



OPEN

# Polyproline type II helical antifreeze proteins are widespread in Collembola and likely originated over 400 million years ago in the Ordovician Period

Connor L. Scholl<sup>1</sup>, Martin Holmstrup<sup>2,3</sup>, Laurie A. Graham<sup>1</sup> & Peter L. Davies<sup>1</sup>✉

Antifreeze proteins (AFPs) bind to ice crystals to prevent organisms from freezing. A diversity of AFP folds has been found in fish and insects, including alpha helices, globular proteins, and several different beta solenoids. But the variety of AFPs in flightless arthropods, like Collembola, has not yet been adequately assessed. Here, antifreeze activity was shown to be present in 18 of the 22 species of Collembola from cold or temperate zones. Several methods were used to characterize these AFPs, including isolation by ice affinity purification, MALDI mass spectrometry, amino acid composition analysis, tandem mass spectrometry sequencing, transcriptome sequencing, and bioinformatic investigations of sequence databases. All of these AFPs had a high glycine content and were predicted to have the same polyproline type II helical bundle fold, a fold unique to Collembola. These Hexapods arose in the Ordovician Period with the two orders known to produce AFPs diverging around 400 million years ago during the Andean-Saharan Ice Age. Therefore, it is likely that the AFP arose then and persisted in many lineages through the following two ice ages and intervening warm periods, unlike the AFPs of fish which arose independently during the Cenozoic Ice Age beginning ~ 30 million years ago.

Organisms living in sub-zero environments must adapt to avoid cellular injury due to freezing. Formation of ice crystals within living tissues poses the risks of cell dehydration and the rupture of cellular membranes, leading to death<sup>1</sup>. Organisms occupying niches that experience sub-zero temperatures often produce antifreeze proteins (AFPs), which function by adhering to, and preventing the growth of, ice crystals<sup>2</sup>. Once bound, ice growth is limited to areas around the AFP, causing micro-curvatures to form on the ice surface<sup>3,4</sup>. This makes it energetically unfavourable for water to join the ice lattice resulting in a depression of the freezing temperature below the melting temperature, which is termed thermal hysteresis (TH) and is used to quantify an AFP's potency.

The first AFP characterized was from a teleost fish<sup>5</sup>. Since then, AFPs have been observed in other fishes<sup>6</sup>, insects<sup>7,8</sup>, and microorganisms<sup>9–11</sup>. In fish, AFPs are thought to have arisen during the Cenozoic era beginning 20–40 million years ago when sea ice at the poles was present for the first time in ~ 200 million years<sup>12,13</sup>. In seawater, the high concentration of NaCl (~ 0.45 M) depresses the freezing temperature of seawater to ~ - 1.9 °C. As fish blood has a lower solute concentration and freezes at ~ - 0.8 °C, any contact with ice crystals could nucleate freezing and kill the fish<sup>5</sup>. Consequently, fish AFPs must be produced in adequate quantities and with sufficient activity to decrease the body freezing temperature by at least 1.1 °C. This provides a selective advantage to these fish, as they can safely hunt for food in ice-laden seas where fish lacking AFPs are at risk of freezing. There have been four types of AFPs found in fish: (1) alanine-rich type I AFP, (2) lectin-like type II AFP, (3) type III AFP derived from sialic acid synthase, and (4) antifreeze glycoproteins (AFGPs). With different folds that all perform the same function, it raises the question of how these AFPs first arose<sup>14</sup>.

Alanine-rich alpha-helical type I AFPs have independently evolved on at least four occasions<sup>15</sup>. The simple repetitive AFGPs have arisen on two independent occasions including once from a trypsinogen gene by

<sup>1</sup>Department of Biomedical and Molecular Sciences, Queen's University, 18 Stuart Street, Kingston, ON K7L3N6, Canada. <sup>2</sup>Section of Terrestrial Ecology, Department of Ecoscience, Aarhus University, C.F. Møllers Allé 4, 8000 Aarhus C, Denmark. <sup>3</sup>Arctic Research Center, Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark. ✉email: peter.davies@queensu.ca

duplication and divergence<sup>16</sup>. Type II AFPs have evolved from a C-type lectin progenitor and have been spread to at least two distant taxonomic branches of fish by lateral gene transfer<sup>17,18</sup>. Duplication and divergence of a sialic acid synthase gene has given rise to the type III AFP gene family, which has only been found in one branch of fishes<sup>19</sup>. It is thought that these fish AFP folds have arisen within the last few tens of millions of years in response to polar glaciation<sup>20</sup>. In other branches of organisms like insects<sup>21,22</sup> and microorganisms<sup>23,24</sup> there are also examples where different AFP folds have independently arisen to perform the same task.

Collembola are the most abundant terrestrial arthropods and are found on all continents<sup>25</sup>. These small organisms (most are commonly only a few millimeters long) are typically soil dwelling, but some species live in trees, on ponds or wet surfaces surrounding stones, or on glaciers. Collembola got their name from an abdominal organ, the colophore, which is involved in osmo- and ion regulation<sup>26</sup>. A second defining abdominal anatomical feature is a spring-loaded organ known as a furca, which allows them to escape predators by “jumping”. This has given them their nickname “springtails”. These primitive organisms arose around 450 million years ago and there are close to 10,000 species classified to date<sup>27</sup>. Those that live in sub-zero environments have cold tolerance mechanisms to survive in these harsh conditions, one of which is the production of AFPs. Previously a few species of Collembola have been found to have TH activity<sup>28–31</sup>, and cold hardiness and supercooling abilities<sup>32</sup> but no systematic study of the AFP types present and their distribution in Collembola has been made to date.

The first collembolan AFP characterized was the small 6.5-kDa isoform from *Hypogastrura harveyi* (*HhAFP*)<sup>29</sup>. *HhAFP* was predicted to have a glycine-rich polyproline type II (PPII) helical bundle fold<sup>33</sup> a structure that was later confirmed by X-ray crystallography<sup>34,35</sup>. This fold, which appears unique to Collembola, consists of two layers of antiparallel PPII helices connected by loop regions. Each rotation around the helix is exactly three residues in length with tripeptide repeats of G-X<sub>1</sub>-X<sub>2</sub>, where X<sub>1</sub> is often glycine. The core of the protein contains the inward facing glycine residues, allowing for tight packing due to the absence of side chains<sup>33</sup>. Compact packing of the helices allows for a hydrogen bonding network to develop between the helix backbones that increases the stability of the fold<sup>36</sup>. The two layers both have an outward pointed face. One surface is flat, contains small, hydrophobic residues, and is thought to function as the ice-binding surface (IBS)<sup>37</sup>. This surface is alanine-rich, but also contains serine, threonine, and valine residues. The opposing surface is uneven and contains larger residues, some of which are polar or charged. *H. harveyi* also produces a larger 15.6-kDa AFP isoform that is proposed to have 13 polyproline helices<sup>38</sup>.

A putative homolog of *HhAFP* was recently characterized from *Megaphorura arctica* (*MaAFP*) collected in Iceland<sup>31</sup>. Although the 6.5-kDa *MaAFP* shares high similarities in coding sequence to *HhAFP*, their untranslated regions (UTRs) are divergent to the extent that common ancestry could not be definitively established. Additionally, 9.6-kDa AFP isoforms from *Granisotoma rainieri* (*GrAFP*) were studied<sup>39</sup>. The structure of *GrAFP* was modelled to be a polyproline type II bundle with nine helices, which was confirmed using X-ray crystallography. When the UTRs of the five *GrAFP* cDNAs were compared, the most dissimilar shared only 69% sequence identity. Therefore, a lack of similarity between the UTRs of different species does not disprove homology.

Here we have characterized the AFPs from 18 species of Collembola from numerous families within two of the four extant orders of Collembola, collected from four different continents. The extent of the analysis was determined by the amount of biomass available. Small quantities of springtails (< 100 mg of freeze-dried tissue) were used to determine TH activity and ice-shaping. AFPs were purified from larger samples (> 100 mg of freeze-dried tissue) by ice affinity purification (IAP) and characterized by MALDI-MS, amino acid composition, and/or tandem mass spectrometry. Transcriptomes were generated from some species to deduce AFP sequences at the nucleic acid level, and in some cases to recombinantly express the encoded proteins. AFPs present in the different Collembola tested here all inhibited ice growth on the basal plane, suggesting that they could be hyperactive. Where more detailed analysis was possible, all the AFPs examined had the same glycine-rich tripeptide repeating pattern indicative of the PPII helical bundle. To date, this fold has only been found in Collembola and its presence across distant species suggests that the PPII helical bundle fold originated in a basal collembolan species, shortly after the group arose.

## Materials and methods

**Collembola collection.** Here, 20 new species representing five families were tested for TH (Supplementary Table 1). Most species were collected in the field by the authors, whereas a few were supplied from cultures of other laboratories. Collembola were typically maintained for several years in Petri dishes with moist plaster of Paris mixed with charcoal, and fed dried baker’s yeast and/or green algae ad libitum. All species were kept at 20 °C with 12 h light and 12 h darkness. Exceptions from this procedure were *M. arctica* and *Entomobrya nivalis*, that were field collected and cold acclimated in the laboratory (Supplementary Table 1), and *Cryptopygus antarcticus* that were frozen immediately after being field collected. *H. harveyi*<sup>29</sup> and *G. rainieri*<sup>39</sup>, which completed the study group of 22 species, were collected as previously described.

**Sample preparation.** To induce AFP synthesis, specimens were acclimated in darkness and cold at 10 °C for 19 days, followed by 5 °C for 13 days, and 1.5 °C for 28 days. Animals were then freeze-dried for 2 days. Dried animals were added 1:8 (w/v ratio) to buffer (50 mM Tris–HCl (pH 7.8), 150 mM NaCl, 1 mM phenylthiocarbamide and 1 × EDTA-free Roche protease inhibitor cocktail) and homogenized by hand using a disposable plastic pestle in a 1.5-mL microcentrifuge tube. All manipulations and temporary storage of samples were done on ice or at 4 °C to prevent thermal denaturation of the AFP. The homogenates were centrifuged at 16,300×g for 30 min at 4 °C. The aqueous fraction beneath the lipid layer was removed for TH measurements.

