This is to certify that the
thesis entitled

INVESTIGATION OF THE PROTEIN OSTEOCALCIN OF *CAMELOPS HESTERNUS*: SEQUENCE, STRUCTURE, AND PHYLOGENETIC IMPLICATIONS

presented by

JAMES FREDERICK HUMPULA

has been accepted towards fulfillment
of the requirements for the

M.S. degree in Geological Sciences

Major Professor's Signature

12/11/06

Date

MSU is an Affirmative Action/Equal Opportunity Institution

INVESTIGATION OF THE PROTEIN OSTEOCALCIN OF *CAMELOPS HESTERNUS*: SEQUENCE, STRUCTURE, AND PHYLOGENETIC IMPLICATIONS

By

James Frederick Humpula

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Geological Sciences

2006

ABSTRACT

INVESTIGATION OF THE PROTEIN OSTEOCALCIN OF *CAMELOPS HESTERNUS*: SEQUENCE, STRUCTURE, AND PHYLOGENETIC IMPLICATIONS

By

James Frederick Humpula

Due to its preservation potential, the bone protein osteocalcin is recognized as a potential source of genetic information for use in molecular paleontology, and potentially allow ambiguous evolutionary relationships between extant and extinct taxa to be clarified. To this end, osteocalcin was isolated and sequenced from a 21 Kyr bone of *Camelops hesternus* and three modern camelid species (*Camelus bactrianus*, *Camelus dromedarius* and *Llama guanicoe*). Complete sequence coverage for *C. hesternus* was obtained via peptide mass fingerprinting and tandem mass spectrometry. This marks the second complete osteocalcin sequence from a fossil preserved in a temperate climate and the first molecular data for the genus *Camelops*. The data indicate that there are no sequence differences among the ancient and modern camelid species. Although sequence similarity precludes phylogenetic analysis of camelids, we were interested in exploring evolutionary relationships among artiodactyls and among higher order taxa using osteocalcin sequences. The character analysis of six artiodactyls provided a low level of taxonomic resolution. A maximum likelihood tree based on osteocalcin sequences from 25 taxa was produced and analyzed in comparison to a cytochrome b based maximum likelihood tree. The osteocalcin tree showed a higher amount of homoplasy and lower taxonomic resolution. A more complete character analysis of osteocalcin must be carried out to more clearly understand the affect of homoplasy at specific taxanomic levels.

This thesis is dedicated to the memory of my grandfather, Fred Tithof. I'll never let the van get away again.

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Peggy Ostrom for her input, understanding, and constant support; without her prodding, I might have never finished. Dr. Hasand Gandhi for his priceless knowledge of the methods and machines used in this study. Dr. John Strahler for carrying out the MS/MS analyses included here. Dr. Joe Leykam for conducting my Edman Sequencing. Dr. Jim Smith for assistance with phylogenetics. Dr. R. George Corner and Dr. Michael Voorhies for supplying fossil material. The MSU Museum, the American Museum of Natural History, and the University of Michigan Museum for providing modern samples. The MSU Mass Spectrometry Facility for the use of their equipment. Dr. Thomas Stafford and Dr. Russel Graham for their help and ideas. The National Science Foundation, because without grant EAR-0309467 I would not have been able to do much of anything. My friends Aaron and Chuck, for keeping me sane. My parents, for always believing in me.

Finally, I would like to thank the molecule $C_8H_{10}N_4O_2$ and the chemical reactions it induced, as I would be sleeping right now without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The Family Camelidae originated in the early Tertiary of North America (Kurtén and Anderson, 1980). Belonging to the order Artiodactyla, the camelids are characterized by a lack of horns, long slender necks, and long limbs with a much-reduced ulna and fibula (Kurtén and Anderson, 1980). Most of camelid evolution occurred in North America, with both camels and llamas being common during the Pleistocene (Kurtén and Anderson, 1980; Lange, 2002). North American camelids became extinct at the end of that epoch. The family is currently comprised of two extant tribes, Lamini and Camelini, which split from a common ancestor approximately 11 Ma ago (Webb, 1974). Lamini and Camelini are comprised of four New World and two Old World species, respectively (Webb, 1974).

The genus *Camelops*, belonging to the extinct tribe Camelopini, is comprised of five separate species which existed in northern North America between two million and 10,000 years ago (Savage, 1951). This genus can be divided into two groups based on size and dental characters. The larger of the two groups is comprised of the species *Camelops hesternus* and *Camelops heurfanensis* (Savage, 1951). *C. hesternus*, or Yesterday's Camel, first appeared 300,000 years ago, and was abundant throughout the western United States, the Yukon, and Alaska (Kurtén and Anderson, 1980). For unknown reasons, it went extinct between 12,600 and 10,800 years ago. Similar in relative proportion to *Camelus dromedarius*, or Dromedary camel, the legs of *C. hesternus* were approximately 20 percent longer than the modern camelid. *C. heurfanensis* is very similar to *C. hesternus*, and is differentiated from the other large *Camelops* species by the location of the postpalatine foramen and the placement of the

lower border of the mandible from the symphysis to a point below the last molar (Savage, 1951). Kurtén and Anderson (1980) suggest that these two species may be conspecific due to the small number of morphological differences between them.

Recently, the bone protein osteocalcin has been identified and implemented as a potential source of genetic information from fossils (Ostrom et al., 2000; Nielsen-Marsh et al., 2002; Nielsen-Marsh et al., 2005; Ostrom et al., 2006). Because the likelihood of survival for some proteins is greater than that of DNA (Collins et al., 2000), protein sequences may assist in extending molecular records into the past. Osteocalcin, also known as Bone Gla Protein or BGP, is a 46-50 amino acid long acidic protein whose potential for survival is attributed to a high affinity for hydroxyapatite (Hauschka et al., 1989; Lian et al., 1989 Cancela et al., 1995). Osteocalcin's conserved region contains a single disulfide bridge connecting amino acids 23 and 29 and, in all but one species, three gama-carboxyglutamic acid (Gla) residues that occur at positions 17, 21, and 24 (Hauska et al., 1989). The marine fish *Argyrosomus regius* also contains Gla_{25} (Frazao et al., 2005). The Gla residues and the disulfide bridge are key factors in the stabilization of osteocalcin and maintaining its association with the hydroxyapatite mineral phase (Hoang et al., 2003).

Our laboratory has been exploring the preservation of DNA relative to osteocalcin. While ancient DNA is an important and established source of data for determining phylogenetic relationships of extinct taxa (Wayne et al., 1999; Barnes et al., 2002), it is most frequently obtained from cold climates and from samples less than 50 ka (Wayne et al., 1999; Hofreiter et al., 2001). If protein sequences are to expand the number of extinct taxa from which molecular data can be obtained, they must be

phylogenetically informative. Evolutionary comparisons between sequenced osteocalcin proteins have been explored (Nielson-Marsh et al., 2002), but very little work has been done to determine the taxonomic level of resolution and the strength of the hypothesized phylogenetic relationships derived based on the osteocalcin protein (Laizé et al., 2005). To address these issues, a phylogenetic analysis of osteocalcin sequences should be compared to a robust, well-supported DNA-based tree. Zardoya and Meyer (1996) determined the level of performance of all mitochondrial protein-coding regions in returning two expected topologies. Cytochrome b was among the highest performing mtDNA-coding sequences found, and therefore appropriate for evaluating the phylogenetic efficacy of the osteocalcin protein sequence.

The complete osteocalcin sequence for *Camelops hesternus* is presented here, as well as complete osteocalcin sequences for both modern camels and guanaco. In addition, the evolution of osteocalcin within artiodactyls is investigated, and a phylogenetic analysis of osteocalcin based on 24 species of vertebrates is presented. Comparisons between the phylogenetic tree based on osteocalcin sequence data and a tree based on Cytochrome b allow for determination of the degree of phylogenetic resolution and examination of specific phylogenetic relationships exhibited in the osteocalcin tree.

