

# DNA sequence coding for an antifreeze protein precursor from winter flounder

(preproprotein/cDNA/recombinant DNA)

PETER L. DAVIES\*, ARTHUR H. ROACH\*†, AND CHOY-L. HEW‡

\*Department of Biochemistry, Group in Eukaryotic Molecular Biology and Evolution, Queen's University, Kingston, Ontario, Canada K7L 3N6; and †Department of Biochemistry, Memorial University of Newfoundland, St. John's, Newfoundland, Canada A1C 5S7

Communicated by Norman H. Horowitz, September 28, 1981

**ABSTRACT** A cDNA made to antifreeze protein mRNA of the winter flounder was cloned in the plasmid pBR322 and its sequence was determined by the method of Maxam and Gilbert. Its sequence codes for a precursor protein that is 82 amino acids in length. This precursor has both a signal polypeptide and a pro-sequence before the mature protein of 38 amino acid residues. The mature protein matches in composition one of the alanine-rich serum antifreeze proteins that was purified by ion-exchange and reverse-phase chromatography. The composition of the pro-sequence is similar to that of the native protein except that it contains five prolines. The mature protein, but not the pro-sequence, contains three of the 11-residue repeats previously observed [Lin, Y. & Gross, J. K. (1981) *Proc. Natl. Acad. Sci. USA* 78, 2825-2829] in two other antifreeze protein components.

The winter flounder (*Pseudopleuronectes americanus*) is protected from freezing in sea water temperatures of  $-1.5^{\circ}\text{C}$  by the presence of alanine-rich antifreeze proteins in the blood (1, 2). These proteins are produced by the liver (3). Their concentration in serum increases in the fall, is at its highest during the winter, and declines rapidly in late spring (4). Duman and DeVries initially described three distinct antifreeze proteins in flounder serum (nos. 1-3) of molecular weights 12,000, 8000, and 6000 (2). These proteins differed in net charge and were numbered in order of their elution from DEAE-cellulose. From amino acid sequence data (5) it is clear that both gel filtration and NaDodSO<sub>4</sub> gel electrophoresis have given overestimates for the molecular weight of protein 3. In flounder from Newfoundland, Hew and Yip described a fraction in winter serum with antifreeze protein activity that had an apparent molecular weight of 10,000 (3). We report here that this fraction contains at least two different antifreeze proteins and that their molecular weights have also been overestimated.

One feature of the latter proteins that has not been described by other workers (6) is their initial appearance in the serum as part of a larger precursor or proprotein (7). This precursor could be detected during gel filtration chromatography of the serum as a shoulder to the main peak of antifreeze protein. Pulse-labeling experiments in which flounder were injected with [<sup>3</sup>H]alanine showed that the precursor was preferentially labeled but that after 2 days all the label had been chased into the mature protein (7). The amino acid composition of the precursor was similar to that of the mature protein except that it contained a large amount of proline (7).

The cloning and sequence analysis of antifreeze protein cDNA were undertaken in order to learn more about the structure of flounder antifreeze proteins, to characterize the precursor further, and to produce a well-defined hybridization

probe with which to study the structure and regulation of the antifreeze protein genes.

## MATERIALS AND METHODS

**Labeling and Sequence Analysis of the Antifreeze Protein Primary Translation Product.** The preparation of antifreeze protein mRNA and its translation in micrococcal nuclease-treated reticulocyte lysate (8) were carried out as described (9) but with the following modifications. The mRNA (6 μg) was translated in a final volume of 360 μl containing lysate (300 μl) and 360 μCi of [<sup>3</sup>H]leucine (60 Ci/mmol; 1 Ci =  $3.7 \times 10^{10}$  becquerels). After 60 min two aliquots (5 μl) were removed to determine the amount of label incorporated (8). The remainder of the reaction mixture was passed through Sephadex G-25 to remove unincorporated amino acids. The translation product was not further purified. Automatic Edman degradations were performed by using a Beckman 890C sequencer. All sequencer reagents and solvents were obtained from Beckman, and the Quadrol program of Beckman was used without modification. The amount of tritium in aliquots from the first 25 cycles was determined by liquid scintillation counting.

