Collagen sequence analysis of fossil camels, Camelops and c.f. Paracamelus, from the Arctic and sub-Arctic of Plio-Pleistocene North America

DOI:
10.1016/j.jprot.2018.11.014

Document Version
Accepted author manuscript

Collagen Sequence Analysis of fossil camels, *Camelops* and c.f. *Paracamelus*, from the Arctic and sub-Arctic of Plio-Pleistocene North America

M Buckley[1*], C Lawless[2], N Rybczynski[3]

[1]School of Earth and Environmental Sciences, Manchester Institute of Biotechnology, University of Manchester, Greater Manchester, M1 7DN, UK.

[2]School of Biological Sciences, Michael Smith Building, University of Manchester, M13 9PN, UK.

[3]Department of Biology & Department of Earth Sciences, Carleton University, 1125 Colonel By Dr, Ottawa, ON K1S 5B6, Canada

***Corresponding email: m.buckley@manchester.ac.uk**

Abstract

Proteomic analyses of ancient remains are increasing in number and offer great potential to recover phylogenetic information on extinct animals beyond the reach of ancient DNA, but limitations in proteomic techniques remain unclear. Here we carry out LC-MS/MS sequence analysis of a ~3.5 million year old giant camel specimen from Nunavut along with the younger Pleistocene remains of the Yukon giant camel (c.f. *Paracamelus*) and the western camel (*Camelops hesternus*) for comparison with complete sequences to both extant camels (Bactrian and Dromedary) and the alpaca. Although not complete (~75–80% sequence coverage), no amino acid sequence differences were confidently observed between the giant camels and the extant Dromedary, indicative of a closer relationship than that of the extant Bactrian lineage. However, multiple amino acid changes were observed for the western camel (*Camelops*) collagen sequence, placing it as a sister group to these members of the Camelini tribe consistent recent ancient DNA analyses. Although this supports a role for the sequencing of ancient collagen in the understanding of vertebrate evolution, these analyses highlight the limitations in phylogenetic reconstructions based on partial sequence data retrieved from proteomic analyses, particularly, the impact of omitting even only a single peptide on the resulting tree topology. The presence of other non-collagenous proteins, such as biglycan and PEDF, indicates a further resource for phylogenetic information, but none more promising than the degraded camel albumin seemingly observed in the Pliocene specimen.

Keywords: Paleoproteomics, Giant camels, Ancient Collagen, Arctic camel, Paracamelus, Camelops, Ancient Albumin.

Introduction

Camelidae, the artiodactyl Family that includes llamas and camels, originated in North America during the Eocene period (~45 Ma) and diversified in the early Miocene (~23-16 Ma) giving rise to at least 20 genera [1]. These are primarily separated into two modern tribes, Aucheniini and Camelini, which diverged from one another approximately 17 million years ago [2]. Initially, members of the Camelini dispersed into Eurasia and then Africa by the Late Miocene, whereas the Aucheniini later dispersed to South America by the late Pliocene, approximately three million years ago [3]. Among the last surviving camels in North America was the western camel Camelops, the largest of the late Pleistocene camels becoming extinct ~13,000 years ago after being notably hunted by early humans [4]. Interestingly, morphological studies suggested *Camelops* as being a highly derived member of the Aucheniini [1,5], whereas more recent ancient DNA (aDNA) analysis of three Camelops sub-fossils from the Yukon indicates a closer affinity with Camelini [6].

North American fossil Camelini include *Procamelus*, which was roughly the size of a modern llama, and the "giant" forms, *Titanotylopus*, *Megatylopus*, *Megacamelus* and *Gigantocamelus* [1]. Beginning in the late Miocene, the Eurasian fossil record for camels is considered to preserve only two genera, *Camelus* and *Paracamelus*, with the latter being the most likely ancestor to *Camelus* [7]. Camelus has been recovered from Asia, Europe and Africa, and the oldest-known representative member is from the late Miocene of Spain (Vento del Moro; ca 7.5–6.5 Ma) [7]. In North America, the Yukon giant camel, a rare camel from Late Pleistocene deposits of the Old Crow Basin located just north of the Arctic circle, is considered likely to be *Paracamelus* (cf. *Paracamelus gigas*) [8]. The first specimen of the giant Yukon camel, a proximal phalanx, was discovered in 1913 [9] but numerous finds from the famous Old Crow basin have subsequently been recovered. The North American lineage most closely related to *Paracamelus* and *Camelus*, is debated, with possible candidates including *Megacamelus*, *Procamelus*, or *Megatylopus* [7]. Notably all of these taxa are known from localities that are south of 55°N. Other northern fossil camels include the western camel, *Camelops hesternus*, recovered from Late Pleistocene deposits in the southern half of the Yukon [8], and recently ascribed to Camelini based on aDNA [6]. The most northern fossil camel known, identified with the help of collagen peptide mass fingerprinting, comes from directly-dated Pliocene deposits located 78°N on Ellesmere Island, Nunavut [10]. The Nunavut camel, known only from fragmentary preservation of a partial limb bone, offers limited opportunity for comparative research based on morphology. However, the exceptional biochemical preservation offers the option to explore its affinities with other northern camels using ancient proteins such as collagen.

Collagen, the dominant protein in bone has been shown to survive longer than other genetically informative biomolecules, including other non-collagenous proteins (NCPs) such as osteocalcin as well as DNA [11] and has been recently shown to be sufficiently variable between genera to be useful taxonomically [12]. Although proteomic evidence for the presence of collagen in samples ranging throughout the Pleistocene period is relatively common, even in temperate sites well over one million years old [13,14], reports of older protein sequences are much rarer. The highly sensitive approach of using LC-tandem mass

spectrometry for the in-depth sequencing of ancient proteins was reported to yield collagen peptides from fossilized Mesozoic remains [15,16], but these were initially met with suspicion [17,18] with more recent evidence suggesting that sample cross-contamination cannot be ruled out [19]. However, despite the exact longevity of ancient proteins, the presence of collagen in subfossil Pleistocene material has long been accepted [20] and has been exploited for reconstructing ecological relationships using stable isotopes [21] in addition to determining the phylogenetic relationships of extinct taxa [22,23].

