J. Biochem. **121**, 991-995 (1997)

Clarification of an Uncertain Intron within the cDNA Sequences of Arrowhead Proteinase Inhibitors A and B[1]

Ming-Juan Luo,* Wu-Yuan Lu,† and Cheng-Wu Chi*,2,3

**State Key Laboratory of Molecular Biology, Shanghai Institute of Biochemistry, Chinese Academy of Sciences, 320 Yue-Yang Road, Shanghai 200031, China; and †Department of Cell Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037, USA*

Received for publication, September 24, 1996

An uncertain intron of 87 bp within the cDNA sequences of arrowhead proteinase inhibitors A and B was clarified. By site-directed mutation with either a stop codon inside the uncertain intron or mutated codons at both its 5′ and 3′ ends, it was proved that there was neither a translation intron nor a protein intron present in the cDNA sequences of proteinase inhibitors A and B. The primary structure of inhibitor B was then reexamined by mass spectrometry molecular weight determination and partial amino acid sequencing. A 38 residue peptide was derived by degradation of inhibitor B with lysylendopeptidase, and purified, which was not found in the previous work, and its N-terminal part was none other than the missed 29 residue peptide encoded by the uncertain intron. The 38 residue peptide was very hydrophobic, while the 29 residue peptide it included was even more hydrophobic. The N-terminal part of the missed peptide was also aligned within a BrCN-degraded fragment of the inhibitor. In this paper the cause of the overlooking of this 29 residue peptide in the previous work and some unexpected problems which arose during the former sequence analysis are explained.

Key words: amino acid sequence determination, arrowhead proteinase inhibitor, site-directed mutation.

Arrowhead proteinase inhibitors A and B are both double-headed and multifunctional serine proteinase inhibitors, which were first purified and characterized in our laboratory (*1*). Their primary structures, and the locations of the three pairs of disulfide bonds elucidated in 1992 showed that inhibitors A and B exhibit 91% sequence identity, however, they exhibit very little sequence homology with any other known proteinase inhibitors, suggesting that these inhibitors may belong to a new inhibitor family (*2*). The cDNA and genomic structures of both inhibitors A and B were also clarified in 1993 (*3*). Comparison of the amino acid sequences with their cDNA and genomic structures suggested the possible existence of a special intron of 87 bp in both the cDNA and genomic structures downstream of the 97th Lys codon, AAG, and flanked by an usual GT/AG donor-acceptor pair, coding for 29 amino acid residues.

By that time, two new concepts of regulatory elements in gene expression had been put forward, namely, the translation intron was bypassed during gene translation (*4*), and the protein intron was processed at the protein level (*5*). We wondered whether the uncertain intron of the arrowhead proteinase inhibitors may belong to one of these two special introns. With the secretory expression vector, pVT102u/α, used in our laboratory for the expression of other proteinase inhibitors (*6*), inhibitor B was successfully expressed in *Saccharomyces cerevisiae* strain S-78. Using site-directed mutation, a stop codon, TAA, was substituted for the first Tyr codon, TAC, within the 87 bp uncertain intron (Fig. 1). In the case of a translation intron, this kind of mutation will have no effect on the gene expression of the inhibitor, as this intron even with an in-frame stop codon would be bypassed during the gene translation. Our results showed that the gene expression of the mutated inhibitor was eliminated, so obviously it should not be a translation intron. Then we decided to mutate both the 5′ and 3′ ends of this uncertain intron, namely, the 97th Lys codon, AAG, and the 126th Arg codon, AGG, were both replaced by the Glu codon, GAG. If it is a protein intron, the mutated gene product would not be processed as the drastic charge change in K97E and R126E would interrupt the splicing at the level of protein processing. However, the results showed that the mutated inhibitor was expressed well with full activity. This implied that there was not a protein intron in the inhibitor cDNA. In other words, neither a translation intron nor a protein intron exists in the cDNA or genomic structure of the arrowhead proteinase inhibitors.