**Synthesis and Cloning of Double-Stranded cDNA.** All experiments involving recombinant DNAs were done under P1-EK2 containment conditions. Double-stranded cDNA was synthesized from antifreeze protein mRNA by the method of Wickens *et al.* (10). After treatment with S1 nuclease, the cDNA was inserted into the *Pst* I site of pBR322 by homopolymeric tailing with terminal deoxynucleotidyltransferase. Approximately 15 dGMP residues were attached to the ends of pBR322 and 25 dCMP residues to the ends of the cDNA. Plasmid and cDNA were then annealed in equimolar amounts and transfected into *Escherichia coli* HB101 (11). Tetracycline-resistant colonies were selected and were assayed on nitrocellulose filters (12) for hybridization to antifreeze protein cDNA.

**DNA Sequence Determinations.** Recombinant plasmid CT5 was digested with *Hpa* II and the DNA fragment containing the cDNA insert [477 base pairs (bp)] was recovered by electroelution after electrophoresis of the digest on a 6% polyacrylamide gel. The nucleotide sequence of the insert within this *Hpa* II fragment was determined by the method of Maxam and Gilbert (13). DNA was end-labeled with the Klenow fragment of DNA polymerase I.

## RESULTS

**Antifreeze Protein Components.** When flounder antifreeze protein was purified one major peak of antifreeze activity was

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviation: bp, base pair(s).

† Present address: California Institute of Technology, Pasadena, CA 91125.

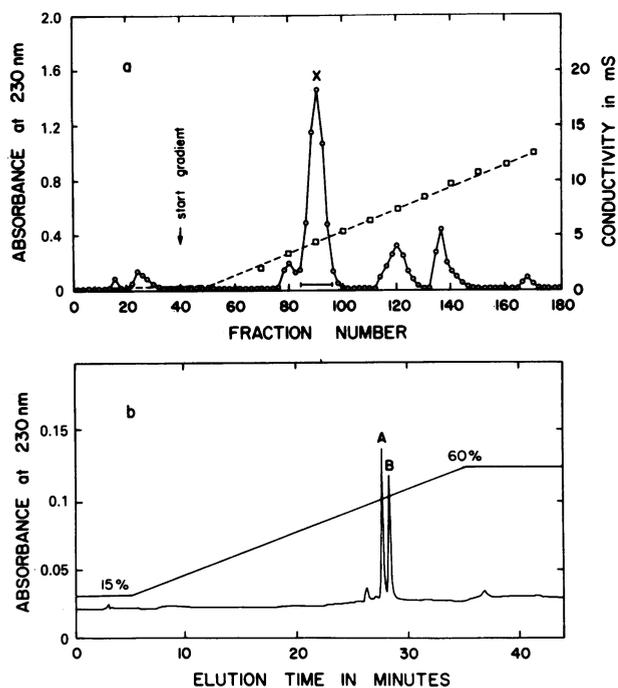


FIG. 1. Purification of serum antifreeze protein components A and B. Antifreeze protein from 8 ml of serum was partially purified by two cycles of chromatography on Sephadex G-75 (3) before application to QAE-Sephadex A-25 in *a*. Fractions (2.5 ml) were eluted by a linear gradient (400 ml) of 0–0.2 M NaCl in 5 mM Tris-HCl (pH 9.5). ○,  $A_{230}$ ; □, conductivity. Fractions spanning the bar (peak X) accounted for the bulk of the antifreeze protein activity in the sample. (*b*) The peak material from *a* was fractionated into components A and B by reverse-phase chromatography on an Altex Ultrasphere octadecylsilica column at 22°C by elution with an acetonitrile gradient at pH 3.0 (14).

eluted from the QAE-Sephadex A-25 column (Fig. 1*a*). Although this protein behaved as a single species on both ion-exchange and gel filtration chromatography, it was reproducibly separated into two components (A and B) by reverse-phase chromatography (Fig. 1*b*). On the basis of absorbance measurements components A and B made up almost 60% and 40% of the mass of the unfractionated antifreeze protein, respectively. Both components were active in freezing point depression assays. Their amino acid compositions are very similar and differ only in their content of aspartic and glutamic acids (Table 1). On the basis of there being one lysine in each protein, component A was calculated to have one glutamic and four aspartic residues, whereas component B has no glutamic residues but an extra aspartic residue. The content of glycine is far below unity in both components, and glutamic acid is present in B in trace amounts.