METHODS

Sample Description and Preparation

A metapodial from a *Camelus bactrianus* (modern bactrian camel) (MSU-9669) and the right femur of *Llama guanicoe* (modern guanaco) (UM 157201) were obtained from Michigan State University Museum and University of Michigan, respectively. A vertebral fragment of dromedary camel (*Camelus dromedarius*) (AMNH 10734) was obtained from the American Museum of Natural History. The University of Nebraska State Museum provided a phalanx of *Camelops hesternus* from Isleta Cave, New Mexico. The sample was dated to 21,190 ± 110 RC yr (CAMS-22182) by Stafford Research Laboratories. The Isleta Cave site is located within a Quaternary lava flow approximately 13 km west of Isleta, New Mexico, at an elevation of 1,716 m. The fossil material found within the cave includes 42 mammalian and six reptilian genera (Harris and Findley, 1964).

Samples were mechanically cleaned with a Dremel tool (Moto-tool model 395), cut into 20-50 mg pieces using a jeweler's saw, and powdered at -190^{O}C in a freezer/mill (SPEX CertiPrep 6750). The sample was then demineralized (sodium EDTA, 4 hrs, 25 °C), centrifuged (14,000 rpm, 10 min) and the resulting supernatant applied to a fresh C_{18} gravity column (60 Å, Fisher, 0.5 x 2 cm). The column was eluted with eight, 1 mL aliquots of a mixture of solvent A and solvent B, with increasing concentration of solvent B (20%, 25%, 30%, 32%, 34%, 36%, 38%, and 40%) and the eluent from each of the eight fractions was collected, concentrated (Speedvac, Eppendorff), and reconstituted using 10 µL of 1% *n*-octyl-β-D-glucopyranoside (OGP, Sigma) in 50 mM Tris-HCl, pH 8.0. The reconstituted eluent was diluted 1:10 with solvent A and 0.5 µL was spotted on a

MALDI target, with 0.5 μL of a 1% solution of TFA and 0.5 μL of 4-hydroxy-α-cyanocinnamic acid (4-HCCA) matrix. At Michigan State University, MALDI-MS from an Applied Biosystems DE-STR (ABI-STR) was used to identify gravity column fractions which contained a peak consistent with the *m/z* of osteocalcin (~5.7 kDa) (Hauschka et al., 1989). Spectra were externally calibrated using the 1+ and 2+ peaks of bovine insulin (average mass 5734.6 Da; Sigma). Resolution was ca. 530 and mass accuracy at *m/z* of 2800 was better than 630 ppm (range 92-630 ppm).

Purification/ Digestion

Osteocalcin was purified using rpHPLC fitted with a peptide trap (1 x 10 mm, Michrom BioResources) and a 1 x 150 mm C_{18} column (300 Å, 5 μm, Reliasil, Michrom BioResources) (Nielsen-Marsh et al., 2002). The peptide trap was equilibrated with 20% B and the sample injected. After the trap was washed with 300 μL of solvent A, the flow from the trap was diverted to the column. Osteocalcin was eluted with a gradient of 20% solvent B, gradually increasing to 30% over 15 min, held at 30% for 5 min, increased to 35% over 15 min, held at 35% for 10 min, increased to 95% over 5 min, and held for 5 min. All peaks were collected and MALDI-MS was used to identify putative osteocalcin.

Peaks from the rpHPLC containing putative osteocalcin were dried, reconstituted with 10 μL of 1% OGP, and digested with Trypsin (Promega) or Asp-N (Sigma). The digest products were purified using rpHPLC (gradient 5% solvent B, gradually increasing to 40% over 40 min, held at 40% for 10 min, increased to 95% over 1 min, held at 95% for 5 min, and decreased to 5% over 1 min) and analyzed using MALDI-MS. The 20-43 peptide of *C. dromedarius* was then subjected to an MMTS reduction to remove the

cystine bridge and allow MS/MS sequencing of the amino acids between the cystine residues.

MS/MS and Edman Sequencing

MALDI-MS was carried out on intact osteocalcin and both MALDI-MS and tandem mass spectrometry (MS/MS) were conducted on the digest products using an Applied Biosystems ABI-4700 (University of Michigan). MALDI-MS spectra were acquired using 1500-3000 laser shots. Spectra of intact osteocalcin were externally calibrated using the 1+, 2+, and β-chain peaks of bovine insulin (monoisotopic mass 5730.6, Sigma). Resolution was ca. 13,000 and mass accuracy was better than 34 ppm (range 9-34 ppm). MS/MS spectra were obtained using 8,000 to 10,000 laser shots, and were calibrated using fragmentation of angiotensin II. External calibration of the MS/MS spectra was done using the Mass Standards Kit for the 4700 Proteomics Analyzer (Applied Biosystems). Atmosphere was used as the collision gas (pressure: 6×10^{-7} torr, collision energy: 1 kV). Edman sequencing was performed on an Applied Biosystems 494 CLc.

Cytochrome b Reference Tree

The mtDNA data set consists of 24 complete sequences from the Cytochrome b protein-coding region of mtDNA. Taxa used to create the osteocalcin tree were also used to create the mtDNA tree (Table 1). Sequences were aligned individually using ClustalX with default settings then manually corrected in MacClade 4.06 (Maddison and Maddison, 2000). The aligned sequences were then analyzed using maximum likelihood,

Table 1. Sequence reference numbers for osteocalcin and cytochrome b for all species used in this study.

Scientific Name	Common Name	Osteocalcin Reference Number	Cytochrome b Reference Number
Bison bison	Bison	P83489	AY689186
Bos taurus	Cow	NP_776674	NC_006853
Camelus dromedarius	Dromedary Camel	Seuqence Included	X56281
Canis familiaris	Dog	P81455	NC_002008
Capra hircus	Goat	P0C225	NC_005044
Dromaius novaehollandiae	Emu	P15504	NC_002784
Equus caballus	Horse	Ostrom et al., 2006	NC_001640
Felis catus	House Cat	P02821	NC_001700
Gallus gallus	Chicken	NP_990718	NC_001323
Gorilla gorilla	Gorilla	P84349	NC_001645
Homo sapiens	Human	NP_954642	AC_000021
Macaca sp.	Macaque	P02819	AJ309865
Mus musculus	House Mouse	AAA39856	NC_005089
Oryctolagus cuniculus	European Rabbit	P39056	NC_001913
Ovis aries	Sheep	ABD83814	NC_001941
Pan troglodytes	Chimpanzee	P84348	NC_001643
Pongo pygmaeus	Orangutan	Q5RDP6	NC_001646
Rana sp.	Bull Frog	BAD16774	AF205094
Rattus norvegicus	European Rat	NP_038200	NC_001665
Setonix sp.	Quokka	1005180C	
Sus scrofa	Pig	AAN73020	NC_000845
Takifugu rubripes	Fugu Puffer	AAO24898	NC_004299
Tetraodon nigroviridis	Spotted Green Puffer	AAO24897	NC_007176
Xenopus laevis	African Clawed Frog	AAB36024	NC_001573
Xenopus tropicalis	Western Clawed Frog	NP_001006689	NC_006839

maximum parsimony bootstrap, and neighbor-joining bootstrap methods. *Takifugu rubripes* and *Tetradon nigroviridis* were chosen as outgroups due to the fishes' early divergence from all other taxa included in the study.

Neighbor joining, maximum parsimony, and maximum likelihood analyses were carried out using PAUP* 4.0b10 (Swofford, 2002). Modeltest 3.7 (Posada and Crandall, 1998) was used to determine an appropriate model for DNA substitution in the maximum likelihood analysis. The GTR+I+G model of substitution was used, and the heuristic search type was used to find the most likely tree. For the neighbor joining analysis, the distance type analysis was chosen. Uncorrected ("p") values were used for distance, with ties between trees broken randomly. 1,000 replicates were used in the bootstrap analysis, with all other settings at default. The maximum parsimony analysis consisted of a heuristic search with uninformative characters excluded and the initial tree found using stepwise addition. 1,000 replicates were performed in the bootstrap analysis.

The maximum likelihood tree was chosen as the tree that best represents the accepted tree topology (ncbi.nlm.nih.gov), and the maximum parsimony bootstrap, neighbor joining bootstrap, and Bayesian trees were used as support.