The goal of our initial study of Yukon and Nunavut sub-fossil giant camels was to help identify the Nunavut camel, known only from a fragmentary, partial tibia [10]. In the original study, samples of the Nunavut fossil, were compared with fossil samples of the extinct Yukon camels, c.f. *Paracamelus*, along with a database of 37 modern mammals, including both extant camels, and llama [10]. Sequencing of selected biomarkers was obtained by MALDI-ToF/ToF mass spectrometry, and the resulting collagen fingerprints for the Nunavut and two sub-Arctic c.f. Paracamelus specimens were found to most closely match the dromedary camel (*Camelus dromedarius*) with only a few distinct peaks present that could not be assigned sequence information [10]. Interestingly, the same analyses found the Camelops samples to be most similar to the Aucheniini group (in terms of the typical species biomarkers used), as expected from the morphological evidence [1] but counter to the more recently published aDNA results [6]. Here we provide in-depth proteomic sequencing of the collagen recovered from these specimens in order to assess the phylogenetic affinities of the Nunavut camel relative to the Yukon c.f. *Paracamelus* and *Camelops*.

Materials and Methods

The palaeontological materials are predominantly those described previously by Rybczynski et al. [10], including the ~3.5 Ma giant camel from the Fyles Leaf Beds site (Ellesmere Island, Nunavut). The Fyles Leaf Beds site refers to a ~1 km-long natural steep exposure of more than 90 m of Beaufort-equivalent high terrace deposits, located in the vicinity of Strathcona Fiord, on west central Ellesmere Island (N78° 30′ W82° 38′) (Nunavut) and yielded ~30 fragments of a large right tibia. The remains were recovered from a steeply sloping (~45°) colluvium surface which extended vertically over 12 m, from a point source located in the upper levels of the section.

Also sampled were Yukon c.f. *Paracamelus* and *Camelops* material attributed to the Late Pleistocene [8]. *Camelops hesternus*: CMNFV 42390 – a left metatarsal of subadult (lacks distal epiphysis) from Sixtymile Area Loc. 3 (P8203), Sixtymile River, Yukon Territory, Canada; CMNFV 46728 - a right radius (distal end) from Sixtymile Area Loc. 3 (P8203), Sixtymile River, Yukon Territory, Canada; the Yukon camels (c.f. *Paracamelus*): CMN 27266 – a proximal phalanx from the Old Crow River Locality 29, Yukon, Canada; CMN 48096 – a distal fragment of metapodial from the Old Crow River Locality 11A, Yukon, Canada. All analyses presented here were carried out on the protein digests already extracted from the specimens described in the 2013 study, including those of the Nunavut specimen, for which powder was drilled from several millimeters below the original surface and collected onto fresh foil.

Collagen Sequence Analysis by LC-MS/MS

Collagen was originally extracted from the modern and fossil samples as described by Rybczynski et al. [10] by agitation in 0.6 M hydrochloric acid (HCl) for 4 h and then subsequently exchanged into 50 mM ammonium bicarbonate (ABC; Sigma-Aldrich, UK) using ultrafiltration (30 kDa molecular weight cut-off (MWCO; Vivaspin, UK)). Additionally, an acid-insoluble component was also extracted (initially only for the Nunavut specimen, but in later runs for all specimens analysed) via the addition of 6 M GuHCl (Sigma-Aldrich, UK) overnight prior to ultrafiltration into ABC as above). From these fractions the proteins were then digested by sequencing grade trypsin (37^{o}C for 18 h; Promega, UK), acidified to 0.1% trifluoroacetic acid (TFA; Sigma-Aldrich, UK) and concentrated using C18 ZipTips following Buckley et al. [24]. Peptide Mass Fingerprints (PMFs) were then obtained by MALDI-ToF-MS and MS/MS spectra acquired for selected peptides as published elsewhere [10]. However, these digested samples were then analysed by an LC-MS/MS (Waters nanoAcquity Ultra Performance Liquid Chromatography (UPLC) instrument coupled to a Thermo Scientific Linear Trap Quadrupole (LTQ) Velos Dual Pressure mass spectrometer) on which the peptides were concentrated on a pre-column (20 mm x 6180 mm) and then separated on a 1.7 µm Waters nanoAcquity BEH (Ethylene Bridged Hybrid) C18 analytical column (75 mm x 250 µm i.d.), using a gradient from 99% buffer A (0.1% formic acid (FA; Sigma-Aldrich, UK) in H_2O)/1% buffer B (0.1% FA in acetonitrile (ACN)) to 25% B in 45 min at 200 nL/min with peptides selected for fragmentation automatically by data dependent analysis (DDA); dynamic exclusion involved the exclusion for 15 s after one occurrence.

Proteomics data files were searched using Mascot against a local database that included Bactrian camel, Dromedary camel and alpaca collagen sequences obtained from protein BLAST searches of cattle (*Bos taurus*) COL1A1 and COL1A2 sequences. Error Tolerant searches with up to two missed cleavages, error tolerances of 0.5 Da (MS and MS/MS) and the oxidation of Proline (P) and Lysine (K) included as variable modifications (along with carbamidomethyl C as a fixed modification); this form of Error Tolerant search allows the algorithm to search the UniMod database for further modifications that could occur as a result of amino acid substitution or amino acid modification during diagenesis. Following revision of the sequences from these results, the searches were carried out as normal searches but with oxidation (P, K and Methionine (M)) and deamidation of asparagine (N) and glutamine (Q) modifications (see Supplementary Table S1 for ordered list of analyses). LC-MS collagen peptide sequence matches from the revised sequences were selected with a peptide ion score greater than that determined by Mascot as the threshold for homology (~16 in these analyses when searched against the local database alone; ~51 when searched against this as well as SwissProt) and are listed with their respective ion scores in the Supplementary Information (Supplementary Tables S2-7); searches using the semi-trypsin parameter were also carried out on the c.f. Paracamelus specimens, including the Nunavut sub-fossil, in order to evaluate levels of degradation (Supplementary Tables S8-10). In order to evaluate carry-over, as well as to include analysis of extraction blanks, two further sets of analyses were carried out, one with and one without reduction and alkylation. In the third set of analyses, the 6 protein standard that is normally run at the start and end of each

batch was this time also ran prior to every set of blanks, in between every pair of samples (see Supplementary Table S1). This was carried out just prior to the tryptic digest by addition of 100 mM dithiothreitol (DTT; Sigma-Aldrich, UK; 2.1 µL per 50 µL sample) for at 60 ^{o}C for 10 minutes, followed by 100 mM iodoacetamide (Sigma-Aldrich, UK; 4.2 µL per 50 µL sample) for a further 45 minutes at room temperature in the dark, before finally being quenched by the same amount of DTT as noted above prior to digestion.