Table 1. Amino acid compositions of components A and B compared to the composition deduced from the cDNA sequence

Amino acid residue	Compositions from amino acid analysis				Composition from cDNA sequence
	Nanomoles		Ratio		
	A	B	A	B	
Asp	146	116	4.3	5.3	4*
Thr	136	88	4.0	4.0	4
Ser	38	30	1.1	1.4	1
Glu	36	4	1.1	0.2	1
Gly	4	7	0.1	0.3	1
Ala	769	494	22.6	22.5	23
Leu	71	44	2.1	2.0	2
Lys	34	22	1.0	1.0	1
Arg	34	21	1.0	1.0	1

Samples were hydrolyzed in 6 M HCl at 110°C for 24 hr. Analyses were performed on a Beckman 121M amino acid analyzer. The compositions of A and B are the average of two determinations. Ratios were calculated on the basis of there being one lysine residue in each protein.

\* Asparagine is tabulated as aspartic acid.

**Cloning and Sequence Analysis of cDNA.** In the clone selected for sequence analysis (CT5) the *Pst* I sites flanking the cDNA insert were regenerated. However, the presence of two other *Pst* I sites within the cDNA made this enzyme of little value in recovering the insert for sequence analysis. Therefore, the enzyme *Hpa* II was used to cut out the cDNA insert from CT5 along with short flanking sequences of pBR322 (Fig. 2). The *Hinf* I, *Dde* I, and *Sau*3A sites were mapped independently on this DNA and used along with the *Hpa* II sites to produce uniquely end-labeled fragments for sequence analysis. A second *Sau*3A site 21 bp away from the first was found (Fig. 3). The sequences of both strands of the cDNA were determined at least once except for a region of  $\approx 40$  bp flanking the first *Sau*3A site. However, the sequence of the sense strand of this region was determined three times in separate experiments (Fig. 2) and there were no ambiguities in base assignments. The antifreeze cDNA sequence is 324 bp long. It is present in CT5 in the opposite orientation to the ampicillin gene and is flanked by poly(dG)-poly(dC) homopolymeric tails of 11 and 32 bp at its 5' and 3' ends, respectively (Fig. 3).

**Preprotein Sequence.** In order to define the amino terminus of the antifreeze protein primary translation product, antifreeze protein mRNA was translated in the presence of [ $^3$ H]leucine. Thereafter, the leucine residues in the translation product were located by protein sequencing (Fig. 4). The distribution of leucine residues in this protein sequence matches the leucine codon assignments in the top line of Fig. 3. In this way the reading frame of the cDNA sequence was defined and

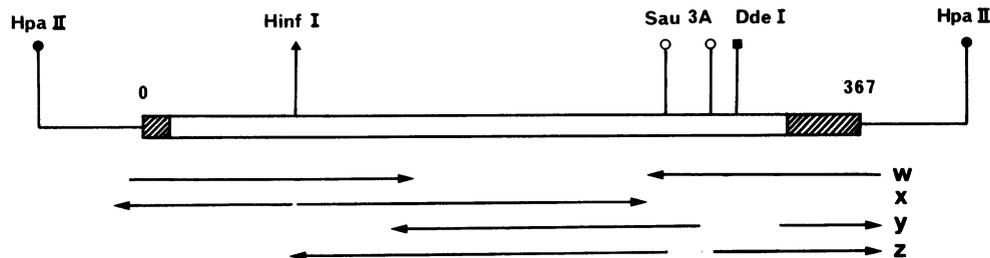


FIG. 2. Restriction map of the cDNA insert in plasmid CT5. The cleavage sites of enzymes used in determining the sequence of the insert (double line) are shown above the map. The pairs of arrows w, x, y, and z show the direction and extent of DNA sequence determined from the *Hpa* II, *Hinf* I, *Dde* I, and *Sau*3A sites, respectively. The hatched areas on the map represent the homopolymeric tails at the ends of the cDNA insert.