Osteocalcin Comparison Tree

The data set of osteocalcin sequences analyzed in this study is comprised of 25 complete protein sequences. Sequences were aligned with ClustalX (Thompson et al., 1997) using the Gonnet series matrix. For the pairwise parameters, the gap opening penalty was set to 35.00 and the gap extension penalty was set to 0.75. For the alignment parameters, the gap opening penalty was set to 15.00 and the gap extension penalty was set to 0.30. The alignment was unaffected by changing the parameters and needed no further modification.

The aligned osteocalcin sequences were then analyzed using a maximum likelihood analysis, with *T. rubripes* and *T. nigroviridis* used as outgroups. Maximum likelihood analyses were performed using Tree-puzzle 5.0 (Schmidt et al., 2002). The analysis used the quartet puzzling algorithm to perform a search of the tree space. 10,000 puzzling steps were performed, parameter estimates were set to exact, and the JTT (Jones et al., 1992) model of amino acid substitution was used.

RESULTS

Mass spectra were obtained using both ABI-STR and ABI-4700 MALDI-TOF mass spectrometers. Ions from the ABI-STR are reported as average *m/z* and nominal mass while those from the ABI-4700 are reported as monoisotopic and specified to one decimal place. Gravity column fractions of *C. hesternus* contain $[M+H]^+$ ions of *m/z* 5625 and 5668, which are within the range of osteocalcin from modern vertebrates (M_r = 5210–5889) (Hauschka et al., 1989). The 43 Da difference between these analytes is consistent with carbamylation as previously observed in fossil osteocalcin (Ostrom et al., 2006).

For sequencing by PMF and tandem mass spectrometry (MS/MS), the gravity column fractions were purified using rp-HPLC and digested with trypsin or ASP-N. Because no sequences for osteocalcin from extant camelids exist, interpretations were made by comparison to predicted tryptic fragments of bison osteocalcin (Nielsen-Marsh et al., 2005) (*Bison bison*: YLDHGLGAPA PYPDPLEPKR EVCELNPDCD ELADHIGFQE AYRRFYGPV: with decarboxylation of Gla_{17}, Gla_{21}, Gla_{24}; Hyp_9; disulfide bond between Cys_{23} and Cys_{29}).

The tryptic digest of the *m/z* 5668 analyte produced several peptides that were re-purified using rp-HPLC and subjected to MALDI-MS/MS. These included peaks at *m/z* 744.4, 2110.9, 2828.2 and 3002.3 (Figure 1, A). The primary digest product of ASP-N, *m/z* 1987.7, was also purified and subjected to MS/MS. It was hypothesized that these peaks represented residues 44-49 (744.4), 34-49 (*m/z* 1987.7), 1-19 (*m/z* 2110.9), 20-43 (*m/z* 2828.2), and 20-44 (*m/z* 3002.3), but contained modifications or substitutions relative to the bison sequence. A tryptic digest peak *m/z* 2067.7 was also recovered,

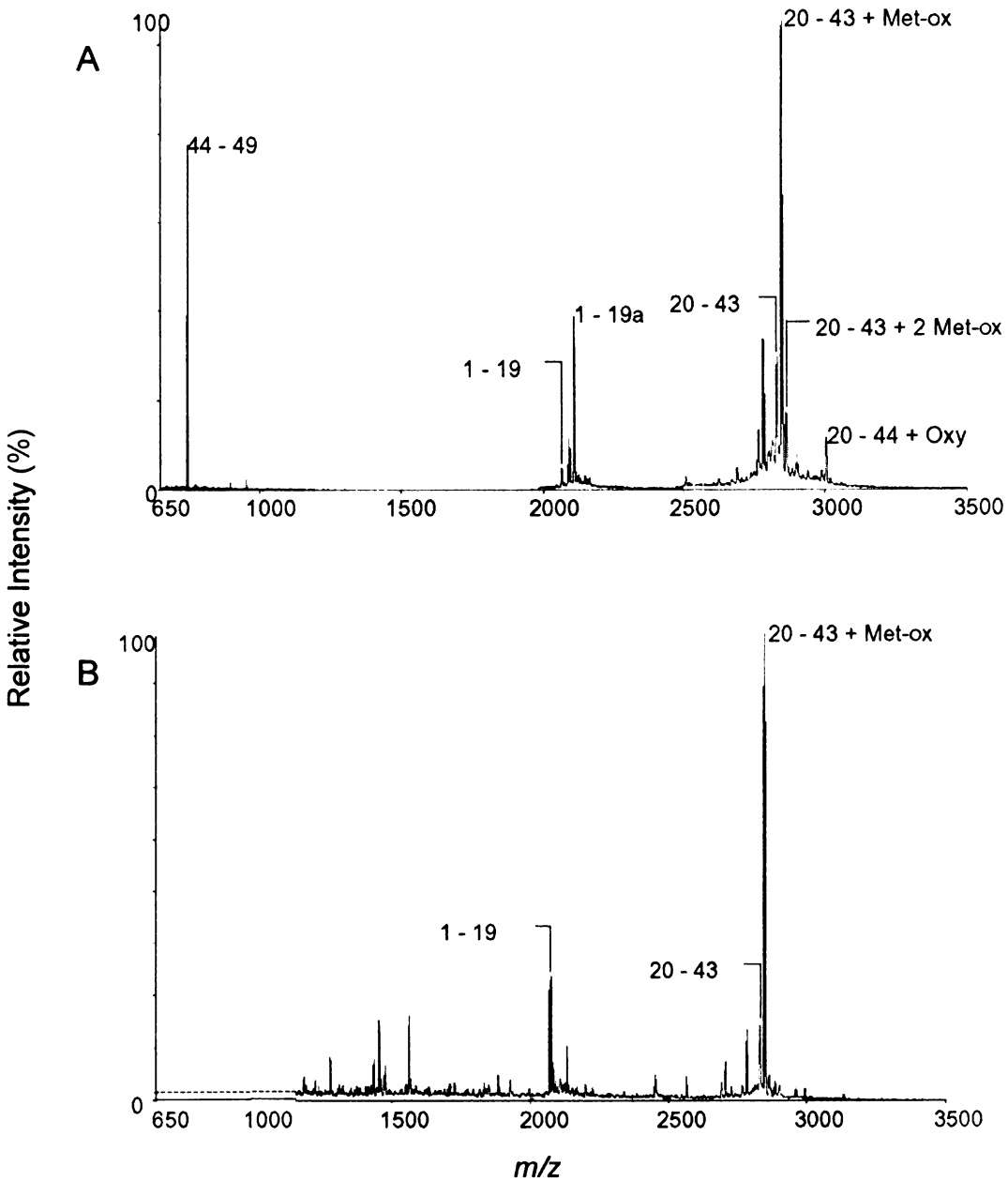
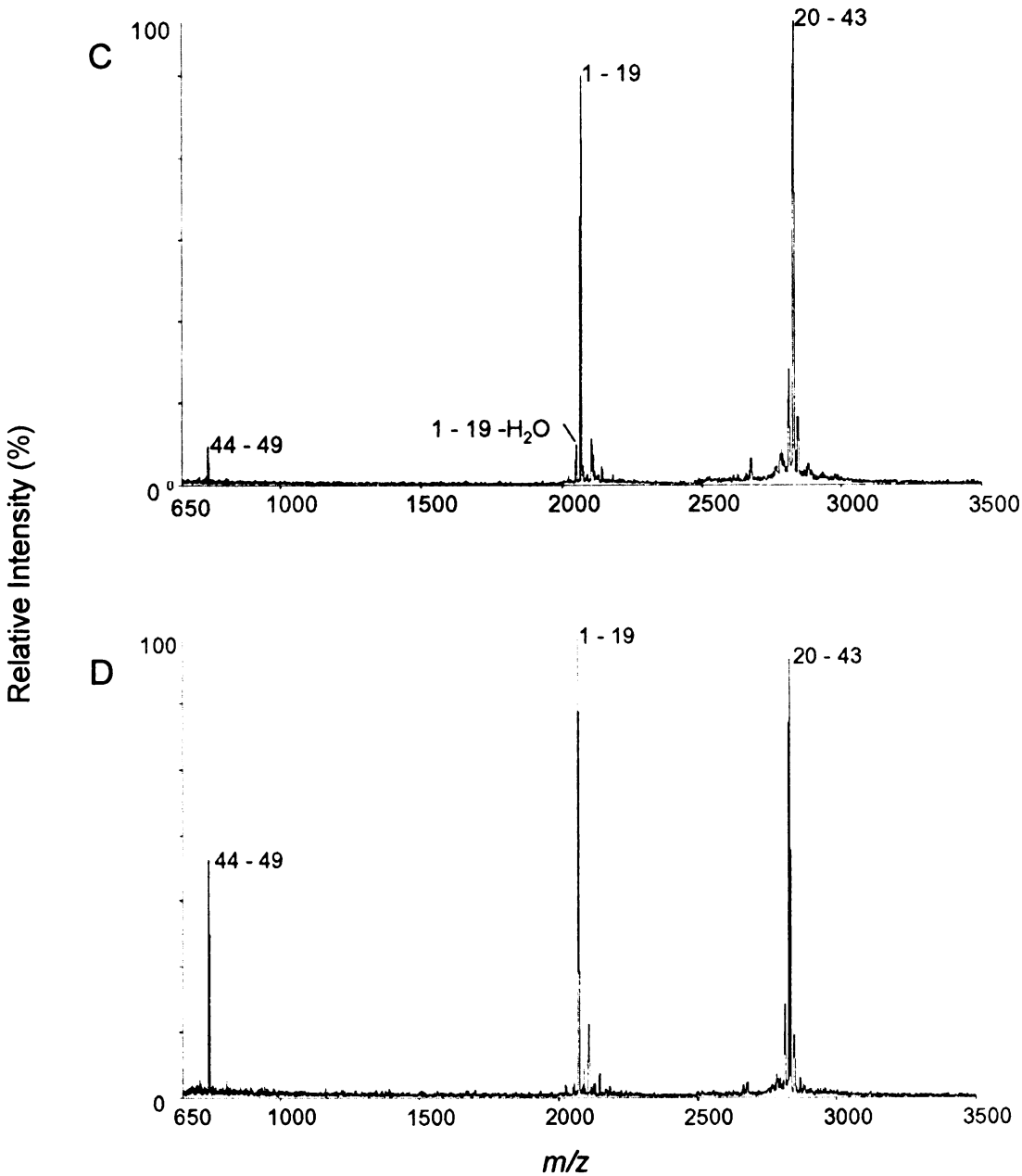