Finally, we tried to precisely determine the molecular weights of both the natural and recombinant inhibitor B by matrix-assisted laser-desorption mass spectrometry instead of the rather rough method of SDS-PAGE, which only gives an estimated figure. Unexpectedly, the molecular weights of both inhibitors were not 16 kDa as determined before, but 19 kDa (Fig. 2, a and b), showing the existence of an internal 29 residue peptide. As a result, we have to

[1] This work was supported by a State Biological High Technology Research Grant of China.

[2] To whom correspondence should be addressed. Tel: +86-21-64374430, Fax: +86-21-64338357.

[3] Senior Research Fellow at De Monfort University Norman Borlaug Center for Plant Science.

```
AIA                                                                                             TCTAGCTGAGCCCAATCCGATCCCCAACATC
AIB                                                                                                                      AACATC
                          -20                 -15                 -10                 -5              -1                  5
          Met Ala Ala Ser Asn Ala Leu Leu Leu Ile Ser Gly Val Leu Leu Ile Ser Leu Ala Val Leu Cys His Gly Asp Pro Val Val Asp
AIA   ACC ATG GCG GCC TCC AAC GCT CTC CTC CTC ATA TCC GGC GTT CTC CTC ATC TCG CTA GCC GTT CTT TGC CAC GGA GAT CCC GTC GTC GAC
AIB   ACC ATG GCG GCC TCC AAC GCT CTC CTC CTC ATC TCC GGC GCT CTC CTC ATC TCG CTA GCC GTT CTT TGC CAA GGA GAT CCC GTC GTC GAC
          Met Ala Ala Ser Asn Ala Leu Leu Leu Ile Ser Gly Ala Leu Leu Ile Ser Leu Ala Val Leu Cys Gln Gly Asp Pro Val Val Asp
                      10                  15                  20                  25                  30                  35
      Ser Asp Gly Asp Ala Val Gln Leu Asn Leu Gly Gly Asn Tyr Pro Leu Tyr Thr Ile Gln Ser Ala Ala Ile Gly Phe Arg Gly Gly Leu
AIA   AGC GAT GGC GAT GCG GTC CAG CTC AAC TTG GGT GGC AAC TAC CCG CTA TAC ACC ATC CAG AGT GCT GCC ATA GGC TTC CGC GGT GGG CTT
AIB   AGC GAT GGC GAT GCG GTC CAG CTC AAC TTG GGT GGC AGG TAC CCG CTA TAC ACC ATC GAA AGT GCT GCC ATA GGC TTC CAC GGT GGG CTT
      Ser Asp Gly Asp Ala Val Gln Leu Asn Leu Gly Gly Arg Tyr Pro Leu Tyr Thr Ile Glu Ser Ala Ala Ile Gly Phe His Gly Gly Leu
                      40                  45                  50                  55                  60                  65
      Ser Thr Leu His Lys Asp Ala Cys Lys Ser Tyr Val Tyr Glu Ala Pro Glu Thr Asp Arg Gly Leu Pro Val Gly Phe Ser Ala Ser Ala
AIA   TCC ACA TTG CAC AAG GAC GCC TGC AAG AGC TAC GTC TAC GAG GCC CCC GAG ACT GAC CGC GGC TTG CCG GTG GGG TTC TCG GCA TCG GCG
AIB   TCC GCA TTG CAC AAG GAC GTC TGC AAG AGC TAC GTC TAC GAG GCC CCC GAG ACT GAC CGC GGC TTG CCG GTG TCG TTC TCG GCA TCG GCG
      Ser Ala Leu His Lys Asp Val Cys Lys Ser Tyr Val Tyr Glu Ala Pro Glu Thr Asp Arg Gly Leu Pro Val Ser Phe Ser Ala Ser Ala