In order to avoid potential issues with column carry-over, we ensured that all three taxa of interest (the Nunavut giant camel, the Yukon camel and Camelops) were placed first in different analytical runs (see Supplementary Table S1). However, these additional runs (Supplementary Tables S11-42 inclusive of blanks) were both carried out using an UltiMate 3000 Rapid Separation LC (RSLC, Dionex Corporation, Sunnyvale, CA, USA) coupled to an Orbitrap Elite (Thermo Fisher Scientific, Waltham, MA, USA) mass spectrometer (120k resolution, Full Scan, Positive mode, normal mass range 350–1500). Using the same type of column, a gradient was applied from 92% A (0.1% formic acid in water) and 8% B (0.1% formic acid in acetonitrile) to 33% B in 44 min at a flow rate of 300 nl min^{-1}. Peptides were then automatically selected for fragmentation by data-dependent analysis; 6 MS/MS scans (Velos ion trap, product ion scans, rapid scan rate, Centroid data; scan event: 500 count minimum signal threshold, top six) were acquired per cycle, dynamic exclusion was also employed and one repeat scan (2 MS/MS scans total) was acquired in a 30 s repeat duration with that precursor being excluded for the subsequent 30 s (activation: CID, 2 + default charge state, 2 *m*/*z* isolation width, 35 eV normalized collision energy, 0.25 Activation Q, 10.0 ms activation time). For the additional sets of analyses (Runs 2 and 3), much larger (~5-10 times as much) amounts of starting material was used for each sample, typically ~300 mg.

Phylogenetic analyses of the concatenated collagen alpha 1 and alpha 2 sequences (via an R residue; yielding a total length of 2098 amino acid residues) of the *Paracamelus*, 'Nunavut' and *Camelops* partial sequences were carried out using the PhyML plugin for MEGA [25] version X with 22 other mammalian type 1 collagen sequences (with and without concatenation of these two chains) obtained from protein BLAST that include three extant camelids (sequence sources are listed in Supplementary Table S43 and those for non-collagenous proteins in Supplementary Table S44; all sequences available in FASTA format in Supplementary Table S45). Following Buckley *et al.* [24], the JTT + I + G model was used and trees were rooted to the African elephant as an Afrotherian out-group. 500 bootstraps were carried out to estimate support. Bayesian analyses were also carried out using the MrBayes 3.2.2 [26] with 10,000,000 MCMC generations, discarding the first 2,500,000 as burn-in, estimated invariable gamma distribution (4 categories), 4 chains (3 heated, 1 cold) with unconstrained branch lengths and also rooted to the African elephant (*Loxodonta*).

Results

Variation amongst known camelid collagen sequences

Complete sequences were retrieved for both extant species of the Camelini, the dromedary camel and the Bactrian camel (with no amino acid substitutions observed between the Bactrian camel *Camelus bactrianus* and its wild ancestor *Camelus ferus*) as well as one

species from the Aucheniini, the alpaca (*Vicugna pacos*) in addition to the NCPs most frequently observed in proteomic analyses of ancient bone [14]. With searching against at least 22 other NCPs (Supplementary Table S44) in addition to collagen (I), collagen was the only protein matched with multiple unique peptides in every sample and therefore forms the sole focus of this study. From the over 2,000 amino acids that form the two collagen chains, only one substitution can be observed between the dromedary and Bactrian camel's COL1A1 sequence, and only a further three substitutions between their COL1A2 sequences. By comparison there are at least 12 substitutions between the tribes Aucheniini and Camelini (although one of these peptides is so short that it is only observed as part of a missed cleavage, but combined with a neighbouring variation-containing peptide possesses two such substitutions). Only one of these 12 is observed on the COL1A1 chain (different between the alpaca and dromedary, but not between the alpaca and Bactrian camel), and eight are found as separating the two groups uniformly (the remaining three are all shared between the alpaca and dromedary, each distinct from the Bactrian camel; Table 1).

Species	1t1	2t2	2t6-8	2t23	2t31	2t33	2t37-38	2t43	2t67	2t77	2t82
V. pacos	**M**	**S**	**P**	**S**	**N**	**I**	**A-H**	**P**	**P**	**I**	**P**
C. dromedaries	**L**	**G**	**A**	**S**	**S**	**V**	**T-N**	**P**	**A**	**I**	**S**
C. bactrianus	**M**	**G**	**A**	**G**	**S**	**V**	**T-N**	**S**	**A**	**T**	**S**
***Yukon giant camel* 27266**	**X**	**G**	**X**	**S***	**S***	**V**	**T-N**	**P**	**A**	**I**	**S**
***Camelops* 42390**	**X**	**G**	**X**	**S***	**S***	**V**	**T-N**	**P**	**P**	**I**	**S**
***Camelops* 46728**	**X**	**G**	**X**	**S***	**S***	**V**	**T-N**	**P**	**P**	**I**	**S**
***Yukon giant camel* 48091**	**X**	**G**	**X**	**S***	**S***	**V**	**T-N**	**P**	**A**	**I**	**S**
Nunavut giant camel	**X**	**G**	**X**	**S***	**S***	**V**	**T-N**	**P**	**A**	**I**	**S**

Table 1 – Sequence variations known between extant camel collagen sequences, and those observed in the fossils (x indicates not observed, *only observed from missed cleavage). Sequence labels following [27].

Camelops surprisingly yielded nine of the ten observed states as identical to the dromedary camel, with the only unique match to the alpaca as deriving from the one peptide biomarker (2t67 at *m/z* 3033.4) described from the collagen peptide fingerprint analysis [10](Fig. 1), being alongside the three substitutions in common between dromedary and alpaca. No differences were observed between the Yukon or Nunavut giant camel, nor were any further substitutions discovered through our Error Tolerant searches.