Figure 1. Peptide mass fingerprint for osteocalcin tryptic digest from: (A) 21 ka *Camelops hesternus* produced from an ABI-4700 (confirmed on an ABI-STR); (B) *Camelus dromedarius*; (C) *Camelus bactrianus*; and (D) *Llama guanicoe*. Peptide 1–19a, 43 Da greater than 1–19, is osteocalcin modified by carbamylation.

Figure 1 **Continued.**



which differs from the primary hypothesized 1-19 peak, *m/z* 2110.9, by 43 Da. This suggests that *m/z* 2110.9 was modified by carbamylation.

MS/MS was used to resolve the location of carbamylation and obtain sequence information for each of the digest products. MS/MS of *m/z* 2110.9 (putative 1-19) produced a series of y-ion fragmentations whose sequence is equivalent to bison 2-14 with a substitution of Val_{13} for Pro_{13} (Figure 2, A). Asp_{14} is further confirmed by b_{13} and b_{14}. The mass to charge ratio of the y_{18} ion (*m/z* 1905.1) differs from that of its parent ion (*m/z* 2110.9) by 205.8 Da, indicating carbamylation of Tyr at the N-terminus (Tyr = 163.1). A series of peaks 43 Da greater than the *m/z* expected for several b ion fragments denoted b* are consistent with N-terminal carbamylation. The carbamyl-modified b ions confirm the y ion assignments for residues 4-8 and identify Pro_{18} and Lys_{19}.

The ion fragmentation pattern produced from the MS/MS of the tryptic digest product *m/z* 2828.2 (putative 20-43) indicates residues 37-43 are the same as bison (Figure 2, B). The *m/z* 286.3 and 357.1 are equivalent to the b_2 and a_3 ions of bison. There are two residue pairs with a mass of 286.1 Da: Arg-Glu and Trp-Val. Residues 20 and 21 of all other mammal species with known osteocalcin sequences are Arg-Glu, suggesting that this is also the case for residues 20 and 21 of *C. hesternus*. MS/MS data for the tryptic digest product *m/z* 3002.3 (putative 20-44) confirms that residues 37-43 are the same as bison and identifies Arg_{44}.

The tryptic digest product *m/z* 744.4 (putative 44-49) has a peak at *m/z* 304.2, consistent with the residue pair RF (Figure 2, C), the only unmodified residue pair with that mass. The b ion series identifies residues Tyr_{46}, Gly_{47}, Thr_{48}, and Thr_{49}.