**Antifreeze protein isolation.** AFP extractions for protein characterization were performed using > 100 mg of freeze-dried animals. Tissue samples were homogenized in buffer (50 mM Tris–HCl (pH 7.8), 150 mM NaCl,

1 mM phenylthiocarbamide and 1 × EDTA-free Roche protease inhibitor cocktail) using an IKA ULTRA-TUR-RAX disperser (Staufen, Germany). The homogenate was centrifuged at 22,000×g for 30 min and the supernatant was filtered through glass wool to remove lipid. AFPs in the filtered supernatant were recovered using four rounds of ice-shell purification as previously described<sup>40</sup>. The final ice fraction for each preparation was concentrated to <500 µL using an Amicon Ultracel 3 K filter (MilliporeSigma, Burlington, MA, USA) spun in a Sorvall ST16R centrifuge at 3000×g.

**Amino acid analysis.** Amino acid compositions were determined at the SickKids Proteomics, Analytics, Robotics & Chemical Biology Centre (SPARC, The Hospital for Sick Children, Toronto, ON, Canada) by acid hydrolysis as previously described<sup>40</sup> and were also used to calculate protein concentration.

**Mass spectrometry.** MALDI-TOF MS was performed at the Protein Function Discovery Facility (Queen's University, Kingston, ON, Canada) using alpha-cyano-4-hydroxycinnamic acid matrix on dried droplets with a SCIEX Voyager DE Pro in Linear mode. Protein sequencing by tandem mass spectrometry was done at the SickKids Proteomics, Analytics, Robotics & Chemical Biology Centre (SPARC, The Hospital for Sick Children, Toronto, ON, Canada).

**RNA extraction and transcriptome sequencing.** RNA samples were extracted from 30 to 50 mg of tissue stored in RNAlater (Thermo Fisher, Waltham, MA, USA). Extraction was performed as previously described<sup>39</sup>. The transcriptome of *C. antarcticus* was assembled at the Sequencing & Genotyping Center (University of Delaware, Newark, DE, USA), and those of *G. rainieri* and *E. nivalis* were assembled at the Institute for Genome Sciences (University of Maryland, Baltimore, MD, USA).

**Thermal hysteresis measurements.** The AFP-containing samples were injected into a grid filled with immersion oil on a Peltier unit. The temperature was controlled using a nanoliter osmometer (Micro-Ice Ltd, Alon Shvut, Israel) and a model 3040 temperature controller (Newport, Irvine, CA, USA). The samples were flash-frozen and melted slowly until a single ice crystal remained. The temperature was held just below the melting point and then decreased at a rate of 0.075 °C/min until ice growth began. Videos of the ice crystals during TH measurements were recorded either using Panasonic WV-BL200 CCTV camera or a DMK 33UX249 USB 3.0 monochrome industrial camera (The Imaging Source, Charlotte, NC, USA).

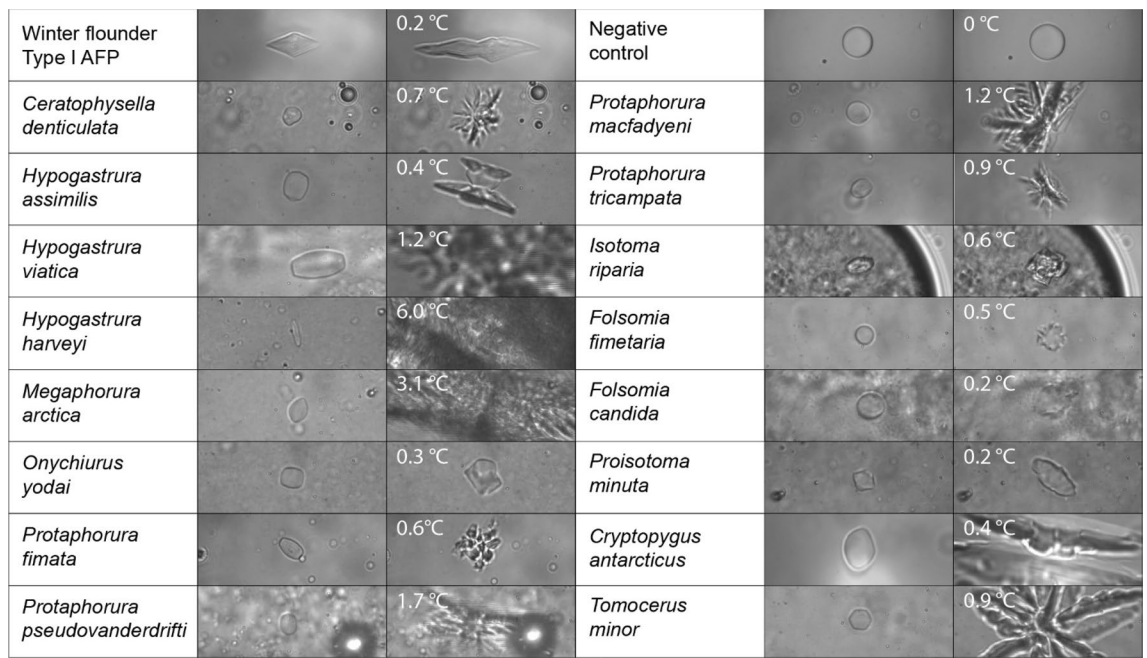
**Gene synthesis and expression.** Codon-optimized, synthetic genes for *G. rainieri* (QQY00623.1) and *Folsomia candida* (OXA44825.1) AFPs were ordered from GeneArt (Thermo Fisher, Waltham, MA, USA). The DNA encoding the signal peptide was removed and an N-terminal methionine residue was encoded to help introduce an NdeI cut site. To the C terminus, leucine and glutamate codons were introduced to add a XhoI cut site. The genes were subcloned into pET-24a vectors<sup>39</sup>. The resulting plasmids were transformed into TOP10 competent cells (Invitrogen, Carlsbad, CA, USA) for isolation using a GeneJET Plasmid Miniprep kit (Thermo Fisher, Waltham, MA, USA). DNAs were sequence checked before the plasmids were retransformed into BL21 (DE3) expression cells (Invitrogen, Carlsbad, CA, USA). Cell cultures were grown in lysogeny broth medium with 100 µg/mL kanamycin at 37 °C. Upon reaching an OD<sub>600</sub> of 0.6–0.8 the cell cultures were cooled to 20 °C, and 1 mM isopropyl β-D-1-thiogalactopyranoside was added to induce the cell culture overnight. Cells were centrifuged at 4500×g for 30 min and resuspended in 50 mL of lysis buffer (20 mM Tris/HCl (pH 7.8), 500 mM NaCl, 5 mM imidazole, 0.1 mM phenylmethylsulfonyl fluoride, and one dissolved tablet of cComplete™ ultra protease inhibitor cocktail). Cells were sonicated 16 times at 10 s per round and cooled to 4 °C between cycles to prevent protein denaturation.

**Recombinant protein isolation.** His-tagged recombinant AFPs were separated from the lysate supernatant using Ni-affinity chromatography. Fractions containing AFP were pooled, loaded into a 250-mL round-bottom flask seeded with an ice shell, and ice-affinity purified<sup>39</sup>.

**Phylogenetic tree assembly.** The taxonomy IDs for each species were extracted from the NCBI taxonomy database. The phylogenetic tree was assembled using phyloT (<https://phyloT.biobyte.de/>).

## Results and discussion

**Antifreeze activity is present in multiple genera of Collembola.** Whole homogenate supernatants from 22 different species of Collembola were assessed for TH activity and ice crystal shaping (Fig. 1). Of the 22 species tested, 18 had TH activity. The single ice crystals monitored in active homogenates melted into defined oblong shapes that were symmetrical around the *c*-axis, suggesting that the AFPs are binding to and stabilizing several ice planes, including the basal plane. In contrast, crystals formed in the presence of an AFP that does not bind to the basal plane, namely type I AFP, melt into discs that grow into hexagonal bipyramids as they are being cooled (Fig. 1, top left panels). When the freezing point was exceeded in collembolan samples, dendritic ice growth emanated from the *a*-axes and grew rapidly in samples that had high TH activity. In *Hypogastrura viatica* and *M. arctica* the dendritic burst covered the field-of-view within one frame (1/12 s). In contrast, ice crystals in the buffer control were disk-shaped and kept growing with the same shape, while the burst with type I AFP occurred along the *c*-axis (Fig. 1 top right panels). Samples were homogenized at the same w/v ratios making comparisons of relative TH activity possible, and the activity in different species ranged from 0.2 to 1.7 °C. These differences could arise from variation in gene copy number, expression levels and/or the activity of the AFPs.