                      70                  75                  80                  85                  90                  95
      Thr Ser Gln Pro Val Met Gln Leu Gly Ser Arg Tyr Lys Phe Ser Phe Ser Met Pro Val Pro Leu Ile Cys Asp Thr Ala Trp Ser Ile
AIA   ACT TCT CAG CCC GTC ATG CAG CTG GGG TCC CGC TAC AAG TTC TCC TTC TCG ATG CCG GTA CCG CTC ATC TGC GAC ACC GCG TGG TCC ATC
AIB   ACT TCC GAG CCC GTC ATG CAG CTG GGG TCC CGC TAC AAG TTC TCC TTC TTG ATG CCG GTA CCG CGC ATC TGC GAC ACC GCG TGG TCC GTC
      Thr Ser Glu Pro Val Met Gln Leu Gly Ser Arg Tyr Lys Phe Ser Phe Leu Met Pro Val Pro Arg Ile Cys Asp Thr Ala Trp Ser Val
                      100                 105                 110                 115                 120                 125
      Gly Lys Ser Glu Thr Asn Gly Gly Ile Ser Phe Gln Pro Ile Thr Ala Gly Asp Tyr Phe Tyr Leu Asn Asn Phe Ser Trp Phe Glu Ala
AIA   GGC AAG TCG GAA ACG AAC GGT GGA ATC TCC TTC CAG CCG ATC ACC GCC GGG GAC TAC TTT TAC CTG AAC AAC TTT AGC TGG TTC GAG GCG
AIB   GGC AAG TCG GAA ACG AAC GGT GGA ATC TCC TTC CAG CCG ATC ACC GCC GGG GAG TAC TTT TAC CTG AAC GAC TTT AGT TGG TTC GAG GTG
      Gly Lys Ser Glu Thr Asn Gly Gly Ile Ser Phe Gln Pro Ile Thr Ala Gly Glu Tyr Phe Tyr Leu Asn Asp Phe Ser Trp Phe Glu Val
                      130                 135                 140                 145                 150                 155
      Arg Ser Thr Glu Glu Thr Gly Val Tyr Lys Leu Ala Ala Cys Ser Cys Glu Phe Cys Lys Ile Ala Cys Pro Glu Val Gly Ser Phe Asn
AIA   AGG AGC ACC GAG GAA ACC GGC GTG TAT AAG CTC GCT GCC TGC TCC TGT GAG TTC TGC AAG ATA GCT TGC CCC GAA GTA GGC TCC TTT AAT
AIB   AGG AGC ACC GAG GAA ACC GGC GTG TAT AAG CTC GCC GCC TGC TCC TGT GAG TTC TGC AAG ATA GCT TGC CCC GAA GTG GGC TCC TTT AAT
      Arg Ser Thr Glu Glu Thr Gly Val Tyr Lys Leu Ala Ala Cys Ser Cys Glu Phe Cys Lys Ile Ala Cys Pro Glu Val Gly Ser Phe Asn
                      160                 165                 170                 175                 180                 185
      Val Asn Gly Arg Thr Leu Leu Gly Ile Gly Gly Glu His Phe Thr Val Arg Phe Gln Lys Phe Asp Ala Leu Ala Met Lys Thr Ala Pro
AIA   GTC AAC GGC CGT ACC TTG CTG GGC ATC GGA GGG GAG CAC TTC ACC GTC CGG TTT CAG AAG TTC GAC GCA CTC GCA ATG AAA ACT GCA CCG
AIB   GTC AAC GGC AAA ACC TTG CTG GGC ATC GGA GGG GAG CAC TTC ACC GTC CGG TTT CAT AAG TCC GAC GCA CTC GCA ATG AAA ACT GCA CCG
      Val Asn Gly Lys Thr Leu Leu Gly Ile Gly Gly Glu His Phe Thr Val Arg Phe His Lys Ser Asp Ala Leu Ala Met Lys Thr Ala Pro

      Gln stop
AIA   CAG TAG CTAGCTAGCTGGCTGGCTACGTTCAGCAAGGGTGGTCGTGTCGTGGAACTAAGCAGTGATCGCTTGTTCGCTCATATGAATGAAGCTACAAGTTCTAAATAAGGCCGGTTC
AIB   CAG TAG CTAGCTAGCTGGCTGGTTACTTTCAGCAAGGGGGGTCGTGTCGTGGAACTAAGCAGTTAGCTCTTGTTCGCTCATATGAATGAACCTACAAGTTCTAAATAAGGCCGGTTC
      Gln stop

AIA   TCATCCTCCATGAGGTTACTACCAACTGTGATGGAAGAAGTGAAGAAAACATAAATAACAATGTCACCCGGACTCTTCGGTATAT(A)n
AIB   TCATCCTCCATGAGGTTACTACCAACTGTAATGGAAGAAGTGAAGAAAACATAAAAAACAATGTCACCCGGACTCTTCGGTATAT(A)n
```