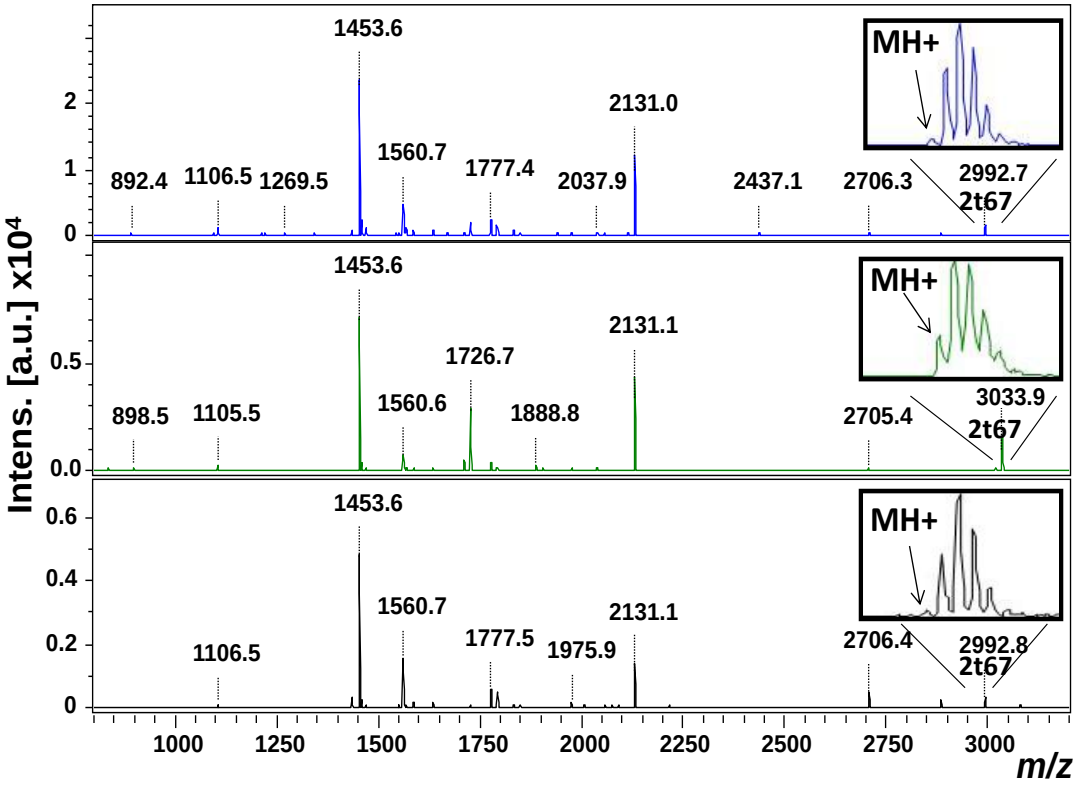

Figure 1 - MALDI-ToF mass spectra of collagen peptide mass fingerprints from the Yukon giant camel (top) and *Camelops* (middle) in comparison to the Nunavut giant camel (bottom; note high levels of deamidation in peptide 2t67 for the giant camels (inset) proposed as an indicator of relative age (see Buckley et al. [28])).

Proteomic Sequences and Phylogenetic Analysis

It is clear from both ML and Bayesian analyses that the collagen sequence analysis supports the previously published aDNA results [6] placing *Camelops* as a sister group to the extant and sub-fossil camels, within Camelini (Fig. 2). However, in order to explore the limitations in paleoproteomics further we carried out analyses in which one or more of the observed peptides were consecutively removed (e.g., replaced with question marks). In this case, even with the removal of a single peptide, we observed minor changes in topology with the loss of any one of the peptides exhibiting variation, whereby the two *Camelops* specimens would form a grade of taxa sister to the camels, rather than as a monophyletic grouping themselves. Despite this, the observations supporting the placement of *Camelops* remained supported. However, it is worthy of note that the mere omission of even only a single peptide can result in substantial changes to the topology; in this case the omission of the 2t67 peptide from *Camelops* moving it from being a sister clade to the extant camels and c.f. *Paracamelus* to being within this clade (Fig. 3B). This has significant implications for ensuring the completeness of a palaeoproteomic sequence. We also evaluated the phylogenetic signal produced with the COL1A1 and COL1A2 sequences separately from their typically concatenated form. With Bayesian analyses (Supplementary Figure S3), suids formed a sister group to the camelids with COL1A1, whereas with both the COL1A2 and concatenated sequences the camelids were basal to the suids and remaining cetartiodactyls. With Maximum Likelihood analysis (Supplementary Figure S4), although the COL1A1 topology

was the same as for Bayesian, the COL1A2 topology placed Carnivora as sister to the Perissodactyla; however, the topology of the concatenated form appeared consistent with that of the Bayesian analysis (i.e., with Perissodactyla as sister grouping to Cetartiodactyla) with suids as sister to the remaining cetartiodactyls to the exclusion of the camelids (note that this latter relationship is also compromised when fewer cetartiodactyl sequences are used, e.g., Supplementary Figure S3).

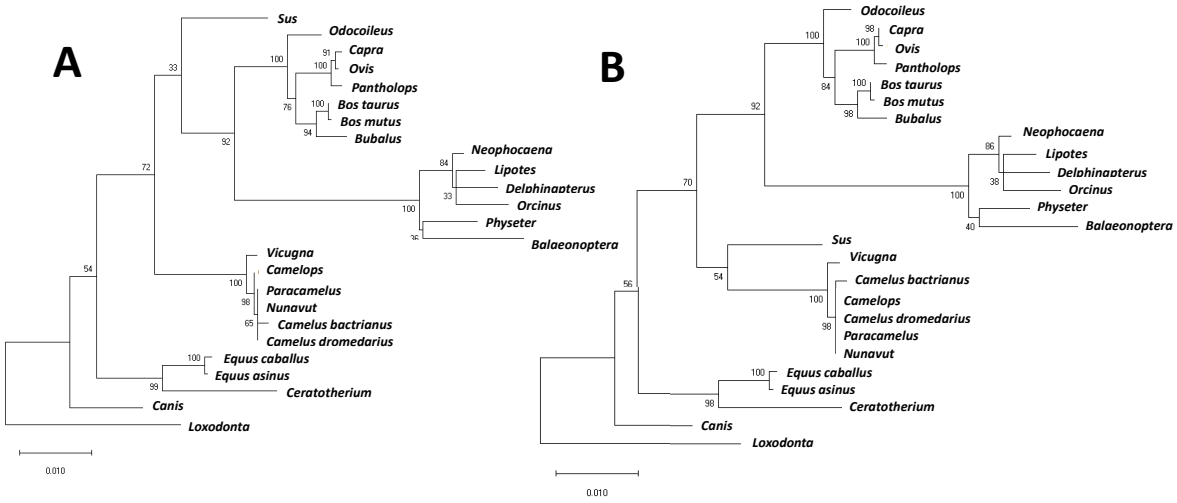

Figure 2 - Maximum Likelihood analysis of *Camelops* and giant camels (*c.f. Paracamelus*) with other ruminant taxa rooted to the African elephant (*Loxodonta*) for (A) all matched ancient sequences and (B) to the exclusion of the peptide GPSGEPGTAGPPGTPGPQGLLGAPGFLGLPGSR (2t67; residues 1826-1858).