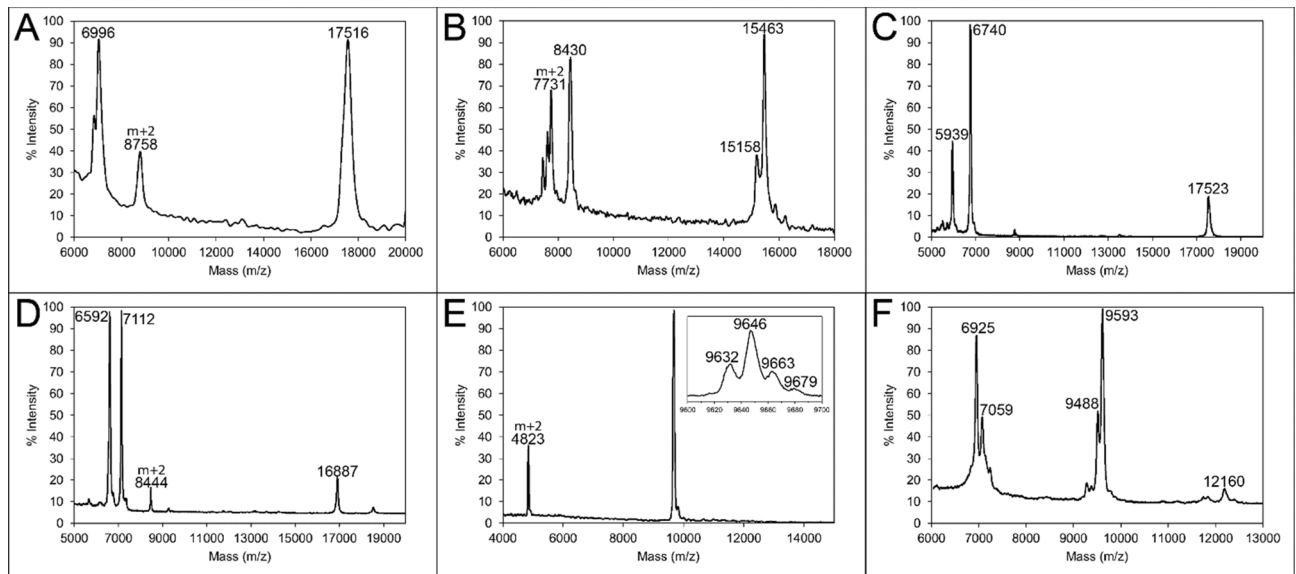
**Figure 1.** Comparison of ice-shaping in collembolan homogenates. Freeze-dried Collembola were gently homogenized in buffer (8:1 v/w) and the supernatants were assayed for antifreeze activity (TH). For each species (left column), an image was captured from the video during the TH measurement (middle column) and just as the freezing point was exceeded (right column). The positive control is type I AFP from winter flounder (*Pseudopleuronectes americanus*). Negative control is buffer (50 mM Tris-HCl (pH 7.8), 150 mM NaCl, 1 mM phenylthiocarbamide and 1 × EDTA-free Roche protease inhibitor cocktail).

Fluorescent ice plane analysis of the large *H. harveyi* AFP isoform showed binding to both basal and prism planes of ice<sup>38</sup>. Support for basal plane binding was also provided by the X-ray crystal structure of *GrAFP*, as the crystallographic waters could be aligned to both the basal and primary prism planes of ice<sup>39</sup>. In addition, the high activity (> 2 °C) of two of the homogenates suggest that, as in other arthropods such as *Tenebrio molitor*<sup>41</sup>, most Collembola produce hyperactive AFPs. Additionally, the consistent differences in ice shaping and the burst between collembolan AFPs, and type I AFP from winter flounder (Fig. 1) is likely due to basal-plane binding by the collembolan AFPs.

**Multiple springtail species produce glycine-rich AFPs.** Amino acid compositions of AFPs extracted from three species of Collembola (*H. harveyi*, *G. rainieri*, and *M. arctica*) have been previously reported<sup>29,31,39</sup>. Here, AFP extracts from three additional species (*Cryptopygus antarcticus*, *Folsomia candida*, and *Protaphorura pseudovanderdrifti*) were also subjected to amino acid analysis (Supplementary Table 2). All had high abundances of glycine and alanine, which are diagnostic of the PPII helical bundle fold. However, these glycine and alanine proportions were lower than for *H. harveyi*, which had been further purified, suggesting some contamination by trace levels of other proteins. This is to be expected as each round of IAP only reduces non-AFP protein levels ~ 10 fold<sup>40</sup>. Nevertheless, MALDI-MS suggested that the AFPs were the dominant species after IAP (Fig. 2). *P. pseudovanderdrifti* (Fig. 2A), *C. antarcticus* (Fig. 2B), and *Ceratophysella denticulata* (Fig. 2C) extracts all showed a few discrete peaks, corresponding to two or more small isoforms (5.9–8.8 kDa) and one or more large isoforms (15.5–17.5 kDa), similar to what was previously reported for *M. arctica* (Fig. 2D)<sup>31</sup> and *H. harveyi* (6.5 and 15.7 kDa)<sup>29</sup>. In contrast, *F. candida* has one main peak consisting of four sub-peaks that differ by ~ 16 Da (Fig. 2E) and *G. rainieri* (Fig. 2F)<sup>39</sup> has three clusters of isoforms within a narrower range of 6.9–12.2 kDa. There also appear to be small variations in the isoforms as indicated by the shoulder peaks. Much like the five 9.6-kDa isoforms from *G. rainieri*<sup>39</sup>, there are likely isoforms within each population with a few amino acid polymorphisms.

**AFP sequence determination.** Partial sequencing of the purified AFPs by tandem mass spectrometry of tryptic fragments provides a robust link between the isolated proteins and their nucleic acid sequences. This has previously helped deduce full-length *M. arctica* and *G. rainieri* AFP sequences from transcriptome data<sup>31,39</sup>. In this study, tryptic fragments of *C. antarcticus* and *F. candida* AFPs were also rich in glycine and alanine and contained GXX repeat motifs (Table 1).

There are over 3400 *C. antarcticus* ESTs, from two studies<sup>42,43</sup>, in the NCBI database. BLAST searches identified two with full-length coding sequences (GR869204.1 and FF279148.1), and two with incomplete coding sequences (GR870234.1 and FF278983.1). The full-length transcripts encoded a 108-amino-acid protein (8.5 kDa) after the removal of the 19-amino-acid signal peptide (Fig. 3A). GR870234.1 was missing the N-terminal methionine

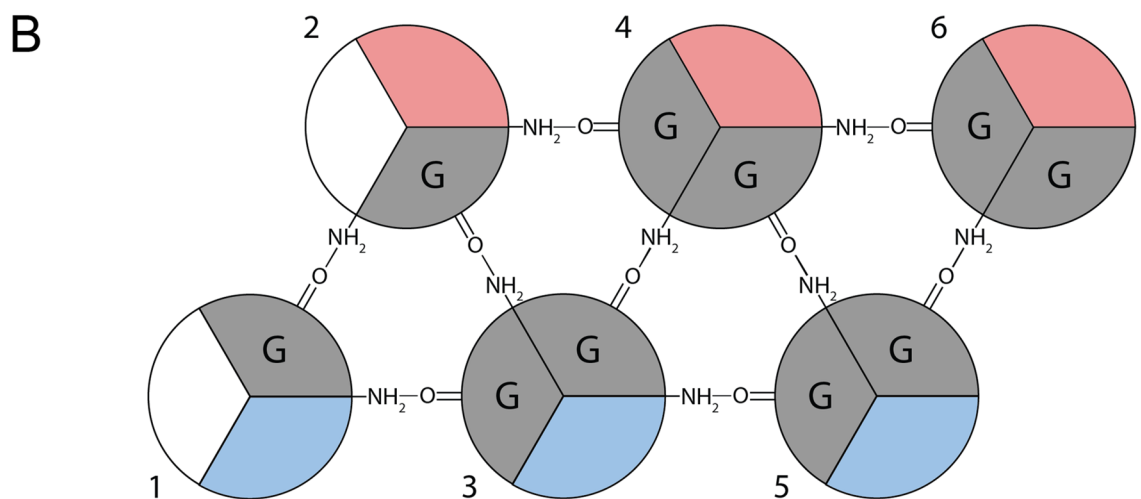
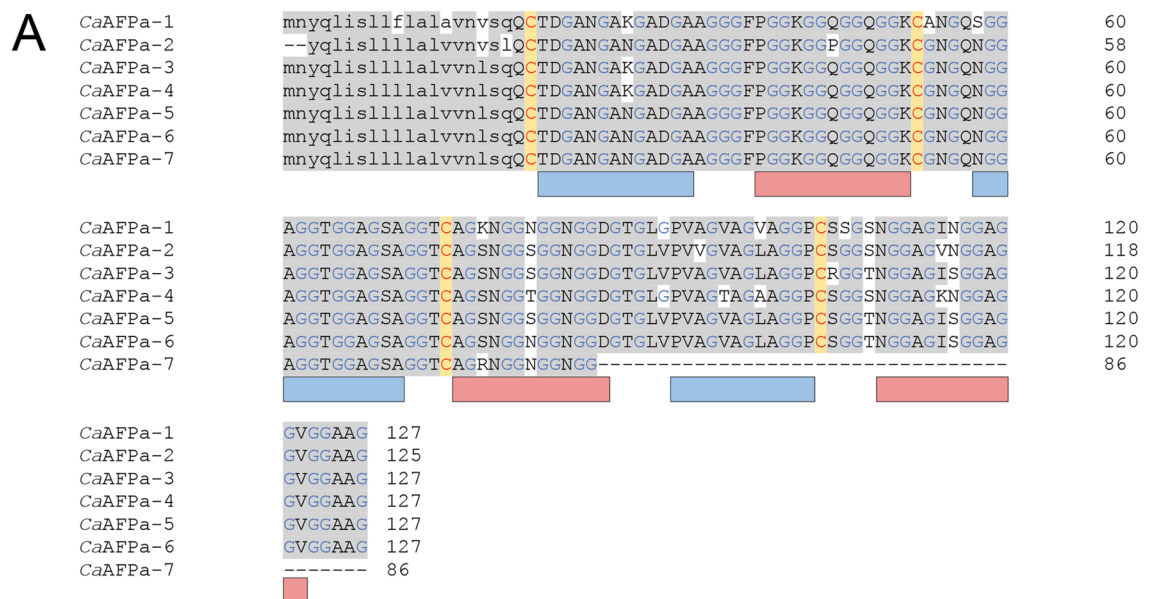


**Figure 2.** MALDI spectra of purified collembolan AFP extracts. Proteins from: (A) *Protaphorura pseudovanderdrifti*, (B) *Cryptopygus antarcticus*, (C) *Ceratophysella denticulata*, (D) *Megaphorura arctica*, (E) *Folsomia candida*, and (F) *Granisotoma rainieri* were purified with four rounds of ice-affinity purification and subjected to MALDI-MS. Major peak masses are labelled.