Fig. 1. **Amino acid and cDNA sequences of arrowhead proteinase inhibitors A and B.** The uncertain intron downstream of the 97th residue Lys is in boldtype. A stop codon, TAA, was substituted for the Tyr114 codon, TAC, and the Glu codon, GAG, for both the Lys97 codon, AAG, and the Arg126 codon, AGG. All these mutations are shaded. The signal peptide and the C-terminal extra part of 7 amino acid residues are both underlined.

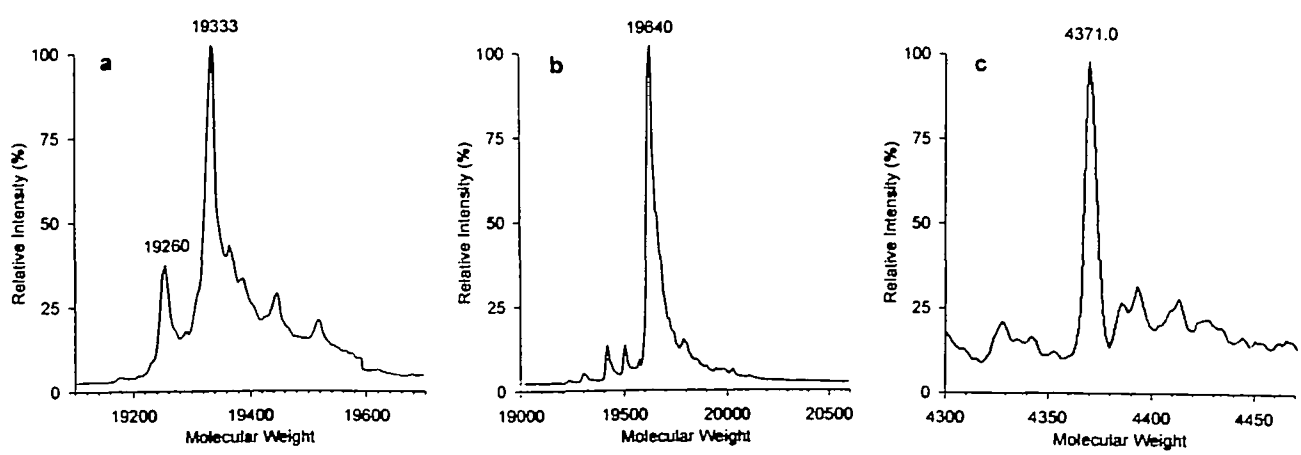

Fig. 2. **Molecular weight determination by matrix-assisted laser-desorption mass spectrometry.** (a) For the natural arrowhead inhibitor B, the several small peaks may be attributed to the polymorphism of the inhibitor (*3*). (b) For the recombinant inhibitor, that its molecular weight is higher than that of the natural one is attributed to the presence of two additional residues, Asn and Ser, at the N-terminus derived from the linker in plasmid pVT102 u/α (*6*). The determined molecular weight was 19,640 compared with the theoretical one, 19,647, the error being less than 0.04%. (c) For the 38 residue Lys-C peptide 98-135 of inhibitor B, its molecular weight is 4,371.0, *i.e.* almost identical with the theoretical value of 4,371.7.

make sure that this so called intron really corresponds to the 29 residue peptide, and to explain why this peptide was missed during the previous sequence determination. We began to reexamine the protein structure of the inhibitor.