Post-translation modifications in ancient collagens

Deamidations of asparagine and glutamine residues are well known post-translational modifications (PTMs) that occur within the laboratory environment, but thought to be a potentially useful measure of relative age in ancient proteins [29, 30]; oxidations of methionine residues also occur spontaneously albeit there are much fewer available positions in collagen (n = 8+4 (12) compared with 12+22 (34) for N and 30+23 (53) for Q). It was also recently reported that the oxidation of proline residues which, alongside that of lysine residues, are a naturally occurring PTM in collagen, could exhibit increase with geological age [31] though this is not evident here. Of the less commonly reported PTMs, we observed the conversion of arginine residues into ornithine (-42.022), the carboxylation (+44.026), galactosylation (+178.048; see [32]) and in some cases myristoylation (+210.198) of lysine residues [33] and the formation of hydroxamic acids (+15.011) discovered through Error Tolerant searches against the *Camelus dromedarius* sequence. Although several of the peptides with modified R or K residues were no longer observed under traditional searches, this affected only a relatively small number of peptides and these were the smaller peptides with least phylogenetic information (e.g., Fig. 3).

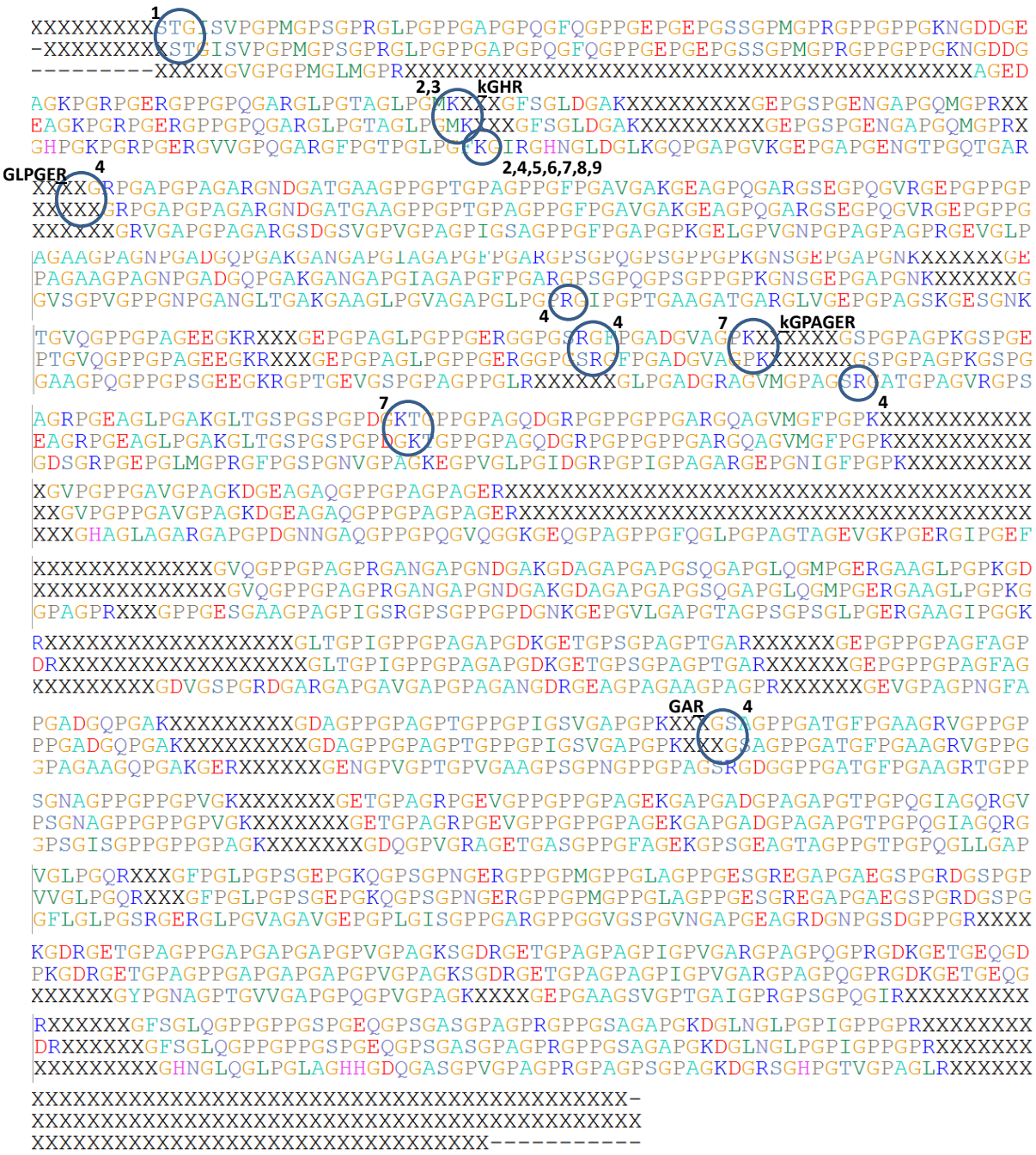

Figure 3 - Sequences obtained from the Nunavut giant camel aligned following [34] showing several of the less common PTMs (excluding oxidations and deamidations; sites marked with open circles): 1) phosphorylation (T+56), 2) glucosylation (K+178), 3) myristoylation (K+210), 4) ornithine formation (R-42), 5) allysine (K-1), 6) formylation (K+28), 7) carbamylation (K+43), 8) carboxylation (K+44), 9) glucosylgalactosylation (+340); sequences present indicate additional observations through Error Tolerant search only.

Although collagen was by far the most dominant protein, and only protein matched in every sample, there were peptides from several NCPs matched in some of the samples across the three analytical runs that were consistent with camelid origins. These included lumican, chondroadherin and PEDF in run 1 in both Yukon samples, albumin and biglycan in run 2 in the Nunavut specimen, and both PEDF and biglycan in the *Camelops* of run 3. However,

although camel albumin was observed in the Nunavut material, making this the oldest reported proteome at ~3.5 Ma, there was also bovine fetuin present, indicating the need for caution with regards understanding potential contamination of this protein in our analyses.