Mass (Da) <sup>a</sup>	Sequence <sup>b</sup>	Isoform match
1953.8	(CA)NGQSGGAGGTGGAGSAGGTCAGK	CaAFP <sub>a</sub> -1
1921.8	(CT)DGANGANGADGAAGGGFPGGK	CaAFP <sub>a</sub> -2,5,6,7
1118.5	(GA)DGAAGGGFPGGK	CaAFP <sub>a</sub> -1-7
1960.9	(AGA)ANTNGNGGIGGAGGAGGACGTK	CaAFP <sub>b</sub> -1-4
699.4	(GG)VGAPGGK	CaAFP <sub>b</sub> -5-7
747.3	(pQC)TTGVRC	CaAFP <sub>b</sub> -5-7
7206.2	GANGVSGITGAAGIGGPGGQPGAGQGGACYGAFNGGVGGHGGNGGAAGAGQSGPNGADG-GAGGDA   GPSG   N   GG   AGGYGGHGGASGASGASK <sup>d</sup>	CaAFP <sub>b</sub> -5-7
1210.6	(GA)DGPVGAASSGAHK	CaAFP <sub>b</sub> -8
1336.6	(GA)GGYGGAGGAGGAAGTK <sup>e</sup>	CaAFP <sub>b</sub> -8
1504.7	(GA)GGYGGAGGAGGAAGTKPA	CaAFP <sub>b</sub> -8
1515.7	AGTAGTHGGFGGAGGAGGVG	CaAFP <sub>b</sub> -8
1618.8	(SN)GIVNAGGTGCNNGGVGK	CaAFP <sub>b</sub> -8
1945.9	(PA)GIGAAGGNGGNAGTGGAGGYGK <sup>e</sup>	CaAFP <sub>b</sub> -8
2303.0	(GGA)GGASGTCNGHDGAAGGAGGVGDSGTK	CaAFP <sub>b</sub> -8
2494.1	(CFGTHNGGAGGA)GAPGGAASSAHDGPAGGK <sup>f</sup>	CaAFP <sub>b</sub> -8
3263.5	(GAG)GYGGAGGAGGAAGTKPAGIGAAGGNGGNAGTGGAGGYGK	CaAFP <sub>b</sub> -8
2493.0	(CS)GAAGADGTSNGQAGLAGTAGGPGCDGGR <sup>f</sup>	FcAFP
3192.4	(SGTGTG)GHGGNGDGHPPGA(PG)AGGA(PN)GL(PG)SA(GNTLP) <sup>f</sup>	FcAFP
3979.8	(GGAG)FPGTNVPGGAGGAGGAGGSGNTAGGAGGHGSSNTLTGGAGGAGGIR <sup>f</sup>	FcAFP

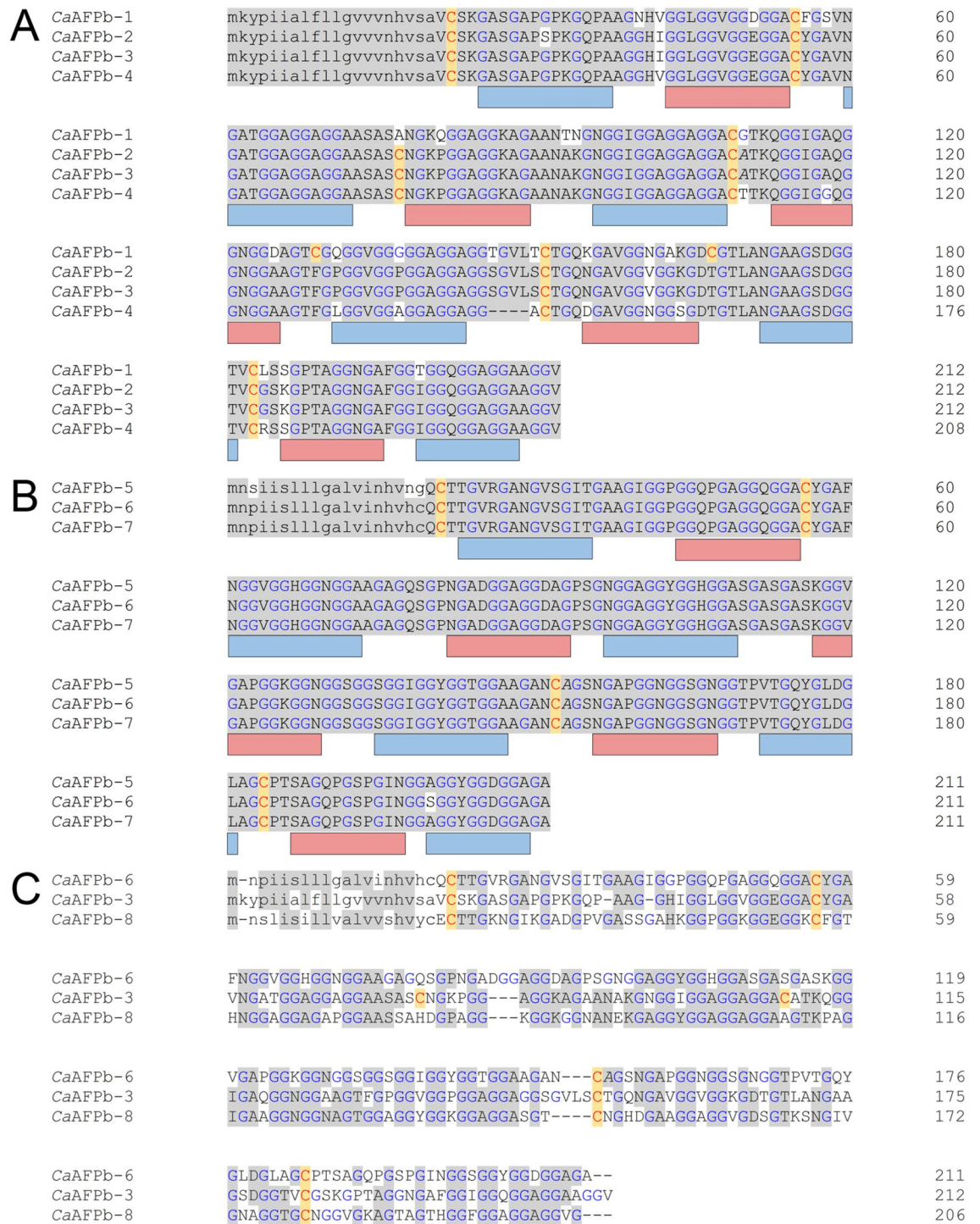
**Table 1.** Tryptic fragment sequences of *Cryptopygus antarcticus* and *Folsomia candida* AFPs determined by tandem mass spectrometry. <sup>a</sup>Monoisotopic [M+H]<sup>+</sup>. <sup>b</sup>Residues in brackets were consistent with the mass of the lightest b-ion observed. <sup>c</sup>Fragment mass consistent with N-terminal Gln cyclization (pyroglutamate). <sup>d</sup>Fragment masses exceeded instrument cutoff, but in-source fragmentation generated a series of C-terminal fragments beginning at Lys and extending to the residues delineated with |. <sup>e</sup>Fragments resulting from cleavage of 3263.5 Da product between Lys and Pro. <sup>f</sup>Bolded proline residues were seen to be occasionally hydroxylated.

start codon and FF278983.1 had a truncated C terminus at residue 86. Peaks with similar masses were seen in the MALDI profile (Fig. 2B). Additional isoforms were identified within the transcriptome generated in this study; three that encode 8.5-kDa isoforms (OQ445583, OQ445586, and OQ445587) and eight that encode 15-kDa isoforms (OQ445584, OQ445585, OQ445588, OQ445589, OQ445590, OQ445591, OQ445592, OQ445593). The 8.5-kDa isoforms had between 91 and 93% sequence identity at the nucleotide and protein levels (Fig. 3A).



**Figure 3.** Sequence alignment of the 8.5-kDa *CaAFP* isoforms. **(A)** The amino acid sequences of *CaAFP* isoforms from ESTs and the generated transcriptome were aligned. Identical amino acid residues are highlighted in grey. Glycine residues are coloured blue. Cysteine residues are coloured red and highlighted yellow. The signal peptides are shown in lowercase and the coding sequence is shown in uppercase. The rectangles below show putative PPII helices for the IBS (blue) and non-IBS (red) faces. *CaAFP*a-1, OQ445586; *CaAFP*a-2, GR870234; *CaAFP*a-3, OQ445587; *CaAFP*a-4, FF279148; *CaAFP*a-5, OQ445583; *CaAFP*a-6, GR869204; *CaAFP*a-7, FF278983. **(B)** Schematic of six polyproline type II helical bundles. The antiparallel helices are connected by loop regions (not shown) and arrange into two layers. The ice-binding surface (blue) and non-ice-binding surface (red) are shown. Hydrogen bonding between helices stabilizes the fold.

Using a schematic to visualize each helix the 8.5-kDa isoforms can be modelled to have 6 helices (Fig. 3B). The strings of 3–4  $GX_1X_2$  repeat motifs can be separated into individual helices separated by variable loop regions that are 3–6 amino acid residues in length. The  $X_2$  residues in one helix and the following helix alternate between a hydrophobic and hydrophilic residue, and this produces the distinctive ice-binding site (Fig. 3 blue surfaces) and non-ice-binding site (Fig. 3 red surfaces). The 15-kDa isoforms clustered into three groups. The first group had three isoforms with a 20-amino-acid signal peptide and 192-amino-acid mature protein (Fig. 4A). *CaAFP*b-4 had a deletion of four amino acids, making the mature protein 188 amino acids in length. The second group had three isoforms with a 19-amino-acid signal peptide and 192-amino-acid mature protein (Fig. 4B). The third type had a single isoform (*CaAFP*b-8) with a 19-amino-acid signal peptide and a 187-amino-acid mature protein (Fig. 4C). Within the first and second groups, the isoforms showed between 88–99% and 98–99% sequence identity, respectively, while between these groups and the third isoform type there was only 50–52% sequence identity. Within each group, the amino acid sequences in the loop and non-IBS regions were less conserved between isoforms. When the three groups of the 15-kDa *CaAFP*s are compared, the number of PPII helices was constant, with 11 predicted (Fig. 4), and the length of each helix was roughly equivalent, with three to four GXX or GGX repeats each. Additionally, the number of disulfide bonds predicted varied, from 2 to 4. Predicted