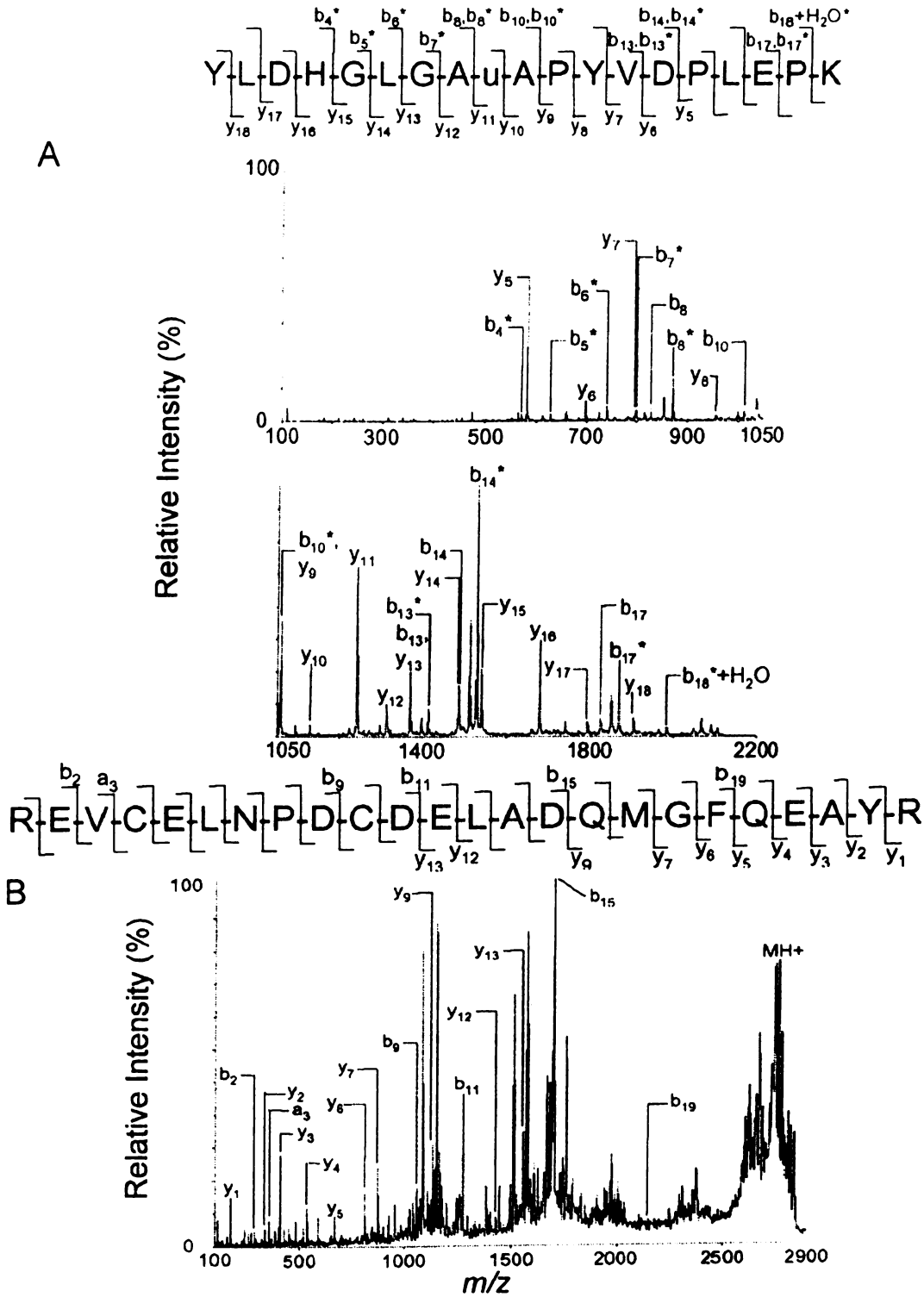
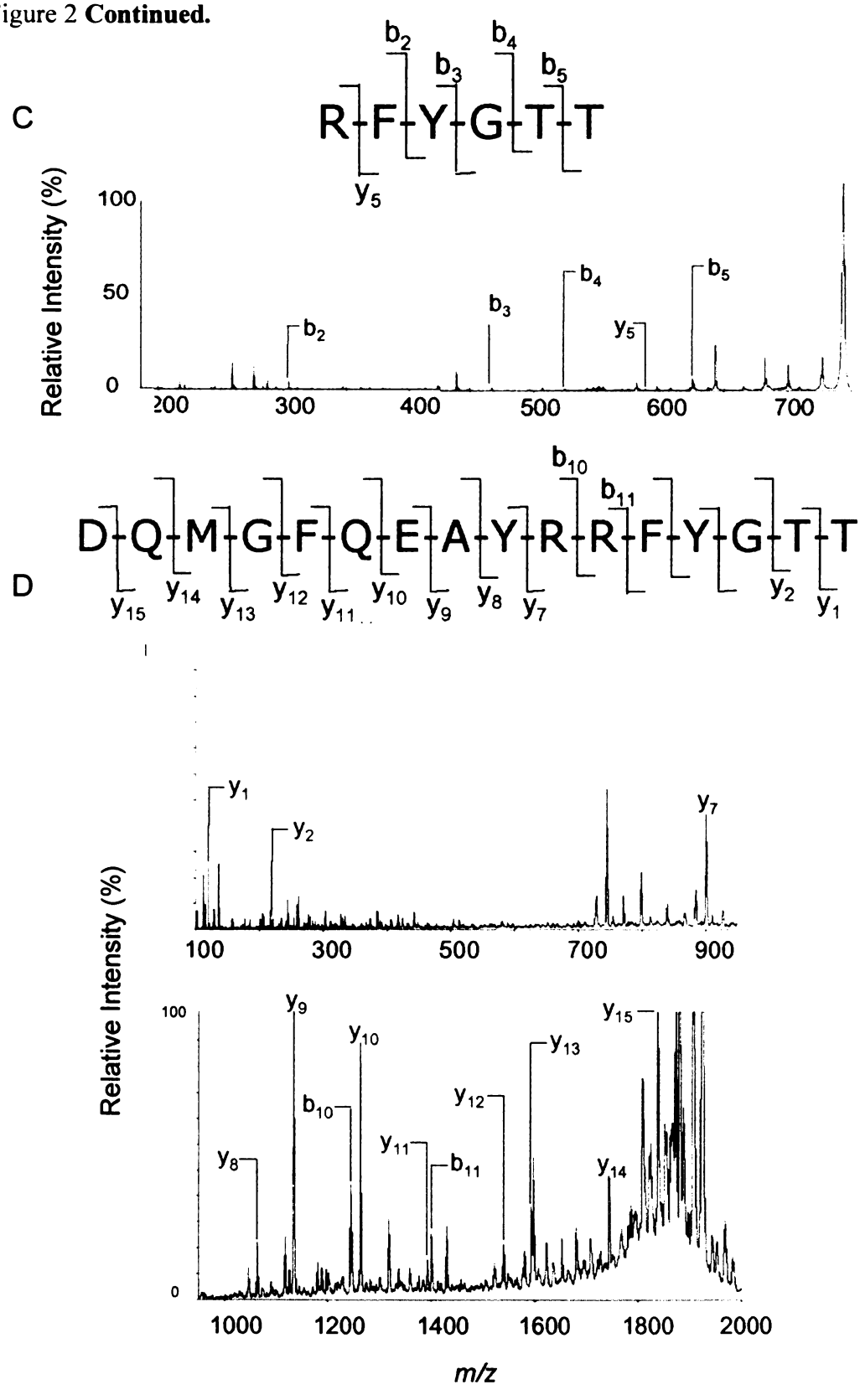


Figure 2. MS/MS spectra for tryptic peptides (A) 1–19, (B) 20–43, and (C) 44-49 and (D) Asp-N peptide 34–49 of *Camelops hesternus*. Spectra produced on the ABI-4700. Y-ions show coverage for residues 1-14 in peptide 1–19, with carbamylated b-ions, identified as b*, providing residues 18 and 19. Together, peptides 20–43, 44-49, and 34–49 provide sequence for residues 22, 31, and 34-49. For each spectrum, right and left panels of each are scaled for clarity. To allow comparison, both panels are scaled relative to the largest ion in the spectrum.

Figure 2 **Continued.**



The MS/MS of ASP-N digest product *m/z* 1987.7 (putative 34-49) confirms Thr_{48} and Thr_{49}. It also produced several peaks at an *m/z* consistent with y_7 to y_{12} of bison, allowing tentative assignments for residues 37-42 (Figure 2, D).

Based on comparison to bison, our data for *C. hesternus* define residues 1-14, 18-21, and 37-49 (YLDHGLGAPA PYVD - - - PKR E - - - - - - - - - - - - - - GFQE AYRRFYGTT) (carbamylation of Tyr_1, decarboxylation of Gla_{17}, Gla_{21}, Gla_{24}, hydroxylated Pro_9, and a disulfide bond between Cys_{23} and Cys_{29}). To resolve the remaining sequence ambiguity and for taxonomic comparison, we sequenced three modern camelids (*C. bactrianus*, *C. dromedarius,* and *L. guanicoe*) by PMF and tandem mass spectrometry. Sequence determinations were based on comparison to *C. hesternus* and *B. bison*. The PMF of tryptic products from modern camelids produced two or three primary peaks: putative 44-49, 1-19, and 20-43 (Figure 1, B-D). The *m/z* of 1-19, 20-43, and 44-49 are the same for the three modern taxa and *C. hesternus*. The *m/z* of 34-49 of *C. bactrianus* and *L. guanicoe* are the same (*m/z* 1970.2) but that of *C. hesternus* (*m/z* 1987.7) and *C. dromedarius* (*m/z* 1986.4) are 16 and 18 Da higher, respectively.

Complete sequence coverage for 1-19 of *C. dromedarius* is provided by MS/MS (Figure 3, A). A 28 Da increase in the y and b ions containing residue 9 in the MS/MS of the formylated 1-19 confirm Hyp_9. As with *C. hesternus*, we observe Val_{13}. The MS/MS of 1-19 of osteocalcin from *C. bactrianus* identifies 3-19 and from *L. guanicoe* identifies 3-10, 13-19. These data suggest that the osteocalcin sequences for 1-19 in modern taxa are identical.