After reduction and carboxamidomethylation of inhibitor B, the Rcam-derivative was digested with lysylendopeptidase at an enzyme/substrate ratio of 1:100, in 50 mM NH_4HCO_3 buffer, pH 8.8, 2 M urea, at 37°C for 12 h. The products were directly subjected to RP-HPLC on a C_{18} column (Fig. 3a). All the peptides thus obtained were further purified on the same column and then subjected to amino acid analysis. The results are shown in Table I, the amino acid compositions of the 9 major peaks being consistent with the amino acid sequences of the 9 lysylendopeptidase peptides derived from Rcam-AIB. The last peak eluted at a rather high acetonitrile concentration (42%), which was not observed in our previous work, corresponded to a 38 residue, peptide 98-135, including the missed 29 residue peptide. Mass spectrometry analysis of this peptide indicated a molecular weight of 4,371.0 Da (Fig. 2c). A computer program was used to automatically search for all peptides possibly cleaved by Lys-C proteinase in the entire inhibitor B sequence. It unambiguously showed that it was none other than peptide 98-135 of 38 residues, which has a calculated MW of 4,371.7, just slightly within experimental error, compared with 4,371.0 Da.

This 38 residue peptide was then subjected to sequence determination. Its partial N-terminal sequence was completely consistent with that deduced from the inhibitor cDNA (Fig. 1). Due to the high hydrophobicity of the peptide, along with the stepwise Edman degradation of some hydrophilic residues in its N-terminal part, the remaining degraded peptide became more and more hydrophobic, and was washed out with the eluting buffer used to extract PTH-amino acid derivatives, the sequencing could only be performed for 6 steps. In order to further confirm