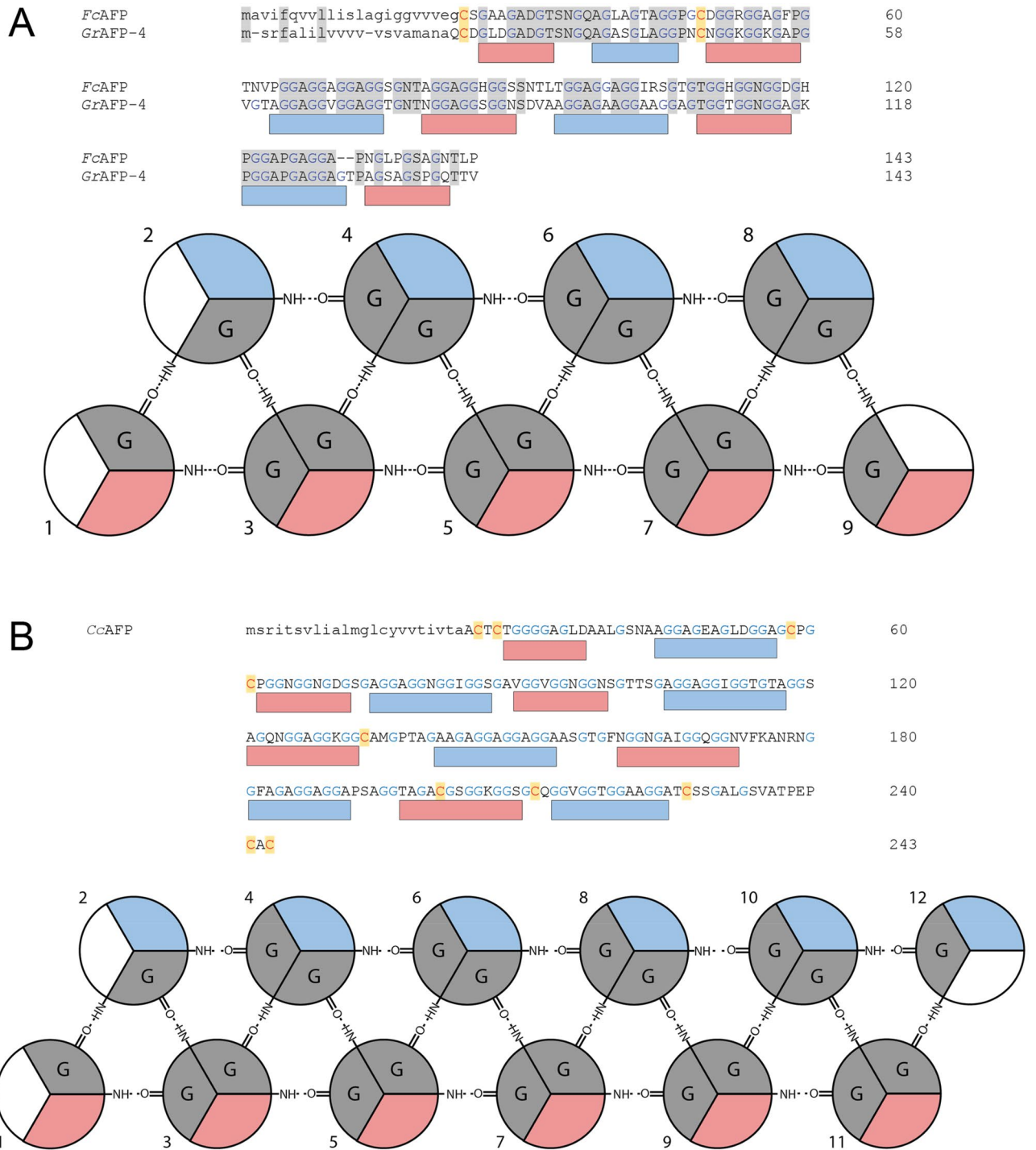
**Figure 4.** Sequence alignment of the 15-kDa CaAFP isoforms. The amino acid sequences of CaAFP isoforms from the generated transcriptome were aligned for the (A) first and (B) second groups. (C) The third CaAFP isoform (CaAFPb-8) is aligned to one sequence from the first group (CaAFPb-3) and the second group (CaAFPb-6). The colouring is the same as in Fig. 3. CaAFPb-1, OQ445590; CaAFPb-2, OQ445584; CaAFPb-3, OQ445585; CaAFPb-4, OQ445589; CaAFPb-5, OQ445588; CaAFPb-6, OQ445592; CaAFPb-7, OQ445593; CaAFPb-8, OQ445591.

average masses of CaAFPb-8 (15.2 kDa) and CaAFPb-6 (15.5 kDa) closely match peaks 15,158 and 15,463 Da seen by MALDI (Fig. 2B).

The annotated assembly of the genome of *F. candida*<sup>44</sup> was found to contain a single gene encoding a glycine-rich sequence (OXA44825.1), herein called FcAFP, resembling other PPII helical AFPs. FcAFP shares 57%

sequence identity with *GrAFP-4* (Fig. 5A). *FcAFP* was predicted to have a structure very similar to that of *GrAFP-4* solved by crystallography<sup>39</sup>, with nine PPII helices forming an ice-binding face made up of the four even-numbered helices.

The tryptic fragments analyzed by tandem mass spectrometry (Table 1) support the contention that *F. candida*, unlike the other species examined, has but one AFP isoform. The dominant spectra matched the three predicted tryptic fragments, and they were sequenced multiple times, over most of their length. As the genome<sup>44</sup> and this sample (Supplementary Table 1) were derived from the same parthenogenetic laboratory strain, this exact match was not unexpected. Similar fragments of different masses arose either from in-source fragmentation of



**Figure 5.** Polyproline type II helical bundle schematics. The protein sequences of *FcAFP* and *CcAFP* arranged into individual polyproline helices. (A) *FcAFP* can be arranged into 9 helices and (B) *CcAFP* can be arranged into 12 helices. Colouring is the same as in Fig. 3.



the tryptic fragment<sup>45</sup>, or via post-translational modification. Fragments were sometimes 16 Da heavier than expected, with the additional mass coinciding with proline residues that were followed by glycine residues (Table 1, in bold).

The gene sequence and tryptic fragment sequences were consistent with the masses observed by MALDI-MS. The dominant peak was at 9646 m/z, with the double-charged species at 4832 m/z, but close examination of the spectrum (Fig. 2E, inset) reveals four peaks differing by ~ 16 Da. The lightest, at 9632 m/z, closely matches the average mass of 9630 Da predicted for the mature protein without modifications, whereas the others at 9464, 9663 and 9679 likely contain 1, 2 or 3 modified proline, respectively. A mass increase of 16 Da is consistent with hydroxylation of proline. Collagen contains X<sub>1</sub>-X<sub>2</sub>-G repeats and the structure consists of three PPII helices in parallel that form the collagen triple helix<sup>46</sup>. Proline is modified to hydroxyproline within X-P-G motifs throughout<sup>47</sup>. Therefore, it is likely that the same process is responsible for modifying, on average, one of the six X-P-G motifs in the AFP, given that the sequence repeats and secondary structure of the two proteins are similar.

The extracted protein from *F. candida* had only 0.2 °C of TH activity at the concentration tested, lower than most other species (Fig. 1). When recombinantly expressed in *E. coli*, FcAFP reached 0.52 ± 0.01 °C of TH at a concentration of only 2.3 μM. This suggests that even after cold acclimation, the levels of AFP produced in the animal from this single gene are well below what can be attained in vitro. Despite producing an AFP, when *F. candida* were exposed to -3 °C for 15 days none of specimens survived<sup>32</sup>. However, when starved their supercooling point is below -15 °C<sup>48</sup>. Arthropods with high TH activity generally have more than one AFP gene, producing different AFP isoforms, as exemplified by the spruce budworm moth with its 16 gene copies<sup>49</sup>, and the other springtails herein.

*Entomobrya nivalis* was previously known to have up to 3.5 °C TH activity within their hemolymph during the winter months<sup>50</sup>. Therefore, the limited number of animals that were collected in the field were acclimated, before their RNA was extracted. A transcriptome was generated from which five AFP isoforms were identified (Supplementary Fig. 1B). Each sequence had a signal peptide between 22 and 24 amino acids in length. The mature proteins had predicted masses of 8.9 kDa, 9.6 kDa, and 11.2 kDa. The 8.9-kDa isoform (EnAFP-2) was 86% identical to the 11.2-kDa isoform (EnAFP-3) with a 29-amino-acid deletion that likely removes two helices. The other sequences had between 64 and 81% identity.