Sequence coverage for 22-31 and 34-43 of *C. dromedarius* was obtained from the MMTS derivatized tryptic peptide 20-43 (Figure 3, B). The difference in *m/z* defined by

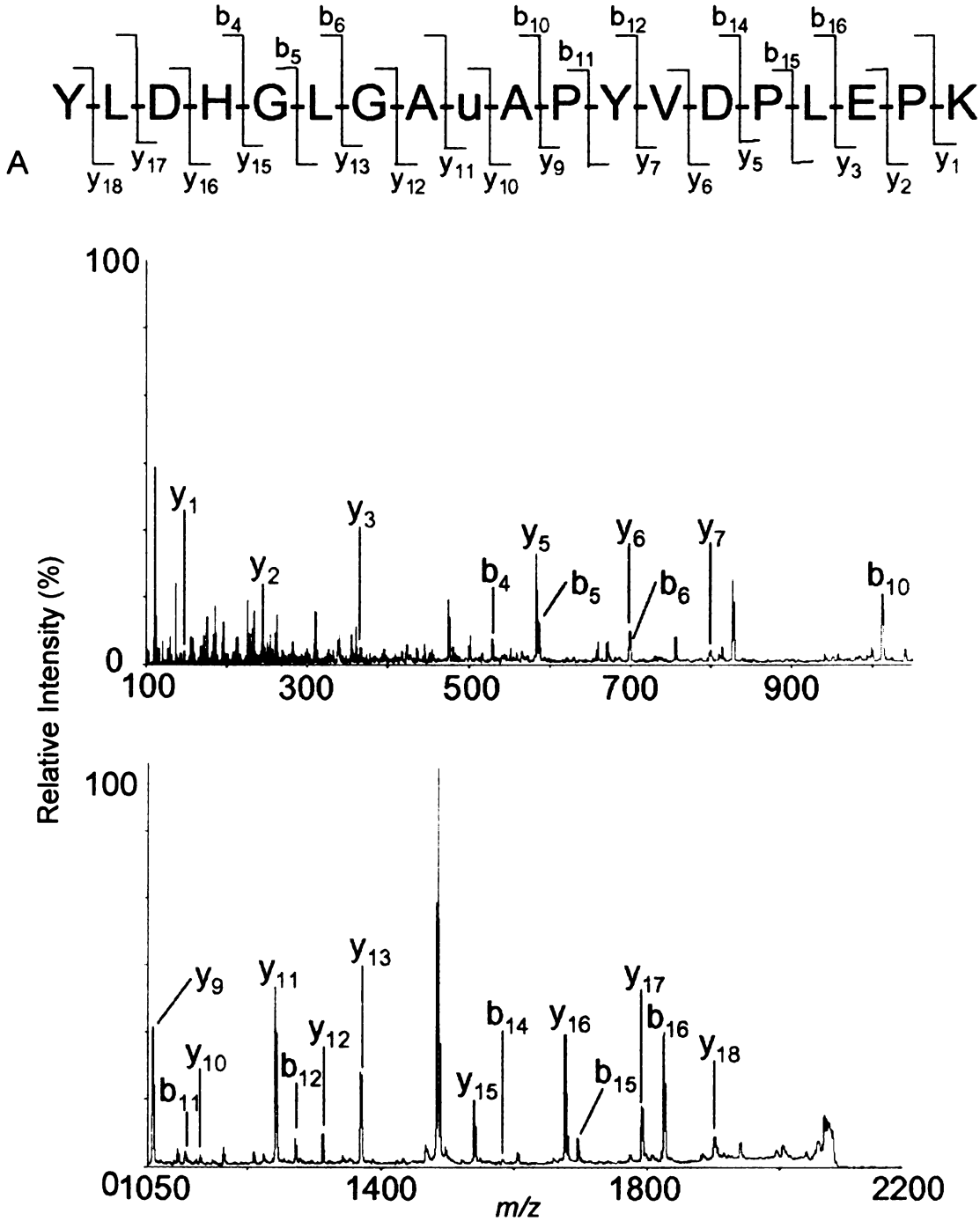
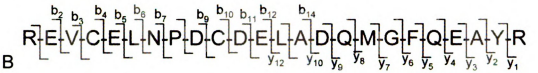
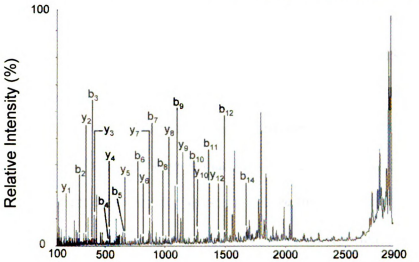


Figure 3. MS/MS spectra for tryptic peptides (A) 1–19 and (B) MMTS-derivitized 20–43, and (C) Asp-N peptide 34–49 of *Camelus dromedarius*. Spectra produced on the ABI-4700. Peptide 1-19 shows complete coverage, and together peptides 20–43 and 34–49 provide sequence for residues 22-31, 34-44, and 47-49. For each spectrum, right and left panels of each are scaled for clarity. To allow comparison, both panels are scaled relative to the largest ion in the spectrum.

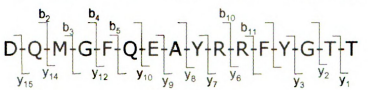
Figure 3 **Continued.**

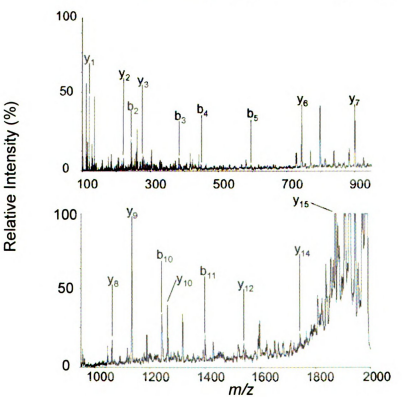
B





C





y_7 and y_8 (147.0 Da) is consistent with either methionine sulfoxide (Met-ox) or Phe in position 36. The *m/z* of b_2 (*m/z* 286.2) is consistent with either residue pairs RE or WV. There is no known osteocalcin sequence with Trp_{20} and Val_{21}, and furthermore Gla_{21} (decarboxylated to Glu_{21} during MALDI-MS) is an important hydroxyapatite-binding site for osteocalcin (Hoang et al. 2003). Assuming Arg_{20} and Glu_{21}, MS/MS data for 20-43 of osteocalcin from *C. bactrianus* provides sequence for 20-24 and 29-43 and identified position 36 as Met. Data for *L. guanicoe* provide residues 20-24, 29-34, and 37-42. These data and those of the PMF suggest that the sequences for 20-43 from the modern and fossil taxa are identical except for Met-ox_{36} in *C. dromedarius* and *C. hesternus*. Met-ox_{36} explains the 16 Da discrepancy in the *m/z* of 20-43 reported in the PMF of tryptic products for *C. dromedarius* and *C. hesternus* relative to the two other modern taxa.

The MS/MS of ASP-N peptide 34-49 of *C. dromedarius* identifies residues 34-44, 48, and 49 as well as Met-ox_{36} (Figure 3 C). MS/MS data for 34-49 of *C. bactrianus* and *L. guanicoe* provides residues 34-47 as well as identifying Met_{36}. Residues 44-49 of *C. bactrianus* and *L. guanicoe* are confirmed by MS/MS of tryptic peptide 44-49 (complete coverage) and Edman sequencing of 35-49 confirms the sequence for residues 34-38 of *C. bactrianus*. These data suggest that with the exception of residue 36 there is no difference in the sequence of 34-49 among modern taxa. The MS/MS of putative 34-49 of *C. hesternus* (*m/z* 1987.7) shows a series of y-ions consistent with Met-ox_{36} and deamidation of Gln_{35} and/or Gln_{39}. The 17.5 Da increase in the *m/z* of 34-49 of *C. hesternus* relative to that of *C. bactrianus* and *L. guanicoe* (*m/z* 1970.2) is also consistent with Met-ox_{36} and deamidation of Gln_{35} and/or Gln_{39}.

Data from the modern camelids allow us to define some of the sequence ambiguities for osteocalcin of *C. hesternus*. The mass of y_5, b_{17}, and the difference in *m/z* between b_{14} and b_{17} (*m/z* 339.0) of tryptic peptide 1-19 is consistent with the tripeptide PLE in positions 15 to 17. For tryptic peptide 20-43, b_9, b_{11}, b_{15}, y_9, y_{12}, and y_{13} are consistent with the sequence of the modern camelids. Similarly, for ASP-N peptide 34-49 y_{13} to y_{15}, b_{10}, and b_{11} are the same as modern camelids with Met-ox$_{36}$. The MS/MS results and the PMF indicate that the sequences of the camelids are identical with either Met$_{36}$ or Met-ox$_{36}$.