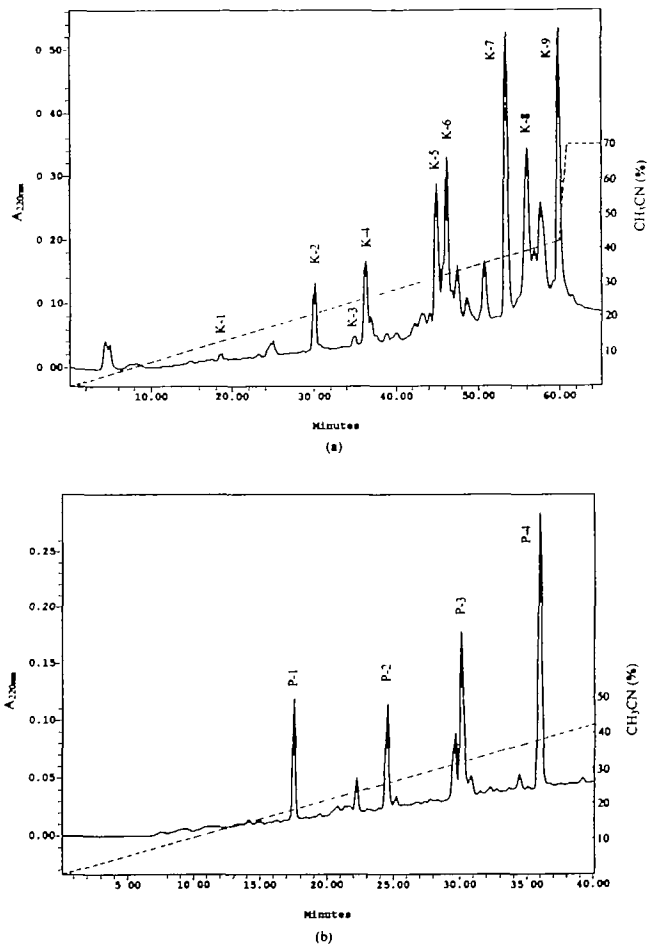

Fig. 3. **Reversed phase HPLC of the lysylendopeptidase peptides and the peptic peptides.** The C_{18} column (4.6×250 mm) was eluted with a TFA-acetonitrile linear gradient, namely, solution A was 0.1% TFA in water, and solution B was 0.1% TFA, 70% acetonitrile in water. (a) Separation of the lysylendopeptidase peptides of Rcam-AIB; (b) RP-HPLC of the peptic peptides of the last peak, K-9, in (a). The peak fractions were pooled and used for amino acid analysis, and the two fractions, K-9 and P-1, were subjected to N-terminal sequencing with an Applied Biosystems model 477 A/120 A protein sequencer and a PTH-analyzer using the program provided by the manufacturer.

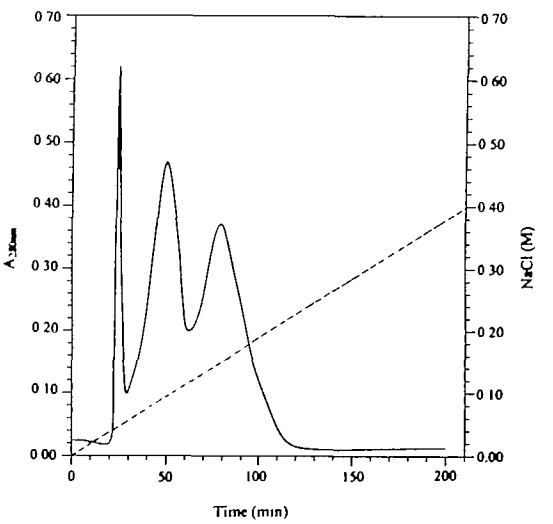

Fig. 5. **Chromatography of the BrCN-cleaved products of Rcam-AIB on a CM-Sepharose (CCF-100) column (12×120 mm).** Elution was carried out with a NaCl gradient, namely, solution A was 20 mM NaAc buffer, pH 4.2, with 7 M urea, and solution B was buffer A with 0.4 M NaCl. The peak fractions were pooled and desalted, and then used for amino acid analysis. The third fraction, C-3, was further subjected to N-terminal sequencing.

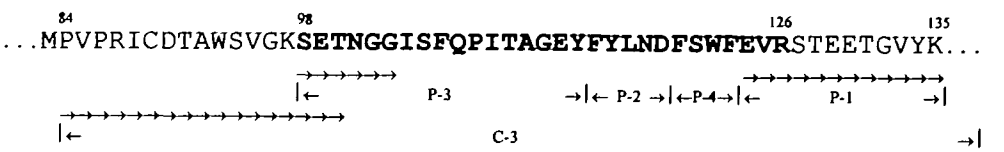

Fig. 4. **Partial amino acid sequences of BrCN peptide 84-197 and lysylendopeptidase peptide 98-135.** The missed 29 residue peptide 98-126 is in boldtype. The BrCN-cleaved and peptic peptides are designated by the prefixes C and P, respectively. The arrows indicate the sequences directly determined with the protein sequencer.

TABLE I. **Amino acid compositions of the lysylendopeptidase peptides (K1-K9) derived from Rcam-AIB.**