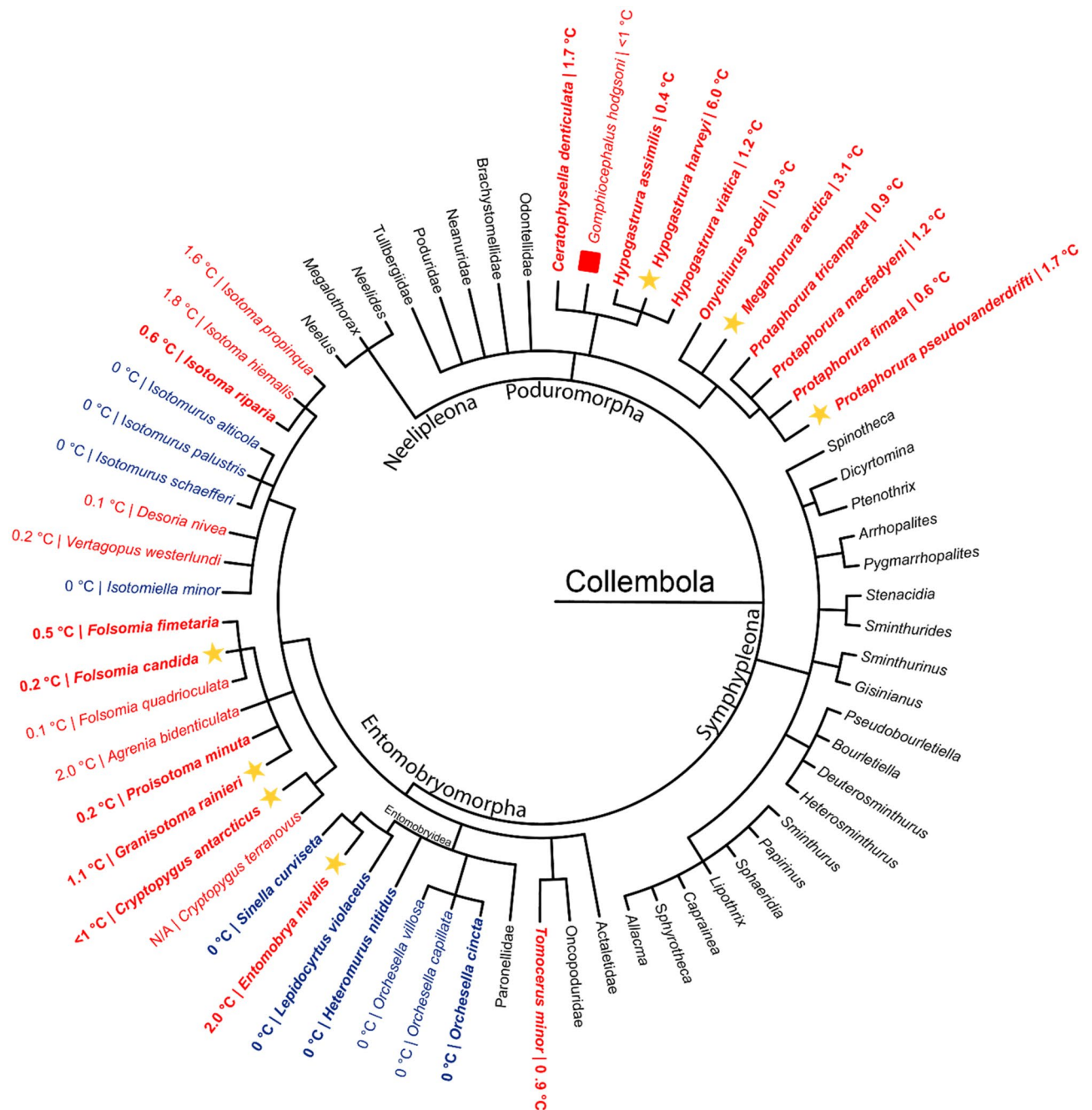
Using the sequences of other collembolan AFPs as a BLAST query, a sequence resembling a PPII helical AFP was found in the genome of *Ceratophysella communis* (VNWX01004235.1). The gene was predicted to contain a single intron<sup>51</sup>, and the resulting 732-bp open reading frame encoded an AFP predicted to have a 23-amino-acid signal peptide<sup>52</sup>. The mature protein was 220 amino acids in length and can be modelled to have 12 PPII helices (Fig. 5B). Although *C. communis* was not collected, *C. denticulata* was, and the homogenate had 0.7 °C TH (Fig. 1). Additionally, the largest peak on the MALDI-MS had a mass of 17.5 kDa, similar to the 17.2 kDa predicted for the mature *C. communis* AFP sequence.

**Polyproline AFPs are spread across orders of Collembola.** Ten of the collembolan species sampled were from Poduromorpha and 12 were from Entomobryomorpha (Fig. 6). All ten poduromorphs and eight entomobryomorphs had AFP activity. The four species lacking AFP activity were from the family Entomobryidae, but *E. nivalis* was inside this family as well and it did have AFP activity. Fourteen other species have been tested for TH activity by others<sup>28,30,53</sup>, for a total of 11 of 11 poduromorphs and 15 of 25 entomobryomorphs testing positive for AFP activity. Of the species that did not produce AFPs, two were again found in Entomobryidae, as well as four from Isotomidae.

The glycine-rich PPII helical bundle was predicted to be the AFP fold of eight species of Collembola, spanning five families and two orders. The presence of PPII helical AFPs in both Entomobryomorpha and Poduromorpha suggests that this protein family originated prior to their divergence (Fig. 7). The exact sequence of collembolan taxonomic diversification is still being debated. The four orders can be arranged into assorted sister clades depending on the datasets used. When using 18S and 28S sequences Neelipleona was basal to (Symphypleona + (Entomobryomorpha + Poduromorpha))<sup>54</sup>. Yet, when using 16S, 28S, and *cox1* sequences the positions of Neelipleona and Symphypleona were reversed and Symphypleona was basal<sup>55</sup>. Regardless of the exact phylogenetic relationship, mitochondrial dating suggests that the four orders diverged between 437 and 421 million years ago<sup>56</sup>. This period coincides with the Andean-Saharan glaciation, an ice age that lasted from 460 to 420 million years ago<sup>57</sup>. Additionally, the lineages corresponding to extant families diverged between 414 and 184 million years ago<sup>56</sup>, in which the Karoo glaciation occurred, between 360 and 255 million years ago<sup>58</sup>. Diversification and the need for freeze resistance in the same timeframe would allow for radiation of the species expressing PPII helical AFP.

The timing of diversification into genera could have led to species lacking AFPs. Within the superfamily Entomobryoidea, *Sinella curviseta*, *Heteromurus nitidus*, *Lepidocyrtus violaceus* and three species from the genus *Orchesella* did not display antifreeze activity AFPs, while *E. nivalis* did (Fig. 6). Analyses of a sample of Collembola from China estimated that five polyphyletic clades of the genus *Entomobrya* diversified between 66 and 34 million years ago during the Paleocene–Eocene thermal maximum<sup>59</sup>. During this period, around 55 million years ago, global temperatures were an average of 5–8 °C higher than today<sup>60</sup>. *Sinella* is estimated to have diverged from one *Entomobrya* clade around 69 million years ago, while *Heteromurus* and *Orchesella* diverged from the family *Entomobryidae* around 100 million years ago<sup>59</sup>. The ancestor of the AFP-lacking species in the superfamily Entomobryoidea might have lost their AFP during this period, leading to radiation of species without an AFP gene, while the *E. nivalis* lineage retained theirs.

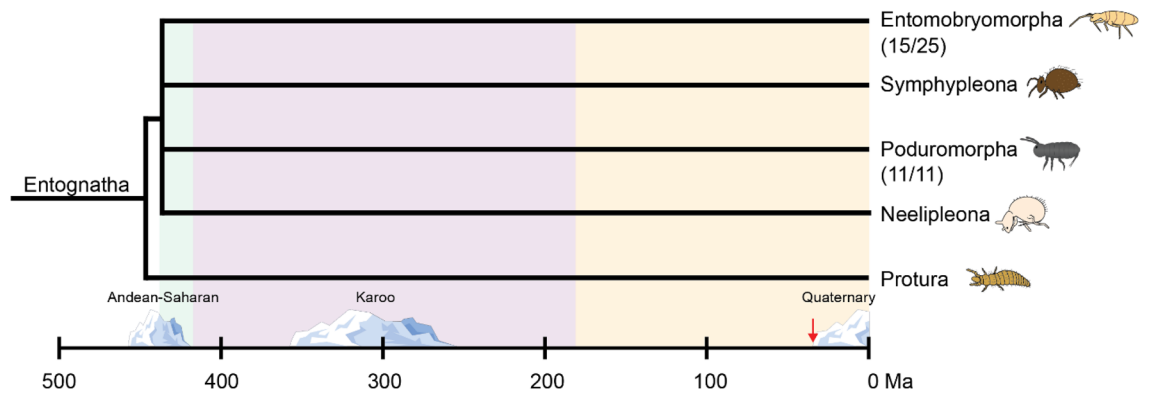
Unlike teleost fish, where AFPs did not originate until the Cenozoic era (~ 30 million years ago) following the Paleocene–Eocene thermal maximum<sup>20</sup>, there is no sign of a sudden diversification of collembolan AFPs during this event or the previous Karoo glaciation. In theory, certain lineages may have lost their PPII AFP type



**Figure 6.** Taxonomic tree of antifreeze-protein-producing Collembola. A taxonomic tree for species from the four orders of Collembola (Entomobryomorpha, Poduromorpha, Symphyleona, and Neelipleona) was generated based on NCBI taxonomy. Species with and lacking TH activity are coloured in red and blue, respectively, and groups not tested are in black. Species assayed in this paper are bolded and all other species not bolded were tested by Zettel<sup>28</sup>, except *Gomphiocephalus hodgsoni*<sup>30</sup> and *Cryptopygus terranovus* (syn. *Gressittacantha terranova*)<sup>53</sup>. Species with a star produced glycine-rich AFPs. *Gomphiocephalus hodgsoni* with a red square is a proposed cystine- and histidine-rich AFP<sup>30</sup>.

gene(s) during an interglacial period, only to evolve a replacement to cope with a new ice age. For this reason, it is worth noting the different amino acid composition of an AFP previously identified in the Antarctic collembolan species, *Gomphiocephalus hodgsoni*<sup>30</sup>. This AFP contained high percentages of histidine and cystine that set it apart from the PPII-type AFPs (Supplementary Table 2). However, the sequence of this AFP has not yet been identified. Interestingly, this species is a member of the Hypogastruridae family in which two species (*H. harveyi* and *C. communis*) produce glycine-rich PPII helical AFPs.

The origins and relatedness of the PPII AFPs are difficult to determine as the repetitive nature of the protein makes comparisons between distantly related species extremely difficult. Homology cannot be easily inferred from repetitive sequences, especially when they are under selection for antifreeze activity. For example, the



**Figure 7.** The relationship between collembolan phylogeny and ice ages. A phylogenetic tree showing the timeline of divergence in Collembola. The shaded green, purple, and yellow show estimated times for divergence of orders, families, and genera/ species, respectively. Number of AFP-producing species are shown below the order name. The red arrow indicates the emergence of fish AFPs during the Cenozoic era. Icebergs display the timespan of each respective ice age.

alanine-rich type I AFPs of fish, some of which have threonine residues at 11-amino-acid intervals, initially appeared homologous, but they are now known to have evolved via convergence within the last 30 million years<sup>15</sup>. Fortunately, the origin of the flounder type I AFPs was traced via their UTRs<sup>20</sup>. It is possible that convergence also played a role in the evolution of collembolan AFPs. However, this seems unlikely, as all but one species examined to date produce glycine-rich AFPs and the known sequences form<sup>34,39</sup>, or can be modelled (Figs. 3, 5)<sup>31,38</sup> as, PPII helical bundles with an ice-binding face. In contrast, when AFPs arose in teleost fishes and insects, a variety of different protein folds were used as AFPs<sup>61–65</sup>. Additionally, the length of the PPII helices do not vary between isoforms or species. When an extra GGX repeat was added to each helix of *GrAFP* the TH activity decreased suggesting that this could be a selective pressure to limit the length of the helices<sup>37</sup>.