With closer inspection in relation to potential contamination, of the five distinct albumin peptides observed in both acid-soluble and acid-insoluble fractions, the first (FVAFVDK) differs from bovine via a preceding amino acid variation leading to a tryptic site change, the second (YLYEIAR) is the same in *Bos*, the third (LPQVSTPTLVEVAR) has two amino acid substitutions and is also observed with an additional missed cleavage, the fourth (LGEYGFQNDILVR) has three but two of which are isobaric, and the fifth (EACFTVEGPLLVAATR) is highly distinct from the bovine equivalent involving five substitutions including one causing a tryptic site change. Most importantly, in the only peptides that are amenable to deamidation (as well as the missed cleavage variant), only deamidated peptides were matched – no non-deamidated forms were observed. Interestingly, the acid-insoluble fraction yielded a greater number of albumin peptides (n=12 distinct peptides, excluding missed cleavages). Of the extra seven, three peptides (AACLLPK, LVNEVTEFAK (with deamidated N), and HPEYAVSLLLR) have one amino acid variation from *Bos* (the latter also one from vicugna), one (MSCAEDYLSLILNR) has two and another (DVFLGMFLHEYAR) has four substitutions. Therefore with at least 18 observable amino acid substitutions specific to the camelids, and only deamidated forms observed, we have a high level of confidence in this attribution. By contrast the PEDF, biglycan and vimentin had very few uniquely camelid peptides to make them of much potential use in phylogenetic studies. Nevertheless, further work is needed to determine whether use of deamidation can adequately rule out cross-contamination from modern samples.

Discussion and Conclusions

The Role of Paleoproteomics in Recovering Phylogenetic Relationships

There are two direct impacts that these results from collagen sequencing have on our understanding of the evolution of the three extinct camels included in this study. One is through supporting the very close relationship between the Yukon and Nunavut giant camels. The other is the support for the DNA-based placement of *Camelops* within Camelini rather than the morphology-based placement within Aucheniini [6]. However, more fundamental to this is the understanding of our limits in phylogenetic interpretation through the use of partial sequence data typically observed through such proteomic studies. It has been claimed that the highly incomplete nature of collagen sequences recovered from dinosaur remains [15, 16] are predominantly due to this blockage of the residues specific to the enzyme being used. Our study does support the notion that some peptide sequences would no longer be observed by standard searches (four short peptides of six or fewer amino acids each; Fig. 3), but still readily resolvable with existing search algorithms such as the Error Tolerant tool used here. However, an approach to overcome such issues altogether would be the use of a different enzyme, not specific to R or K.

One of the remaining obstacles in palaeoproteomic research is the ability to adequately check the quality of interpretation. We have previously relied on the use of only peptides also observed in MALDI analysis as confident enough for use in phylogeny reconstruction. However, the results presented here clearly indicate that for closely related taxa this would result in the loss of too much sequence information. Other criteria in addition to ion score, such as peptide ion count, could be implemented to increase confidence, particularly in the identification of novel peptide sequences. If the aims are to simply identify the closest taxonomic group to the query sample, such as the case with the South American notoungulates [23], this could be attempted through interpretation of the total protein scores but may not always be appropriate. For example in this case, the *Vicugna* collagen has a slightly higher score than that of the *Camelus dromedarius* for one of the *Camelops* samples.

There are other issues in cross-species proteomics that are even more problematic in palaeontological specimens because the range of possible taxa includes species that could be much more distantly related than those for modern. However, this is further complicated by the similar of mass shifts between different PTMs and amino acid substitutions. The transition of alanine to serine being masked by the hydroxylation of a nearby proline is an obvious issue, but this study has noticed issues with apparent changes of proline to asparagine potentially due to the conversion of lysine into allysine that further emphasises the importance of investigating the tandem spectra closely.

Potential Contamination and the Measuring of Endogeneity through PTMS

In proteomic analyses of fossil specimens, some degree of contamination should always be assumed and it is the role of the analyst to express the level of confidence in their observations being not derived from contamination. The four most logical points at which contamination could be introduced are: 1) in the field (i.e., on site either naturally or during excavations), 2) through post-excavation handling, 3) in the laboratory and/or reagents used during laboratory processes, and 4) within the instrumentation. Although the biomolecular scientist rarely has much influence on either (1) or (2), the second two ((3) and (4)) can to some extent be monitored through the analysis of various blanks, the extent to which varies between laboratories for several reasons. In our case, the proteomics core facility would typically only include a pre- and post-run standard (run 1), made up of a protein mixture. However, additional analyses (runs 2 and 3) clearly identify issues with column carry-over that need to be monitored to avoid misleading results. However, the within-instrument contamination (4) is more challenging to monitor because peptides from a previous analytical run might not only remain within the system due to prior overloading (and, as shown here, for several subsequent analyses), but also due to interactions with either tubing or active surfaces within the instrument. Therefore, even when ‘blank’ samples are devoid of proteins of interest, this does not rule out the potential release of peptides from previous analytical runs; therefore the most appropriate approach to assess endogeneity of peptides is through their natural decay phenomena.

Although the most abundant PTM observed was clearly that of the deamidation of both asparagine and glutamine residues, as has been reported frequently before in

archaeological bone (e.g., [28, 29]) and exclusively observed on the albumin observed within this study, here we also see several examples of other rarer PTMs in the Pliocene protein including the conversion of arginine residues into ornithine (Fig. 4A), the carboxylation, galactosylation (and glucosylgalactosylation; Fig. 4B) and in some cases myristoylation of lysine residues and the formation of hydroxamic acids. However, these appear too rare to reproducibly utilise as a means of measuring decay whereas deamidations are relatively ubiquitous. Interestingly, through comparison of the ratios of deamidated peptides to non-deamidated forms there appears to be little relationship with geological age, at least beyond that of the Pleistocene specimens studied here (Table 2), which may have already reached saturation point (e.g., Fig. 1). Yet of some promise is the apparent increase in the relative amounts of peptide hydrolysis measured via the number of semi-tryptic peptides (Table 2) between the Middle Pleistocene specimens and the Pliocene giant camel.

Table 2 - Measure of the relative number of semi-tryptic peptides as well as the relative number of deamidated peptides in the giant camel specimens.

	Yukon27266	Yukon48091	Nunavut
No. SemiTryptic peptides	277	287	402
No. Tryptic peptides	128	124	147
Hydrolysis Ratio	0.46	0.43	0.37
No. Deamidations	1527	1692	2171
No. Non-deamidated	542	562	775
Deamidation Ratio	0.354944335	0.3321513	0.356978

