vertebrates. However, the evolutionary relation of these important functional molecules remains unresolved. Here, the acquisition of an Anseriformes trypsin family fills gaps in trypsin molecular evolution in phylogenetic analysis. Despite the difference between sequences in the S1 pocket of trypsin-IAP and trypsin-IIAP, commonalities of precursor organization, some activity related sites, and three-dimensional structure suggests that they are evolutionarily related and likely originated from a common ancestor by gene duplication.

It was reported that a residue within the activation peptide domain, especially Asp of the penultimate residue, and the number of anionic residues, are closely related to the rate of trypsinogen autoactivation [2, 30]. Interestingly, unlike most other trypsinogens that usually have four anionic residues in the activation domain, trypsin-IAP has five Asp residues (Fig. 3), whilst trypsin-IIAP only contains three Asps adjacent to two alanines, which results in relatively less negative charge of the cleavage site and may accelerate the autoactivation process [2, 31]. The possible fast autoactivation property of trypsin-IIAP owing to its special AADDDK motif is very important for exerting enzymatic function. In addition, the sequences of trypsin-IAP and trypsin-IIAP provide new templates for the study of activation efficiency of trypsinogen and the development of trypsin inhibitor.

This work was supported by grants from the Chinese National Natural Science Foundation (41206153), Growth Project for Liaoning Distinguished Young Scientists of Colleges and Universities (LJQ2013010), and Dalian Distinguished Young Scientists Funding (2013J21DW013).

REFERENCES

1. Rawlings, N. D., and Barrett, A. J. (1994) Families of serine peptidases, *Methods Enzymol.*, **244**, 19-61.
2. Szenthe, B., Frost, C., Szilagyi, L., Patthy, A., Naude, R., and Graf, L. (2005) Cloning and expression of ostrich trypsinogen: an avian trypsin with a highly sensitive autolysis site, *Biochim. Biophys. Acta*, **1748**, 35-42.
3. Wang, K., Gan, L., Lee, I., and Hood, L. (1995) Isolation and characterization of the chicken trypsinogen gene family, *Biochem. J.*, **307**, 471-479.
4. Le Huerou, I., Wicker, C., Guilloteau, P., Toullec, R., and Puigserver, A. (1990) Isolation and nucleotide sequence of cDNA clone for bovine pancreatic anionic trypsinogen. Structural identity within the trypsin family, *Eur. J. Biochem.*, **193**, 767-773.
5. Hermodson, M. A., Ericsson, L. H., Neurath, H., and Walsh, K. A. (1973) Determination of the amino acid sequence of porcine trypsin by sequenator analysis, *Biochemistry*, **12**, 3146-3153.