It is unlikely that an analysis of the UTRs of collembolan AFPs will provide clues as to their origins. Many of the species studied herein diverged much earlier than teleost fish (Fig. 7). This has provided ample time for their non-coding regions to diverge to such an extent as to be unrecognizable as homologous. This is evident even when the 5'- and 3'-UTRs are compared between transcripts from a single species. For example, the 5'- and 3'-UTRs of *EnAFP*s have between 65–93% and 60–84% sequence identity, respectively. The lack of non-coding sequence identity has been previously reported between *HhAFP* and *MaAFP*<sup>31</sup>, and between isoforms of *HhAFP*<sup>38</sup> and *GrAFP*<sup>39</sup>. This suggests that some of these PPII AFPs, even those from the same species, have been diverging for far longer than 30 million years.

One limitation of this study was our inability to sample species from either Neelipleona or Symphypleona. There is currently an underrepresentation of genomic and transcriptomic data for these orders relative to Entomobryomorpha and Poduromorpha, but nine genome sequences are publicly available at NCBI. Although PPII AFPs were not found in the eight Symphypleona and one Neelipleona genome sequences, it should be noted that the repetitive nature of the PPII AFPs, along with the abundance of glycine-rich genes (such as collagen), introns, and repetitive sequences replete with potential glycine codons, make identification of AFP genes difficult. Therefore, identification of these AFPs is heavily reliant on tissue extraction. Unfortunately, to the best of our knowledge, laboratory culturing of Symphypleona and Neelipleona is difficult, complicating detailed studies on AFPs in these two orders.

### Data availability

The datasets generated during the current study are available in the GenBank repository (<https://www.ncbi.nlm.nih.gov/genbank/>) under the accession numbers OQ511494-98 and OQ445583-93.

Received: 31 March 2023; Accepted: 26 May 2023

Published online: 01 June 2023

### References

- Mazur, P. Freezing of living cells: Mechanisms and implications. *Am. J. Physiol.* **247**, C125–142. <https://doi.org/10.1152/ajpcell.1984.247.3.C125> (1984).
- Duman, J. G. Antifreeze and ice nucleator proteins in terrestrial arthropods. *Annu. Rev. Physiol.* **63**, 327–357. <https://doi.org/10.1146/annurev.physiol.63.1.327> (2001).
- Raymond, J. A. & DeVries, A. L. Adsorption inhibition as a mechanism of freezing resistance in polar fishes. *Proc. Natl. Acad. Sci. USA* **74**, 2589–2593. <https://doi.org/10.1073/pnas.74.6.2589> (1977).
- Knight, C. A. Structural biology. Adding to the antifreeze agenda. *Nature* **406**(249), 251. <https://doi.org/10.1038/35018671> (2000).
- DeVries, A. L. & Wohlschlag, D. E. Freezing resistance in some Antarctic fishes. *Science* **163**, 1073–1075. <https://doi.org/10.1126/science.163.3871.1073> (1969).
- Hew, C. L. *et al.* Multiple genes provide the basis for antifreeze protein diversity and dosage in the ocean pout, *Macrozoarces americanus*. *J. Biol. Chem.* **263**, 12049–12055. [https://doi.org/10.1016/s0021-9258\(18\)37891-8](https://doi.org/10.1016/s0021-9258(18)37891-8) (1988).
- Duman, J. G. & Patterson, J. L. Role of thermal-hysteresis-proteins in low-temperature tolerance of insects and spiders. *Cryobiology* **15**, 683–684. [https://doi.org/10.1016/0011-2240\(78\)90106-2](https://doi.org/10.1016/0011-2240(78)90106-2) (1978).

8. Duman, J. G., Horwarth, K. L., Tomchaney, A. & Patterson, J. L. Antifreeze agents of terrestrial arthropods. *Comp. Biochem. Physiol. A Physiol.* **73**, 545–555. [https://doi.org/10.1016/0300-9629\(82\)90261-4](https://doi.org/10.1016/0300-9629(82)90261-4) (1982).
9. Raymond, J. A. The ice-binding proteins of a snow alga, *Chloromonas brevispina*: Probable acquisition by horizontal gene transfer. *Extremophiles* **18**, 987–994. <https://doi.org/10.1007/s00792-014-0668-3> (2014).
10. Vance, T. D. R., Graham, L. A. & Davies, P. L. An ice-binding and tandem beta-sandwich domain-containing protein in *Shewanella frigidimarina* is a potential new type of ice adhesin. *FEBS J.* **285**, 1511–1527. <https://doi.org/10.1111/febs.14424> (2018).
11. Raymond, J. A., Janech, M. G. & Mangiagalli, M. Ice-binding proteins associated with an Antarctic cyanobacterium, *Nostoc* sp. HG1. *Appl. Environ. Microbiol.* <https://doi.org/10.1128/AEM.02499-20> (2021).
12. Near, T. J. *et al.* Ancient climate change, antifreeze, and the evolutionary diversification of Antarctic fishes. *Proc. Natl. Acad. Sci. USA* **109**, 3434–3439. <https://doi.org/10.1073/pnas.1115169109> (2012).
13. Scott, G. K., Fletcher, G. L. & Davies, P. L. Fish antifreeze proteins: Recent gene evolution. *Can. J. Fish. Aquat. Sci.* **43**, 1028–1034. <https://doi.org/10.1139/f86-128> (1986).
14. Davies, P. L. Ice-binding proteins: A remarkable diversity of structures for stopping and starting ice growth. *Trends Biochem. Sci.* **39**, 548–555. <https://doi.org/10.1016/j.tibs.2014.09.005> (2014).
15. Graham, L. A., Hobbs, R. S., Fletcher, G. L. & Davies, P. L. Helical antifreeze proteins have independently evolved in fishes on four occasions. *PLoS One* **8**, e81285. <https://doi.org/10.1371/journal.pone.0081285> (2013).
16. Chen, L., DeVries, A. L. & Cheng, C. H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl. Acad. Sci. USA* **94**, 3811–3816. <https://doi.org/10.1073/pnas.94.8.3811> (1997).
17. Ewart, K. V. & Fletcher, G. L. Herring antifreeze protein: Primary structure and evidence for a C-type lectin evolutionary origin. *Mol. Mar. Biol. Biotechnol.* **2**, 20–27 (1993).
18. Graham, L. A., Lougheed, S. C., Ewart, K. V. & Davies, P. L. Lateral transfer of a lectin-like antifreeze protein gene in fishes. *PLoS One* **3**, 2616. <https://doi.org/10.1371/journal.pone.0002616> (2008).
19. Deng, C., Cheng, C. H., Ye, H., He, X. & Chen, L. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc. Natl. Acad. Sci. USA* **107**, 21593–21598. <https://doi.org/10.1073/pnas.1007883107> (2010).
20. Graham, L. A., Gauthier, S. Y. & Davies, P. L. Origin of an antifreeze protein gene in response to Cenozoic climate change. *Sci. Rep.* **12**, 8536. <https://doi.org/10.1038/s41598-022-12446-4> (2022).
21. Walker, V. K. *et al.* Surviving winter with antifreeze proteins: Studies on budworms and beetles. In *Insect Timing: Circadian Rhythmicity to Seasonality* 199–211 (US, 2001). <https://doi.org/10.1016/B978-0-44450608-5/50048-9>.
22. Graether, S. P. & Sykes, B. D. Cold survival in freeze-intolerant insects: The structure and function of beta-helical antifreeze proteins. *Eur. J. Biochem.* **271**, 3285–3296. <https://doi.org/10.1111/j.1432-1033.2004.04256.x> (2004).
23. Raymond, J. A. & Morgan-Kiss, R. Multiple ice-binding proteins of probable prokaryotic origin in an Antarctic lake alga, *Chlamydomonas* sp. ICE-MDV (Chlorophyceae). *J. Phycol.* **53**, 848–854. <https://doi.org/10.1111/jpy.12550> (2017).
24. Kondo, H. *et al.* Ice-binding site of snow mold fungus antifreeze protein deviates from structural regularity and high conservation. *Proc. Natl. Acad. Sci. USA* **109**, 9360–9365. <https://doi.org/10.1073/pnas.1121607109> (2012).
25. Hopkin, S. P. *Biology of the Springtails: (Insecta: Collembola)* (OUP, 1997).
26. Konopova, B., Kolosov, D. & O'Donnell, M. J. Water and ion transport across the eversible vesicles in the colophore of the springtail *Orchesella cincta*. *J. Exp. Biol.* **222**, 25. <https://doi.org/10.1242/jeb.200691> (2019).
27. Giribet, G. & Edgecombe, G. D. The phylogeny and evolutionary history of arthropods. *Curr. Biol.* **29**, R592–R602. <https://doi.org/10.1016/j.cub.2019.04.057> (2019).
28. Zettel, J. Cold hardiness strategies and thermal hysteresis in Collembola. *Rev. Ecol. Biol. Sol.* **21**, 189–203 (1984).
29. Graham, L. A. & Davies, P. L. Glycine-rich antifreeze proteins from snow fleas. *Science* **310**, 461. <https://doi.org/10.1126/science.1115145> (2005).
30. Hawes, T. C., Marshall, C. J. & Wharton, D. A. A 9 kDa antifreeze protein from the Antarctic springtail, *Gomphiocephalus hodgsoni*. *Cryobiology* **69**, 181–183. <https://doi.org/10.1016/j.cryobiol.2014.07.001> (2014).
31. Graham, L. A., Boddington, M. E., Holmstrup, M. & Davies, P. L. Antifreeze protein complements cryoprotective dehydration in the freeze-avoiding springtail *Megaphragura arctica*. *Sci. Rep.* **10**, 3047. <https://doi.org/10.1038/s41598-020-60060-z> (2020).
32. Holmstrup, M. Screening of cold tolerance in fifteen springtail species. *J. Therm. Biol.* **77**, 1–6. <https://doi.org/10.1016/j.jtherbio.2018.07.017> (2018).
33. Lin, F. H., Graham, L. A., Campbell, R. L. & Davies, P. L. Structural modeling of snow flea antifreeze protein. *Biophys. J.* **92**, 1717–1723. <https://doi.org/10.1529/biophysj.106.093435> (2007).
34. Pentelute, B. L. *et al.* X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers. *J. Am. Chem. Soc.* **130**, 9695–9701. <https://doi.org/10.1021/ja8013538> (2008).
35. Pentelute, B. L., Gates, Z. P., Dashnau, J. L., Vanderkooi, J. M. & Kent, S. B. Mirror image forms of snow flea antifreeze protein prepared by total chemical synthesis have identical antifreeze activities. *J. Am. Chem. Soc.* **130**, 9702–9707. <https://doi.org/10.1021/ja801352j> (2008).
36. Trevino, M. A. *et al.* The singular NMR fingerprint of a polyproline II helical bundle. *J. Am. Chem. Soc.* **140**, 16988–17000. <https://doi.org/10.1021/jacs.8b05261> (2018).
37. Scholl, C. L. & Davies, P. L. Protein engineering of antifreeze proteins reveals that their activity scales with the area of the ice-binding site. *FEBS Lett.* <https://doi.org/10.1002/1873-3468.14552> (2022).
38. Mok, Y. F. *et al.* Structural basis for the superior activity of the large isoform of snow flea antifreeze protein. *Biochemistry* **49**, 2593–2603. <https://doi.org/10.1021/bi901929n> (2010).
39. Scholl, C. L., Tsuda, S., Graham, L. A. & Davies, P. L. Crystal waters on the nine polyproline type II helical bundle springtail antifreeze protein from *Granisotoma rainieri* match the ice lattice. *FEBS J.* **288**, 4332–4347. <https://doi.org/10.1111/febs.15717> (2021).
40. Tomalty, H. E., Graham, L. A., Eves, R., Gruneberg, A. K. & Davies, P. L. Laboratory-scale isolation of insect antifreeze protein for cryobiology. *Biomolecules* **9**, 180. <https://doi.org/10.3390/biom9050180> (2019).
41. Graham, L. A., Liou, Y. C., Walker, V. K. & Davies, P. L. Hyperactive antifreeze protein from beetles. *Nature* **388**, 727–728. <https://doi.org/10.1038/41908> (1997).
42. Burns, G. *et al.* Gene expression associated with changes in cold tolerance levels of the Antarctic springtail, *Cryptopygus antarcticus*. *Insect Mol. Biol.* **19**, 113–120. <https://doi.org/10.1111/j.1365-2583.2009.00953.x> (2010).
43. Purac, J. *et al.* Cold hardening processes in the Antarctic springtail, *Cryptopygus antarcticus*: Clues from a microarray. *J. Insect Physiol.* **54**, 1356–1362. <https://doi.org/10.1016/j.jinsphys.2008.07.012> (2008).
44. Faddeeva-Vakhrusheva, A. *et al.* Coping with living in the soil: The genome of the parthenogenetic springtail *Folsomia candida*. *BMC Genom.* **18**, 493. <https://doi.org/10.1186/s12864-017-3852-x> (2017).
45. Kim, J. S., Monroe, M. E., Camp, D. G. 2nd., Smith, R. D. & Qian, W. J. In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *J. Proteome Res.* **12**, 910–916. <https://doi.org/10.1021/pr300955f> (2013).
46. Shoulders, M. D. & Raines, R. T. Collagen structure and stability. *Annu. Rev. Biochem.* **78**, 929–958. <https://doi.org/10.1146/annurev.biochem.77.032207.120833> (2009).
47. Rappu, P., Salo, A. M., Myllyharju, J. & Heino, J. Role of prolyl hydroxylation in the molecular interactions of collagens. *Essays Biochem.* **63**, 325–335. <https://doi.org/10.1042/EBC20180053> (2019).
48. Christian, E. Induction and detection of moulting synchronization in *Folsomia candida* laboratory populations (Collembola: Isotomidae). *Rev. Ecol. Biol. Sol.* **25**, 469–478 (1988).