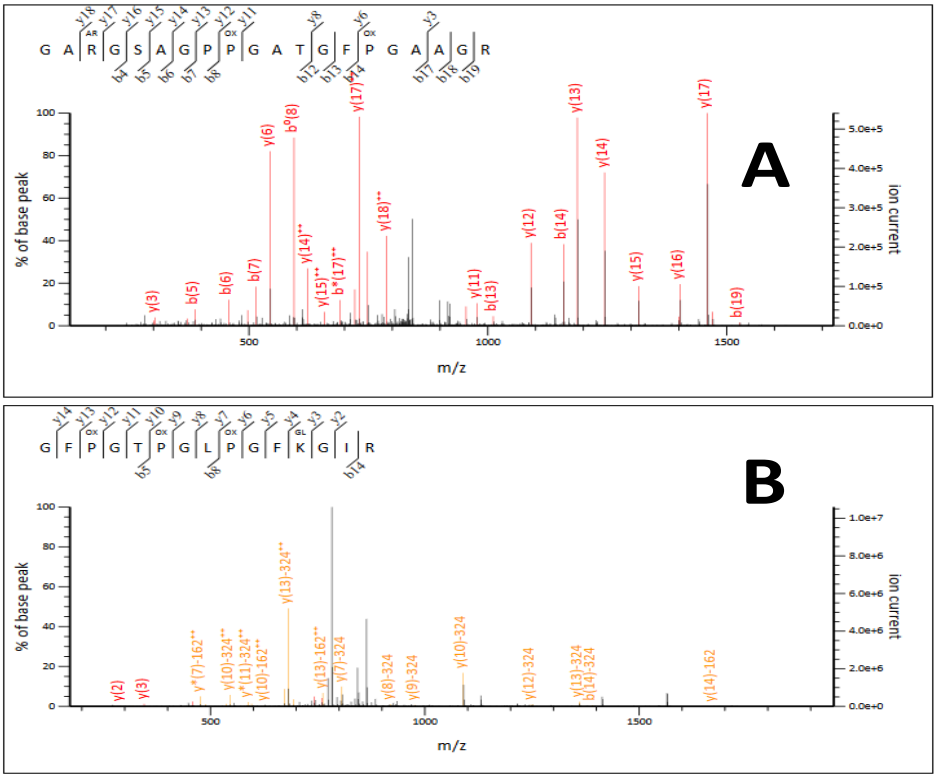

Figure 4 - Example tandem mass spectra of less well-reported PTMs, in this case from the Nunavut giant camel showing A) conversion of an arginine into ornithine, and B) the glucosylgalactosylation of a lysine residue.

The Evolutionary History of Camels

As discussed previously [10], through comparison of cortical thickness with that of tibiae from modern dromedary camel, the Nunvaut giant camel was estimated to be 30-35% larger, placing it similar in size to the other giant camels within Camelini, including the Yukon giant camel c.f. *Paracamelus* [10]. Preliminary studies of the Yukon giant camel reveal morphological similarities with other North American giant camels as well as the Eurasian *Paracamelus gigas*, such as the upper molars having well developed styles and ribs [7]. However, it is particularly the size and morphology of the first phalanx that suggests the giant Yukon camel as having closest affinities with the Eurasian *Paracamelus gigas* [7, 10]. The discovery of giant camel remains on Ellesmere Island extends the range of North American Camelini northward by ~1300 km and represents the first definitive evidence that giant camels inhabited the northern regions of a boreal-type forest. Although it is likely that in the Miocene a biotic continuity existed across the Bering Isthmus, this terrestrial biotic connection between North America and Eurasia was severed when tectonic activity led to the opening of the Bering Strait ~5.5 Ma [35, 36] if not open intermittently over several million years earlier [37]. This would have hindered intercontinental dispersal of terrestrial organisms until the Pleistocene, when sea levels were significantly lowered periodically due to continental glaciation and the re-emergence of the Bering Isthmus. If correct, this scenario would suggest that the Nunavut giant camels, and possibly the younger Yukon giant camels, are relicts of a Miocene biotic province that spanned Eurasia and North America [10].

Conclusions

The results presented here ultimately lend support to the recent aDNA studies that place Camelops with Camelini rather than Aucheniini. However, more importantly they support the capabilities of paleoproteomics as a technique that can recover molecular sequence information consistent with aDNA studies but much further back in time. The potential conflict in topology reconstruction with molecular results has remained one of the greatest concerns of the rising new field of paleoproteomics. It is also revealed here that, despite our previous suggestions that MALDI fingerprint data could be important for supporting endogeneity in ancient protein sequences because of its rapid means of displaying the dominant peptides in a protein digest, too much information could be lost, particularly if we being to consider other non-collagenous proteins that do not dominate the sample's proteome. Therefore further work in authenticating ancient sequences needs to be considered, particularly with older fossilised material of extinct species and the various proteins with false positive matches relating to several avenues of potential cross-contamination. Nevertheless, the conclusions from our previous study showing the greater affinity for the Nunavut and Yukon giant camels with each other remain supported by the in-depth sequencing analysis presented here.

Acknowledgements

We thank the Royal Society for fellowship funding to MB and the BBSRC for grant funding that supports CL (BB/L002817/1). Destructive analysis of Yukon and Nunavut fossils was with permissions from the Canadian Museum of Nature (CMN) and the Culture and Heritage Division of the Government of Nunavut, respectively. We also thank M. Currie (CMN) who sampled the Yukon fossils. Field research support for NR was provided by the Canadian Museum of Nature and the Canadian Polar Continental Shelf Program. The 2006 field season was supported by the National Geographic Foundation Explorer grant (7902-05). NR thanks for field teams from the 2006, 2008, 2010 field seasons, which resulted in the recovery of the Nunavut fossil camel remains. The field research was conducted under palaeontology permits from the Government of Nunavut (Culture and Heritage Division), and with the permission of Qikiqtani Inuit Association, especially the Hamlet of Grise Fiord/ Aujuittuq (Nunavut).

References

[1] J.G. Honey, J.A. Harrison, D.R. Prothero, M.S. Stevens, Camelidae, in: C.M. Janis, K.M. Scott, L.L. Jacobs (Eds) Evolution of Tertiary Mammals of North America: Volume 1, Terrestrial Carnivores, Ungulates, and Ungulate Like Mammals, 1 (1998) pp. 439-462.

[2] A. Hassanin, F. Delsuc, A. Ropiquet, C. Hammer, B. Jansen van Vuuren, et al., Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes, Comptes rendus. Biologies, 335 (2012) 32-50.

[3] S.D. Webb. Late Cenozoic mammal dispersals between the Americas. The great American biotic interchange: Springer; 1985. p. 357-86.

[4] B. Kooyman, L.V. Hills, S. Tolman P. McNeil, Late Pleistocene western camel (Camelops hesternus) hunting in southwestern Canada, Am. Antiq. 77 (2012) 115-24.

[5] J.A. Harrison, Revision of the Camelinae (Artiodactyla, Tylopoda) and description of the new genus *Alforjas*, The University of Kansas Palaeontologicals Contributions, 95 (1979), The Palaeontological Institute, The University of Kansas Paper.