Table 2. Alanine codon usage in the antifreeze preprotein mRNA

Sequence	Alanine codons				Total
	GCU	GCC	GCA	GCG	
Pre-	1	—	—	—	1
Pro-	1	9	4	—	14
Protein	3	17	3	—	23
Preprotein	5	26	7	0	38

tein to proprotein. On this basis the proprotein would have glutamic acid (residue 22) at its amino terminus.

The sequence from residue 45 onwards matches closely the amino-terminal sequence of the mature serum antifreeze protein derived from a mixture of components A and B (16). The carboxyl-terminal sequence was in part derived from the gel in Fig. 5, which shows arginine and glycine codons preceding the TAA termination codon. The nucleotide sequence that follows has termination codons in all three reading frames. This sequence does not contain a poly(dA) tract or a complete version of the putative polyadenylation signal A-A-T-A-A-A, but ends in A-A-T-A. It is likely that the cDNA has been shortened to this point during S1 nuclease treatment.

**Codon Usage.** Alanine codon usage in both the pro- sequence and the sequence corresponding to the mature protein is heavily biased in favor of GCC (Table 2). Some GCA and GCT codons are used but no GCG codons.

## DISCUSSION

The amino acid composition of the antifreeze protein, as determined from the cloned cDNA sequence, matches the composition of component A very closely. The only discrepancy is

in the content of glycine. The nucleotide sequence determination in Fig. 5 clearly shows a glycine codon preceding the termination codon. However, the content of glycine in component A is well below unity, as it is for the three components described by Duman and DeVries (2). We suggest that the carboxyl-terminal glycine residue might be lost from most of the antifreeze protein molecules as a posttranslational modification. In agreement with this suggestion is the observation that the glycine content in the proprotein (7) is consistently higher than that in the mature protein (Table 1) and even equals that of arginine in some determinations, although there is no glycine in the pro- portion of the proprotein sequence (Fig. 3).

Although the cDNA for the proprotein has of necessity a large percentage of G-C base pairs, because its product is alanine-rich, there has been relatively little use made of GCA and GCT codons to reduce the overall G+C content. Instead, more than two-thirds of the alanine codons are GCC and in the non-alanine codons C is again the preferred third base over A and T. Guanine is absent from the third base position throughout the proprotein sequence. This discrimination against G in the third base position eliminates from the DNA sequence many of the restriction endonuclease cleavage sites, such as G-G-C-C, C-G-C-G, and G-C-G-C, which would otherwise have a high probability of occurring in the gene for an alanine-rich protein. These same observations apply to the antifreeze protein cDNA whose sequence was determined by Lin and Gross (6), although in this latter sequence the alanine codon usage is even more biased because GCT is not used. In theory, the alanine codons of these antifreeze protein mRNAs could be decoded by a single tRNA<sup>Ala</sup> with the anticodon sequence IGC.

The cloned cDNA sequence for component A is approximately 200 nucleotides shorter than the average length deter-