DISCUSSION

We isolated and sequenced osteocalcin from a 21 Kyr bone of *Camelops hesternus*, as well as from three modern camelid species (*Camelus bactrianus*, *Camelus dromedarius* and *Llama guanicoe*). Complete sequence coverage for *C. hesternus* was obtained via PMF and tandem mass spectrometry. This marks the second complete osteocalcin sequence from a fossil found in a temperate climate zone and the first molecular data for the genus *Camelops*. The data indicate that there is no sequence difference among the ancient and modern camelid species.

Modifications to the primary structure of osteocalcin were observed in both modern and ancient specimens. A modification of Met_{36} to Met-ox_{36} was observed in *C. hesternus* and *C. dromedarius*. Met, like Cys and Trp, is one of the most easily oxidized residues in proteins. Oxidation occurs when environmental or biological oxidants interact with the sulfur atom in methionine, creating methionine sulfoxide (Vogt, 1995). This modification is reversed in-vivo, and is hypothesized to regulate protein function. It is also possible for the oxidation of methionine to occur during diagenesis or sample preparation. The presence of EDTA and iron can greatly enhance the oxidation of methionine, especially in an acidic environment. Thus, conclusions about the depositional environment of the *C. hesternus* specimen cannot be inferred based on this modification. Despite this, the presence of Met_{36} is notable. The side chain of Met is hydrophobic and its oxidation may influence the hydrophobicity of camelid osteocalcin. Insofar as many functions attributed to osteocalcin have not been unequivocally proven (e.g. chemotaxic activity of the C-terminus), we do not know if oxidation restricts the function of the protein in-vivo (Mundy and Poser, 1983; Vogt, 1995: Dowd et al., 2003).

For example if Met is involved in binding a partner and enhances the interaction by holding together other hydrophobic groups its oxidation may result in loss of binding and/or functionality (Vogt, 1995).

C. hesternus also exhibits deamidation of Gln_{35} and Gln_{39}. Deamidation is the hydrolyzation of the side chain amide linkage to form a free carboxylic acid (Geiger and Clarke, 1987). The optimal environment for deamidation is 37 degrees Celsius, pH 7.4. Consequently, deamidation in fossil material may occur in relatively warm environments. Although deamidation could result from sample handling, it was not observed in modern camelids or modern horse osteocalcin (Ostrom et al., 2006). Deamidation was observed in fossil horse (Ostrom et al., 2006), which was recovered from a temperate environment, but not observed in fossil bison (Nielsen-Marsh, 2002), obtained from permafrost.

Carbamylation was observed in osteocalcin of *C. hesternus* and previously in osteocalcin of a 42 ka horse but not in modern osteocalcin. Carbamylation of an amino group by isocyanic acid is initiated by urea (Stark, 1965). As urea is derived from urine, its presence in cave environments is not unexpected. When carbamylation reacts with the amino terminus of proteins it can block N-terminal sequencing and may be the cause of the N-terminal blockage traditionally identified in ancient proteins (Robbins et al., 1993).

Knowledge of post-translational and diagenetic modifications is essential for understanding protein sequence, structure, function, and, ultimately, the evolution of the protein. Although the phylogeny of camelids cannot be determined due to a lack of sequence differences among the camelids analyzed here, evolutionary relationships between camelids and other artiodactyls can be explored. Five positions can be identified as phylogenetically informative in artiodactyls when they are evaluated within the

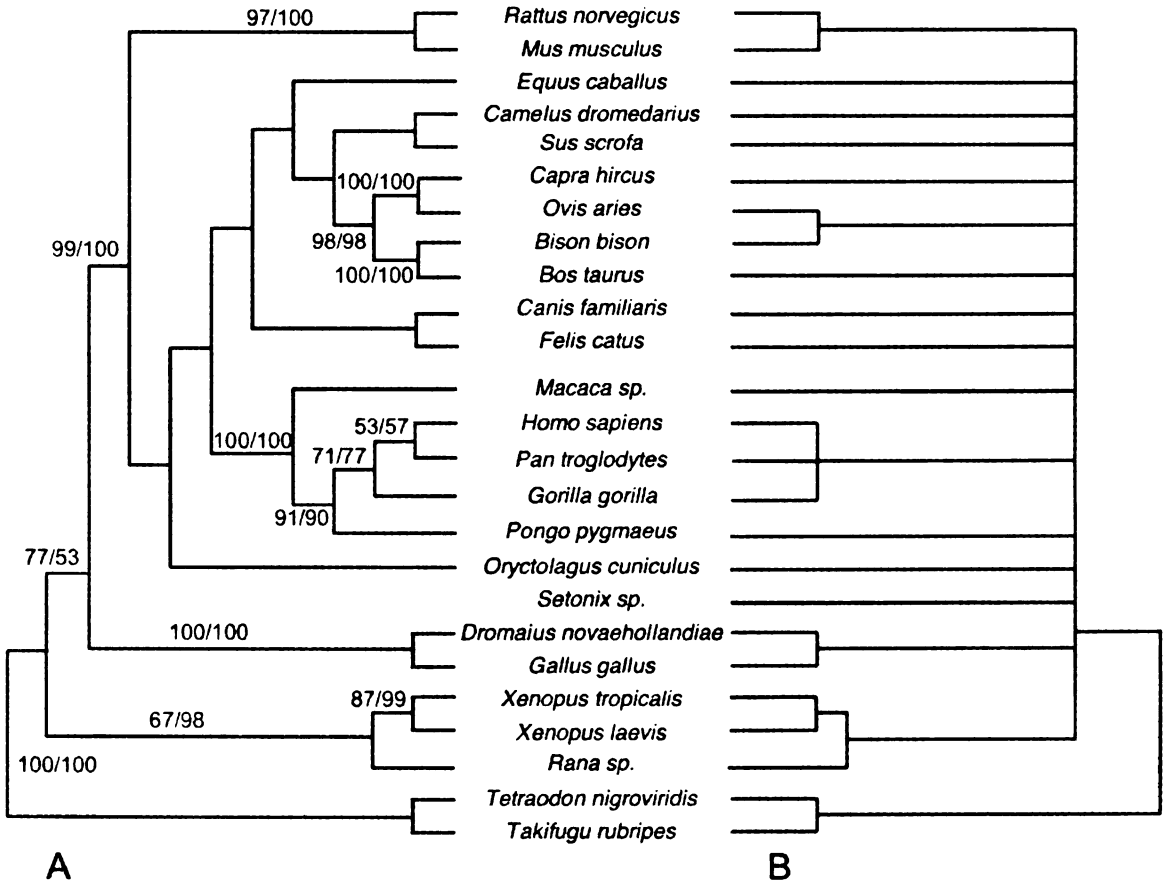


Figure 4. Maximum Likelihood trees for (A) Cytochrome b-coding region of mtDNA and (B) osteocalcin protein. Values above/below branches on Cytochrome b tree are Maximum Parsimony Bootstrap/ Neighbor Joining Bootstrap values. Cytochrome b Maximum Likelihood tree created using PAUP* 4.0b10 with model generated using Modeltest 3.7. Dotted lines indicate taxa which do not possess sequenced Cytochrome b, and have been placed within the tree based on known taxonomic position. Osteocalcin Maximum Likelihood tree created using Tree Puzzle 5.0. Maximum Parsimony and Neighbor Joining bootstrap values obtained using PAUP* 4.0b10. Tree length for tree A = 3267, C.I. = 0.350, R.I. = 0.406. Tree length for tree B = 148, C.I. = 0.655, R.I. = 0.523. Uninformative characters were excluded.

context of the cytochrome b tree structure (Figure 4 A, Figure 5). These characters are located at positions 4, 5, 19, 48, and 49 and exist near the N- or C- terminus. Both of these regions are known to be either variable or less constrained than the alpha-helical regions (Hauschka et al., 1989; Hoang et al.2003).