[6] P.D. Heintzman, G.D. Zazula, J.A. Cahill, A.V. Reyes, R.D. MacPhee, et al., Genomic data from extinct North American Camelops revise camel evolutionary history, Mol. Biol. Evol. 32 (2015) 2433-40.

[7] M. Pickford, J. Morales D. Soria, Fossil camels from the Upper Miocene of Europe: implications for biogeography and faunal change, Geobios, 28 (1995) 641-50.

[8] C. Harington, Pleistocene vertebrates of the Yukon Territory, Quaternary Science Reviews, 30 (2011) 2341-54.

[9] O.P. Hay. Descriptions of species of Pleistocene Vertebrata, types or specimens of most of which are preserved in the United States National Museum 1921.

[10] N. Rybczynski, J.C. Gosse, C.R. Harington, R.A. Wogelius, A.J. Hidy, et al., Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution, Nat Commun. 4 (2013) 1550.

[11] M. Buckley, C. Anderung, K. Penkman, B.J. Raney, A. Gotherstrom, et al., Comparing the survival of osteocalcin and mtDNA in archaeological bone from four European sites, J. Archaeol. Sci. 35 (2008) 1756-64.
[12] M. Buckley, M. Collins, J. Thomas-Oates J.C. Wilson, Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. 23 (2009) 3843-54.
[13] M. Buckley M.J. Collins, Collagen survival and its use for species identification in Holocene-lower Pleistocene bone fragments from British archaeological and paleontological sites, Antiqua, 1 (2011) 1.
[14] C. Wadsworth M. Buckley, Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone, Rapid Commun. Mass Spectrom. 28 (2014) 605-15.
[15] J.M. Asara, M.H. Schweitzer, L.M. Freimark, M. Phillips L.C. Cantley, Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry, Science, 316 (2007) 280-5.
[16] M.H. Schweitzer, W. Zheng, C.L. Organ, R. Avci, Z. Suo, et al., Biomolecular characterization and protein sequences of the Campanian hadrosaur B. canadensis, Science, 324 (2009) 626.
[17] M. Buckley, A. Walker, S.Y. Ho, Y. Yang, C. Smith, et al., Comment on “Protein Sequences from Mastodon and Tyrannosaurus rex Revealed by Mass Spectrometry”, Science, 4 (2008) 33c.
[18] P.A. Pevzner, S. Kim J. Ng, Comment on" Protein Sequences from Mastodon and Tyrannosaurus rex Revealed by Mass Spectrometry", Science, 321 (2008) 1040.
[19] M. Buckley, S. Warwood, B. van Dongen, A.C. Kitchener P.L. Manning, A fossil protein chimera; difficulties in discriminating dinosaur peptide sequences from modern cross-contamination, Proc R Soc B, 284 (2017) 20170544.
[20] W.G. Armstrong, L.B. Halstead, F.B. Reed L. Wood, Fossil proteins in vertebrate calcified tissues, Phil. Trans. Roy. Soc. B. B301 (1983) 301-43.
[21] P. Palmqvist, D.R. Gröcke, A. Arribas R.A. Fariña, Paleoecological reconstruction of a lower Pleistocene large mammal community using biogeochemical (δ13C, δ15N, δ18O, Sr: Zn) and ecomorphological approaches, Paleobiology, 29 (2003) 205-29.
[22] M. Buckley, A molecular phylogeny of Plesiorycteropus reassigns the extinct mammalian order ‘Bibymalagasia’, PloS ONE, 8 (2013) e59614.
[23] M. Buckley, Ancient collagen reveals evolutionary history of the endemic South American ‘ungulates’. Proc Biol Sci 282 (2015) 20142671.
[24] M. Buckley, R.A. Fariña, C. Lawless, P.S. Tambusso, L. Varela, et al., Collagen sequence analysis of the extinct giant ground sloths Lestodon and Megatherium, PloS ONE, 10 (2015) e0139611.
[25] S. Kumar, G. Stecher, M. Li, C. Knyaz K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms, Mol. Biol. Evol. 35 (2018) 1547-9.
[26] J.P. Huelsenbeck F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics, 17 (2001) 754-5.

[27] M. Buckley, Species Identification of Bovine, Ovine and Porcine Type 1 Collagen; Comparing Peptide Mass Fingerprinting and LC-Based Proteomics Methods., Int. J. Mol. Sci. 17 (2016) E445.
[28] M. Buckley, S.W. Kansa, S. Howard, S. Campbell, J. Thomas-Oates, et al., Distinguishing between archaeological sheep and goat bones using a single collagen peptide, J. Archaeol. Sci. 37 (2010) 13-20.
[29] J. Wilson, N.L. van Doorn M.J. Collins, Assessing the extent of bone degradation using glutamine deamidation in collagen, Anal. Chem. 84 (2012) 9041-8.
[30] N. Procopio, A. Williams, A.T. Chamberlain M. Buckley, Forensic proteomics for the evaluation of the post-mortem decay in bones, J. Proteom. 177 (2018) 21-30.
[31] T.P. Cleland, E.R. Schroeter M.H. Schweitzer, Biologically and diagenetically derived peptide modifications in moa collagens, Proc Biol. Sci 282 (2015) 20150015.
[32] R.C. Hill, M.J. Wither, T. Nemkov, A. Barrett, A. D'Alessandro, M. Dzieciatkiwska, K.C. Hansen, Preserved proteins from extinct Bison latifrons identified by Tandem Mass Spectrometry; hydroxylysine glycosides are a common feature of ancient collagen, Mol. Cell Proteomics 7 (2015) 1946-1958.
[33] D.R. Sell, V.M. Monnier, Conversion of arginine into ornithine by advanced glycation in senescent human collagen and lens crystallins, J Biol Chem, 279 (2004) 54173-84.
[34] S.M. Sweeney, J.P. Orgel, A. Fertala, J.D. McAuliffe, K.R. Turner, et al. Candidate cell and matrix interaction domains on the collagen fibril, the predominant protein of vertebrates, J Bio Chem 283 (2008) 21187-97.
[35] A.Y. Gladenkov, Neogene diatoms from the Sandy Ridge section, Alaska Peninsula: significance for stratigraphic and paleogeographic reconstructions, Stratigr. Geol. Correl. 14 (2006) 73-90.
[36] A.Y. Gladenkov, Y.B. Gladenkov, Onset of connections between the Pacific and Arctic Oceans through the Bering Strait in the Neogene, Stratigr. Geol. Correl. 12 (2004) 175-8.
[37] Y.I. Polyakova, Late Cenozoic evolution of northern Eurasian marginal seas based on the diatom record, Polarforschung, 69 (2001) 211-20.