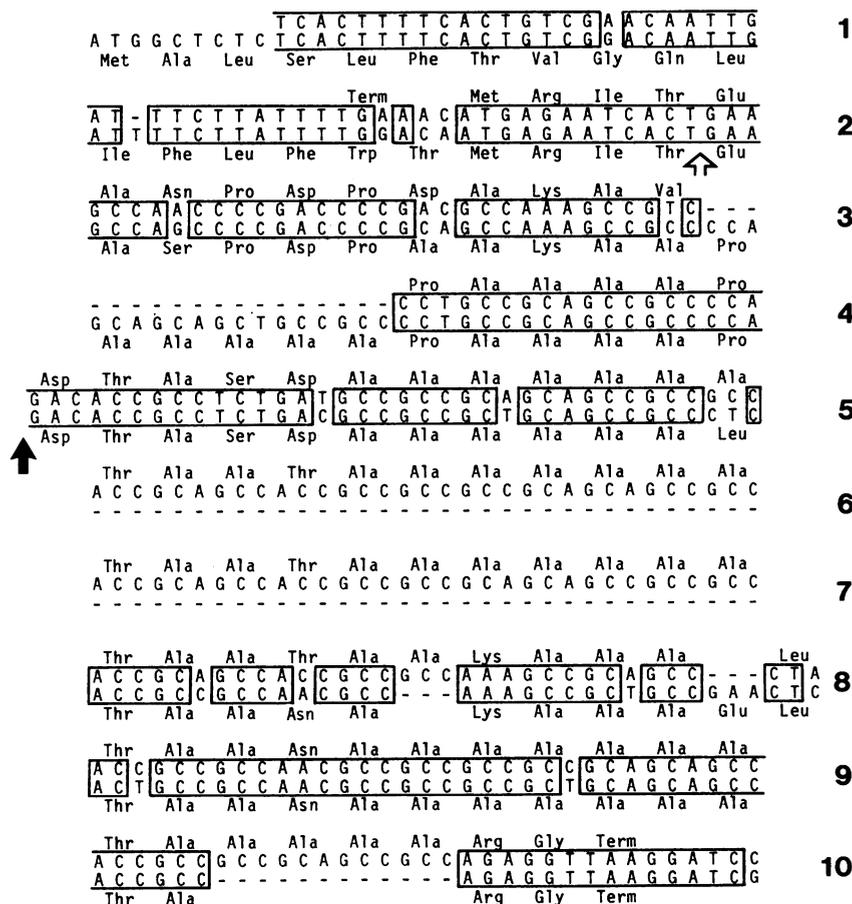


FIG. 6. Comparison of two antifreeze protein cDNA sequences. The upper sequence is from Lin and Gross (6) and the lower sequence is from Fig. 3. Regions of homology are enclosed in boxes. The open arrow marks the putative junction between the signal polypeptide and the prosegment. The solid arrow marks the start of the mature protein. The repeats of 11 amino acid residues are displayed in rows 6-9 and in row 5 beginning at the first threonine.

mined for antifreeze protein mRNA (9). This discrepancy in length could be due to the loss of the poly(A) tract and sequence adjacent to it, and to the loss of all but six of the nucleotides from the untranslated leader at the 5' end of the mRNA.

In Fig. 6 the DNA sequence for component A is compared to the cDNA sequence described by Lin and Gross (6). From radioactive sequencing of the primary translation product we have determined the reading frame of the mRNA for component A and identified the first methionine codon as the point at which translation is initiated. A second methionine is present at residue 18 (Fig. 3) in our sequence. This second methionine has been assumed by Lin and Gross to be the initiating methionine for an antifreeze protein that begins a leader sequence of 21 amino acids preceding the mature antifreeze protein sequence (Fig. 6). This leader sequence does not contain leucine and bears little resemblance to typical signal polypeptide sequences (15). The distribution of proline and hydrophilic amino acids throughout this sequence makes it unlikely that it could function in this capacity. This sequence does, however, resemble the pro-sequence previously described by Hew *et al.* (7) and identified here in component A. Upstream of the second methionine the cDNA sequences are very homologous but with two significant differences. First, in the cDNA described by Lin and Gross (6) there is one T fewer in the sequence coding for Ile-Phe, which changes the reading frame upstream in a region that is otherwise homologous to the signal peptide of component A. Second, in the section immediately preceding the second methionine, a termination codon appears in place of a tryptophan codon. These two differences in the DNAs make it unlikely that the sequence described by Lin and Gross could code for a secreted protein. Moreover, the amino acid composition of such a hypothetical protein does not match the composition of any of the major antifreeze proteins (2, 3). Consequently, their sequence may correspond to that of a pseudogene that is transcribed but not translated.

The junction between the pro-sequence and the mature serum antifreeze protein occurs at a Pro-Asp linkage and there are no basic amino acids in the vicinity. It is not clear whether cleavage here is specific for a particular amino acid sequence or whether it is specific for a structural division between the  $\alpha$ -helical mature protein (17, 18) and the proline-rich pro-sequence.