6. Emi, M., Nakamura, Y., Ogawa, M., Yamamoto, T., Nishide, T., Mori, T., and Matsubara, K. (1986) Cloning, characterization and nucleotide sequences of two cDNAs encoding human pancreatic trypsinogens, *Gene*, **41**, 305-310.
7. Tani, T., Kawashima, I., Mita, K., and Takiguchi, Y. (1990) Nucleotide sequence of the human pancreatic trypsinogen III cDNA, *Nucleic Acids Res.*, **18**, 1631.
8. Wiegand, U., Corbach, S., Minn, A., Kang, J., and Muller-Hill, B. (1993) Cloning of the cDNA encoding human brain trypsinogen and characterization of its product, *Gene*, **136**, 167-175.
9. Titani, K., Ericsson, L. H., Neurath, H., and Walsh, K. A. (1975) Amino acid sequence of dogfish trypsin, *Biochemistry*, **14**, 1358-1366.
10. Stevenson, B. J., Hagenbuchle, O., and Wellauer, P. K. (1986) Sequence organization and transcriptional regulation of the mouse elastase II and trypsin genes, *Nucleic Acids Res.*, **14**, 8307-8330.
11. Craik, C. S., Choo, Q. L., Swift, G. H., Quinto, C., MacDonald, R. J., and Rutter, W. J. (1984) Structure of two related rat pancreatic trypsin genes, *J. Biol. Chem.*, **259**, 14255-14264.
12. Fletcher, T. S., Alhadeff, M., Craik, C. S., and Largman, C. (1987) Isolation and characterization of a cDNA encoding rat cationic trypsinogen, *Biochemistry*, **26**, 3081-3086.
13. Lutcke, H., Rausch, U., Vasiloudes, P., Scheele, G. A., and Kern, H. F. (1989) A fourth trypsinogen (P23) in the rat pancreas induced by CCK, *Nucleic Acids Res.*, **17**, 6736.
14. MacDonald, R. J., Stary, S. J., and Swift, G. H. (1982) Two similar but non-allelic rat pancreatic trypsinogens. Nucleotide sequences of the cloned cDNAs, *J. Biol. Chem.*, **257**, 9724-9732.
15. Kraut, J. (1977) Serine proteases: structure and mechanism of catalysis, *Annu. Rev. Biochem.*, **46**, 331-358.
16. Ma, W., Tang, C., and Lai, L. (2005) Specificity of trypsin and chymotrypsin: loop-motion-controlled dynamic correlation as a determinant, *Biophys. J.*, **89**, 1183-1193.
17. Salameh, M. A., and Radisky, E. S. (2013) Biochemical and structural insights into mesotrypsin: an unusual human trypsin, *Int. J. Biochem. Mol. Biol.*, **4**, 129-139.
18. Douglas, S. E., and Gallant, J. W. (1998) Isolation of cDNAs for trypsinogen from the winter flounder, *Pleuronectes americanus*, *J. Mar. Biotechnol.*, **6**, 214-219.
19. Gudmundsdottir, A., Gudmundsdottir, E., Oskarsson, S., Bjarnason, J. B., Eakin, A. K., and Craik, C. S. (1993) Isolation and characterization of cDNAs from Atlantic cod encoding two different forms of trypsinogen, *Eur. J. Biochem.*, **217**, 1091-1097.
20. Manchado, M., Infante, C., Asensio, E., Crespo, A., Zuasti, E., and Canavate, J. P. (2008) Molecular characterization and gene expression of six trypsinogens in the flatfish Senegalese sole (*Solea senegalensis* Kaup) during larval development and in tissues, *Comp. Biochem. Physiol. B Biochem. Mol. Biol.*, **149**, 334-344.
21. Suzuki, T., Srivastava, A. S., and Kurokawa, T. (2002) cDNA cloning and phylogenetic analysis of pancreatic serine proteases from Japanese flounder, *Paralichthys olivaceus*, *Comp. Biochem. Physiol. B Biochem. Mol. Biol.*, **131**, 63-70.
22. Roach, J. C., Wang, K., Gan, L., and Hood, L. (1997) The molecular evolution of the vertebrate trypsinogens, *J. Mol. Evol.*, **45**, 640-652.
23. Ma, D., Wang, Y., Yang, H., Wu, J., An, S., Gao, L., Xu, X., and Lai, R. (2009) Anti-thrombosis repertoire of blood-feeding horsefly salivary glands, *Mol. Cell. Proteom.*, **8**, 2071-2079.
24. Xu, X., Yang, H., Ma, D., Wu, J., Wang, Y., Song, Y., Wang, X., Lu, Y., Yang, J., and Lai, R. (2008) Toward an understanding of the molecular mechanism for successful blood feeding by coupling proteomics analysis with pharmacological testing of horsefly salivary glands, *Mol. Cell. Proteom.*, **7**, 582-590.
25. Bode, W., and Schwager, P. (1975) The refined crystal structure of bovine beta-trypsin at 1.8 Å resolution. II. Crystallographic refinement, calcium binding site, benzamidine binding site and active site at pH 7.0, *J. Mol. Biol.*, **98**, 693-717.
26. Vajda, T., and Szabo, T. (1976) Specificity of trypsin and alpha-chymotrypsin towards neutral substrates, *Acta Biochim. Biophys. Acad. Sci. Hung.*, **11**, 287-294.
27. Hsu, R. L., Lee, K. T., Wang, J. H., Lee, L. Y., and Chen, R. P. (2009) Amyloid-degrading ability of nattokinase from *Bacillus subtilis natto*, *J. Agric. Food Chem.*, **57**, 503-508.
28. Adair, J. E., Stober, V., Sobhany, M., Zhuo, L., Roberts, J. D., Negishi, M., Kimata, K., and Garantziotis, S. (2009) Inter-alpha-trypsin inhibitor promotes bronchial epithelial repair after injury through vitronectin binding, *J. Biol. Chem.*, **284**, 16922-16930.
29. Himmelfarb, M., Klopocki, E., Grube, S., Staub, E., Klaman, I., Hinzmann, B., Kristiansen, G., Rosenthal, A., Durst, M., and Dahl, E. (2004) ITIH5, a novel member of the inter-alpha-trypsin inhibitor heavy chain family is downregulated in breast cancer, *Cancer Lett.*, **204**, 69-77.
30. Sahin-Toth, M. (2000) Human cationic trypsinogen. Role of Asn21 in zymogen activation and implications in hereditary pancreatitis, *J. Biol. Chem.*, **275**, 22750-22755.
31. Chen, J. M., Kukor, Z., Marechal, C., Toth, M., Tsakiris, L., Raguenes, O., Ferec, C., and Sahin-Toth, M. (2003) Evolution of trypsinogen peptides, *Mol. Biol. Evol.*, **20**, 1767-1777.