Amino acid	K-1 41-44	K-2 136-145	K-3 176-179	K-4 146-159	K-5 160-175	K-6 45-78	K-7 79-97	K-8 1-40	K-9 98-135
Asx	1.16 (1)		1.11 (1)	2.18 (2)		0.83 (1)	1.11 (1)	5.04 (5)	3.08 (3)
Glx		1.14 (1)		1.05 (1)	0.93 (1)	3.44 (4)		2.22 (2)	5.08 (6)
Cam-Cys	1.01 (1)	3.12 (3)		0.98 (1)			0.95 (1)		
Ser		1.04 (1)		0.95 (1)		4.72 (6)	1.96 (2)	2.80 (3)	3.44 (4)
Gly				2.05 (2)	3.39 (3)	1.97 (2)	1.34 (1)	5.33 (6)	4.12 (4)
His					2.28 (2)			2.33 (2)	
Thr					1.63 (2)	1.69 (2)	1.10 (1)	2.36 (2)	3.88 (4)
Ala		2.02 (2)	0.70 (1)	1.02 (1)		3.14 (3)	1.16 (1)	3.20 (3)	1.32 (1)
Arg					1.00 (1)	1.94 (2)	1.08 (1)	1.11 (1)	0.88 (1)
Pro				1.08 (1)		4.00 (3)	1.90 (2)	1.98 (2)	1.36 (1)
Tyr						3.33 (3)		1.91 (2)	2.72 (3)
Val	1.04 (1)			1.78 (2)	0.73 (1)	3.36 (3)	1.88 (2)	2.40 (3)	2.20 (2)
Met						1.22 (1)	0.86 (1)		
Ile				1.00 (1)	0.57 (1)		1.34 (1)	2.18 (2)	2.28 (2)
Leu		0.90 (1)	0.94 (1)		1.95 (2)	2.36 (2)	1.29 (1)	4.76 (5)	1.36 (1)
Phe		1.09 (1)	1.06 (1)	1.03 (1)	1.64 (2)	1.42 (1)	1.88 (2)	1.29 (1)	3.84 (4)
Lys	0.80 (1)	0.80 (1)		0.85 (1)	0.79 (1)	1.17 (1)	1.10 (1)	1.02 (1)	1.04 (1)
Trp							nd (1)		nd (1)
Total residues	4	10	4	14	16	34	19	40	38

TABLE II. **Amino acid compositions of the peptic peptides derived from peptide K-9.**

Amino acid	P-1 124-135	P-2 115-119	P-3 98-114	P-4 120-123
Asx	2.08 (2)	1.25 (1)		
Glx	2.96 (3)		2.79 (3)	
Ser	1.04 (1)		1.92 (2)	1.01 (1)
Gly	1.12 (1)		2.92 (3)	
Thr	1.64 (2)		1.75 (2)	
Ala			1.00 (1)	
Arg	1.08 (1)			
Pro			1.21 (1)	
Tyr	0.88 (1)	0.88 (1)	0.88 (1)	
Val	2.12 (2)			
Ile			2.42 (2)	
Leu		0.96 (1)		
Phe		0.92 (1)	1.00 (1)	1.98 (2)
Lys	1.24 (1)			
Trp				nd (1)
Total residues	12	5	17	4

the remaining sequence of this peptide, especially the sequence including the peptide bond between R126 and S127, which was the joint position between the missed 29 residue peptide and the following peptide KB-II-2 (*2*), TPCK-trypsin and *Staphylococcus aureus* V_8 protease were first used to degrade the peptide. However, the peptide could hardly be dissolved in the reaction buffer, even in that containing 20% DMSO and 2 M urea, and was not thoroughly degraded by these enzymes. Finally, it was found that the peptide could be well dissolved in a 20% HAc solution, and degraded by pepsin at an enzyme/substrate ratio of 1:20 at 37°C for 24 h. The products thus obtained were then separated by RP-HPLC on a C_{18} column (Fig. 3b). Each peak was pooled and rechromatographed on the same column, and then subjected to amino acid composition analysis. The results are shown in Table II. The amino acid composition of the first peak fragment of 12 residues corresponded to the C-terminal part of the 38 residue peptide, in which the junction between the C-terminus of the missed 29 residue peptide, Arg126, and the N-terminus of the following KB-II-2 peptide of 9 residues, Ser127, is involved (*2*). This 12 residue peptide was then sequenced and found to be: Glu-Val-Arg-Ser-Thr-Glu-Glu-Thr-Gly-Val-Tyr-Lys (Fig. 4), *i.e.* completely consistent with the sequence deduced from the inhibitor cDNA (Fig. 1).