Although the nucleotide sequences of the two cDNAs are highly homologous, their predicted amino acid sequences differ significantly in length. Lin and Gross have identified an 11 amino acid residue repeat of Thr-X-X-polar amino acid-X-X-X-X-X-X-X in their sequence (6) and in that of component 3 (5) that they believe to be important in binding to ice crystals. The same repeat is observed in component A in rows 5, 8, and 9 of Fig. 6. The difference in size between the components is largely accounted for by the fact that the repeat appears five times in the sequence described by Lin and Gross (Fig. 6, rows 5-9) but only three times in component A. The few base changes be-

tween the two cDNA sequences are either silent changes or result in amino acid substitutions that can be considered conservative in terms of protein structure according to the rules of Chow and Fasman (19). For example the presence of a glutamic acid or leucine in a run of alanines is unlikely to perturb its  $\alpha$ -helical structure.

It is most probable that these two antifreeze protein sequences arose by gene duplication and that unequal recombination has altered the number of internal repeats. From our analysis of flounder genomic sequences there are at least three separate antifreeze protein gene loci that cross-hybridize extensively to nick-translated plasmid CT5 (20). The organization and sequence of these loci remain to be determined.

We thank Donna Stapleton and Christine Hough-McElhanney for technical assistance and Dr. Geoff Flynn for his advice and his help in performing the automated Edman degradation. We are also grateful to Dr. Joseph Beard and the National Cancer Institute for the gift of reverse transcriptase. This work was supported by grants to P.L.D. and C.-L.H. from the Medical Research Council of Canada and by the award of a Medical Research Council Scholarship to P.L.D. This is contribution 435 from the Marine Sciences Research Laboratory, Memorial University of Newfoundland.

1. Duman, J. G. & DeVries, A. L. (1974) *Nature (London)* **247**, 237-238.
2. Duman, J. G. & DeVries, A. L. (1976) *Comp. Biochem. Physiol. B* **54**, 375-380.
3. Hew, C. L. & Yip, C. (1976) *Biochem. Biophys. Res. Commun.* **71**, 845-850.
4. Fletcher, G. L. (1977) *Can. J. Zool.* **55**, 789-795.
5. DeVries, A. L. & Lin, Y. (1977) *Biochim. Biophys. Acta* **495**, 388-392.
6. Lin, Y. & Gross, J. K. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2825-2829.
7. Hew, C. L., Liunardo, N. & Fletcher, G. L. (1978) *Biochem. Biophys. Res. Commun.* **85**, 421-427.
8. Pelham, H. R. B. & Jackson, R. J. (1976) *Eur. J. Biochem.* **67**, 247-256.
9. Davies, P. L. & Hew, C. L. (1980) *J. Biol. Chem.* **255**, 8729-8734.
10. Wickens, M. P., Buell, G. N. & Schimke, R. T. (1978) *J. Biol. Chem.* **253**, 2483-2495.
11. Lederberg, E. M. & Cohen, S. N. (1974) *J. Bacteriol.* **119**, 1072-1074.
12. Thayer, R. E. (1979) *Anal. Biochem.* **98**, 60-63.
13. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560.
14. Seidah, N. G., Routhier, R., Bengannet, G., Larivieue, N., Gosard, F. & Chretien, M. (1980) *J. Chromatogr.* **193**, 291-299.
15. Thibodeau, S. N., Palmiter, R. D. & Walsh, K. A. (1978) *J. Biol. Chem.* **253**, 9018-9023.
16. Hew, C. L., Yip, C. C. & Fletcher, G. (1976) *Proc. Can. Fed. Biol. Soc.* **19**, 612 (abstr.).
17. Ananthanarayanan, V. S. & Hew, C. L. (1977) *Biochem. Biophys. Res. Commun.* **74**, 685-689.
18. Raymond, J. A., Radding, W. & DeVries, A. L. (1977) *Biopolymers* **16**, 2575-2578.
19. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13**, 222-245.
20. Davies, P. L., Ng, N., White, B. N. & Hew, C. L. (1981) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **40**, 1649 (abstr.).