49. Beliveau, C. *et al.* The spruce budworm genome: Reconstructing the evolutionary history of antifreeze proteins. *Genome Biol. Evol.* **14**, 25. <https://doi.org/10.1093/gbe/evac087> (2022).
50. Meier, P. & Zettel, J. Cold hardiness in *Entomobrya nivalis* (Collembola, Entomobryidae): Annual cycle of polyols and antifreeze proteins, and antifreeze triggering by temperature and photoperiod. *J. Comp. Physiol. B* **167**, 297–304. <https://doi.org/10.1007/s003600050077> (1997).
51. Scalzitti, N. *et al.* Spliceator: Multi-species splice site prediction using convolutional neural networks. *BMC Bioinform.* **22**, 561. <https://doi.org/10.1186/s12859-021-04471-3> (2021).
52. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025. <https://doi.org/10.1038/s41587-021-01156-3> (2022).
53. Hawes, T. C., Marshall, C. J. & Wharton, D. A. Antifreeze proteins in the Antarctic springtail, *Gressittacantha terranova*. *J. Comp. Physiol. B* **181**, 713–719. <https://doi.org/10.1007/s00360-011-0564-4> (2011).
54. Xiong, Y., Gao, Y., Yin, W. Y. & Luan, Y. X. Molecular phylogeny of Collembola inferred from ribosomal RNA genes. *Mol. Phylogenet. Evol.* **49**, 728–735. <https://doi.org/10.1016/j.ympev.2008.09.007> (2008).
55. Schneider, C. Unexpected diversity in Neelipleona revealed by molecular phylogeny approach (Hexapoda, Collembola). *Soil Organ.* **83**, 383–398 (2011).
56. Leo, C., Carapelli, A., Cicconardi, F., Frati, F. & Nardi, F. Mitochondrial genome diversity in Collembola: Phylogeny, dating and gene order. *Diversity-Basel* **11**, 169. <https://doi.org/10.3390/d11090169> (2019).
57. Caputo, M. V. & Crowell, J. C. Migration of glacial centers across Gondwana during Paleozoic Era. *Geol. Soc. Am. Bull.* **96**, 25. [https://doi.org/10.1130/0016-7606\(1985\)96%3c1020:Mogcag%3e2.0.Co;2](https://doi.org/10.1130/0016-7606(1985)96%3c1020:Mogcag%3e2.0.Co;2) (1985).
58. Fielding, C. R., Frank, T. D. & Isbell, J. L. In *Special Paper 441: Resolving the Late Paleozoic Ice Age in Time and Space* 343–354 (2008).
59. Ding, Y. H., Yu, D. Y., Guo, W. B., Li, J. N. & Zhang, F. Molecular phylogeny of Entomobrya (Collembola: Entomobryidae) from China: Color pattern groups and multiple origins. *Insect Sci.* **26**, 587–597. <https://doi.org/10.1111/1744-7917.12559> (2019).
60. McInerney, F. A. & Wing, S. L. The Paleocene–Eocene Thermal Maximum: A perturbation of carbon cycle, climate, and biosphere with implications for the future. *Annu. Rev. Earth Planet Sci.* **39**, 489–516. <https://doi.org/10.1146/annurev-earth-040610-133431> (2011).
61. Graether, S. P. *et al.* Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. *Nature* **406**, 325–328. <https://doi.org/10.1038/35018610> (2000).
62. Hakim, A. *et al.* Crystal structure of an insect antifreeze protein and its implications for ice binding. *J. Biol. Chem.* **288**, 12295–12304. <https://doi.org/10.1074/jbc.M113.450973> (2013).
63. Yang, D. S., Sax, M., Chakrabarty, A. & Hew, C. L. Crystal structure of an antifreeze polypeptide and its mechanistic implications. *Nature* **333**, 232–237. <https://doi.org/10.1038/333232a0> (1988).
64. Liu, Y. *et al.* Structure and evolutionary origin of Ca<sup>2+</sup>-dependent herring type II antifreeze protein. *PLoS One* **2**, e548. <https://doi.org/10.1371/journal.pone.0000548> (2007).
65. Friis, D. S., Kristiansen, E., von Solms, N. & Ramlov, H. Antifreeze activity enhancement by site directed mutagenesis on an antifreeze protein from the beetle *Rhagium mordax*. *FEBS Lett.* **588**, 1767–1772. <https://doi.org/10.1016/j.febslet.2014.03.032> (2014).

## Acknowledgements

We thank David McLeod of the Protein Function Discovery facility at Queen's University for MALDI analysis of AFPs. We are grateful to Marie Boddington for preliminary analysis of *Megaphorura arctica* and *Folsomia candida* AFPs, to David Denlinger for the collection of *Cryptopygus antarcticus*, and to Matty Berg for the collection of *Isotoma riparia*. This work was supported by CIHR Foundation Grant FRN 148422 to PLD and Det Frie Forskningsråd Grant 1026-00055B to MH. PLD holds the Canada Research Chair in Protein Engineering. The funders had no role in study design, data collection/analysis/ interpretation, report writing, or submission of the article for publication.

## Author contributions

Conceptualization, P.L.D., M.H., and L.A.G.; methodology, all authors; investigation, C.L.S. and L.A.G.; resources, P.L.D. and M.H.; writing—original, C.L.S., P.L.D., and L.A.G.; writing—review and editing, all authors; visualization, C.L.S.; funding acquisition, P.L.D. and M.H.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-35983-y>.

**Correspondence** and requests for materials should be addressed to P.L.D.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023