In order to prove the presence of the peptide bond between Lys97 and Ser98, which was the joint position between the missed 29 residue peptide and the upstream peptide, KB-II-6 (*2*), Rcam-AIB was cleaved with BrCN. Three peptide fragments were obtained and then separated on a CM-Sepharose (CCF-100) column (Fig. 5). Each of these three peaks was pooled and desalted by RP-HPLC on a C_3 column, and then subjected to amino acid composition analysis. The results showed that the third peak corresponded to the C-terminal part of the inhibitor, namely, peptide 84-179. The N-terminal part sequence of this fragment was then determined by 17 steps of Edman degradation to the residue, Thr100, which definitely confirmed the presence of the peptide bond between Lys97 and Ser98 (Fig. 4).

Now it is clear that the 29 residue peptide, 98-126, was missed during the sequence determination of inhibitors A and B in our previous work. The mistake was unexpected and unfortunately due to some problems which arose during the sequence determination, as follows: (1) The 29 residue peptide, 98-126, is the most hydrophobic fragment of the inhibitor analyzed with the computer, which was either precipitated and misidentified as an unfolded intact inhibitor, or absorbed on the HPLC C_{18} column and thus could not be eluted even with a high concentration of acetonitrile. (2) The lysylendopeptidase used in the previous work seemed not to be only specific for the Lys residue, the peptide bond, Arg126-Ser127, in inhibitor A being also cleaved, which resulted in the production of the missed 29 residue peptide and the 9 residue peptide, Ser127-Lys135, designated as KA-II-2 (*2*). In general, it is more reasonable to align the KA-II-2 peptide with the Lys residue, Lys97, as we did, instead of the real residue, Arg126, the C-terminus of the missed peptide. (3) The partial N-terminal sequence of KA-II-2, Ser-Thr-Glu, happened to exhibit some similarity with that of the missed 29 residue peptide, Ser-Glu-

Thr, which was then wrongly used as an overlapping peptide. (4) Based on the molecular weights of inhibitors A and B determined by SDS-PAGE, *i.e.* around 16.5 kDa, the inhibitors were then supposed to be composed of around 150 residues. Their amino acid compositions happened to somehow match the 150 residue inhibitors on use of the computer program. (5) The cDNA sequence coding for the missed 29 residue peptide happened to be flanked by a usual GT/AG donor-accept pair. As a result, for a long time this cDNA fragment was mistakenly considered to be an unusual intron, and thus we did not pay attention to the possible loss of the 29 residue peptide. This paper has finally clarified this uncertain intron in the cDNA sequence of arrowhead proteinase inhibitors A and B, and their primary structures have been revised. They are both composed of 179 amino acid residues, as shown in Fig. 1, consistent with their cDNA and genomic structures.

REFERENCES

1. Yang, H.L., Wang, L.X., Zhu, D.X., and Chi, C.W. (1991) Inhibitory property characterization and reactive site exploration of the arrowhead proteinase inhibitor. *Sci. China* **34**, 832-839
2. Yang, H.L., Luo, R.S., and Chi, C.W. (1992) Primary structure and disulfide bridge location of arrowhead double-headed proteinase inhibitors. *J. Biochem.* **111**, 537-545
3. Xu, W.F., Tao, W.K., Gong, Z.Z., and Chi, C.W. (1993) cDNA and genomic structures of arrowhead proteinase inhibitors. *J. Biochem.* **113**, 153-158
4. Engelberg-Kulka, H., Benhar, I., and Schoulaker-Schwarz, R. (1993) Translational introns: an additional regulatory element in gene expression. *TIBS* **18**, 294-296
5. Shub, D.A. and Goodrich, B.H. (1992) Protein introns: a new home for endonucleases. *Cell* **71**, 183-186
6. Chen, X.M., Qian, Y.W., Chi, C.W., Gan, K.D., Zhang, M.F., and Chen, C.Q. (1992) Chemical synthesis, molecular cloning, and expression of the gene coding for the *Trichosanthes* trypsin inhibitor—a squash family inhibitor. *J. Biochem.* **112**, 45-51