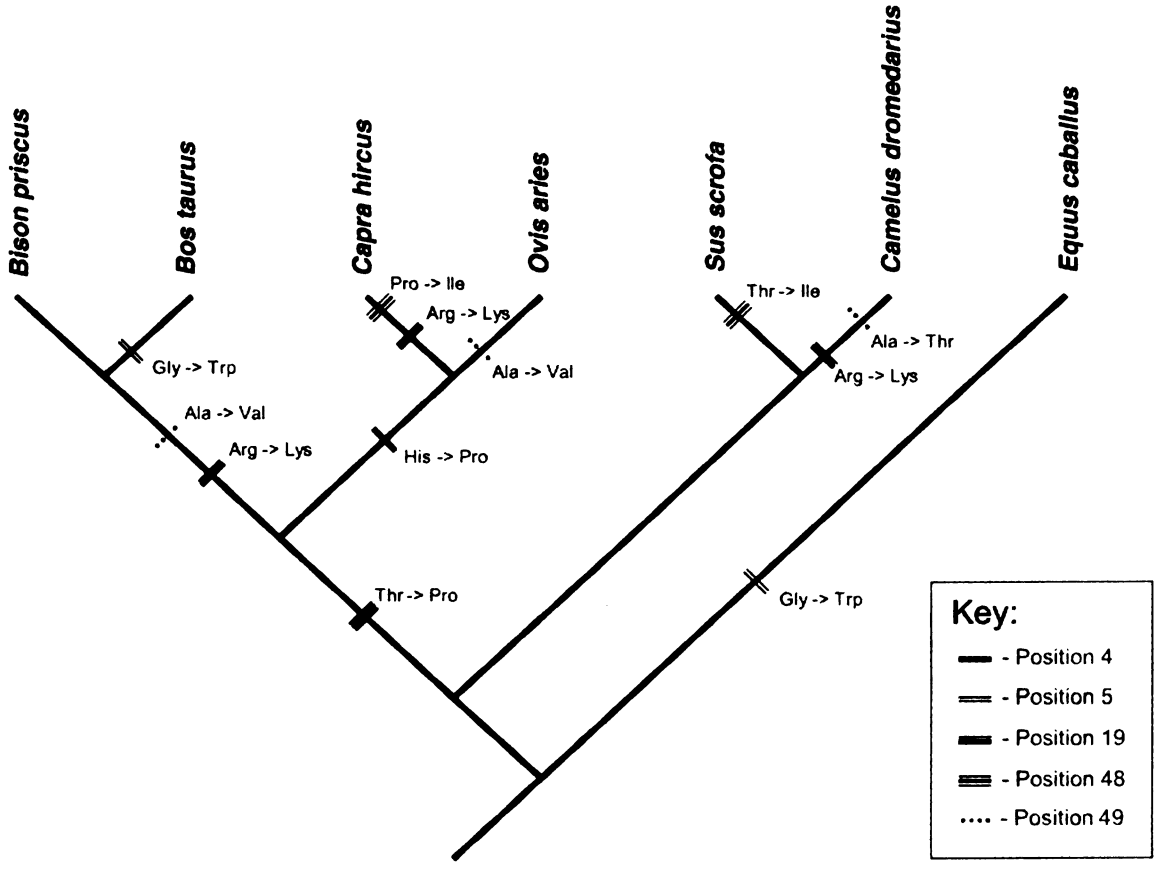


Figure 5. Map of five informative characters within osteocalcin for known artiodactyl sequences, showing four out of the five characters are convergent. *E. caballus* is used as an outgroup. Tree topology was taken from Maximum Likelihood tree of Cytochrome b sequences created using PAUP*4.0b10, and the character evaluation is dependent on the structure of this tree. Characters mapped using MacClade 4.06. Ala was chosen as the ancestral character in the *O. aries – C. hircus* clade for character 48.

By examining each of these characters individually, the evolutionary relationships of the artiodactyls can be inferred. In position 4, Pro defines the *Ovis aries – Capra hircus* clade, with all other species containing His. Within position 5 Trp is present in both *Bos taurus* and *Equus caballus*, with all other species containing Gly. As Gly is the ancestral state, Trp must have evolved at least once within both the artiodactyls and the outgroup (*E. caballus*). Position 19 contains Lys and Arg, with Arg being the ancestral state in this tree. *C. dromedarius, C. hircus, B. taurus, and Bison bison* all contain Lys, indicating

that Lys evolved independently at least twice within artiodactyls. Position 48 contains the amino acids Ile, Pro, and Thr, with Thr as the ancestral state. In this tree, Thr is present in *C. dromedarius* and in *E. caballus*, evolving to Pro in *O. aries*, *B. taurus*, and *B. bison*. Ileu evolves twice within the tree; once in *S. srofa*, and once within *C. hircus*. Ala, Thr, and Val are present within position 49, with Ala as the ancestral state. Ala is present in *E. caballus*, *Sus srofa*, and *C. hircus*, while Val is present within *O. aries*, *B. Taurus*, and *B. bison*. The ancestral character state in the *C. hircus – O. aries* clade is ambiguous, but the presence of both Ala and Val within that clade indicates that one or the other amino acids had to evolve twice within the artiodactyls. *C. dromedarius* is the only taxon to include Thr. Four of the five characters that define the artiodactyl clade exhibit homoplasy in two or more species, making these characters unsuitable for phylogenetic placement of taxa. Position 19 is especially homoplastic, with Arg evolving in three separate instances. The high amount of homoplasy within only a small number of characters can cause conflicting relationships within phylogenetic analyses, and has the potential to cause improper taxonomic grouping or create ambiguous character relationships. The single character that does unambiguously define a clade is position 4. However, the *O. aries – C. hircus* clade is not supported by the other characters due to homoplasy. The homoplasy within poition 4 can overwhelm the relationship between *O. aries* and *C. hircus*, causing it to be suppressed.

To examine the relationship of the artiodactyls, including the camelids, within mammalia and determine what level of effect homoplasy has on the phylogeny, an analysis of all known tetrapod osteocalcin sequences was condcuted (Figure 4 B). This results in a tree that exhibits some lower level taxonomic groupings. Both rodents are

included within a single clade, as are both members of aves. The hominins are also included within a single clade, although the exact relationships within homininae are not defined (a polytomy joins *Pan troglodytes*, *Homo sapiens*, *Gorilla gorilla*). A clade containing *B. taurus* and *O. aries* is also indicated by the tree. Although they are both artiodactyls, the comparison tree indicates that they are more closely allied with *B. bison* and *C. hircus*, respectively. The grouping of these two taxa is most likely due to the homoplasy exhibited within the artiodactyls, as previously mentioned. It is possible that incorrect groupings due to homoplasy did not occur in other taxa because of the limited number of individuals representing those taxa. Of the known osteocalcin sequences, artiodactyls are represented by six separate species. Nearly all others (aves, amphibia, carnivora, perisodactyla, rodentia, lagomorpha, and marsupialia) are represented by three or fewer species. Only the primates have a similar number of species represented, and nearly all are contained within a single subfamily grouping. All four clades mentioned above and all other taxa are included in a single basal polytomy. The polytomy is due to the large conserved region within osteocalcin, which limits the number of informative characters, as well as producing the high degree of homoplasy within the informative characters, as exampled by the artiodactyls.

A large disparity in the level of taxonomic resolution between the amino acid and gene sequences was observed. Rate heterogeneity is an inherent variant in phylogenetics (Meyer, 1994). While the more rapidly evolving codons of the mtDNA genes make them useful for comparisons at low taxonomic levels, the conserved nature of osteocalcin makes its amino acid sequence more amenable to deep divergences (Hauschka et al., 1989; Meyer, 1994). But even this level of resolution is not observed in the osteocalcin-

based tree. Many of the highly conserved characters are completely uninformative phylogenetically, either due to a total lack of mutation or very few single-taxon substitutions. A high degree of homoplasy is observed within characters that are considered informative. This is especially true for the artiodactyls. Further work must be done towards sequencing more taxa, especially among reptiles, birds, and the orders Carnivora and Perissodactyla. The preliminary analysis of osteocalcin evolution provided here is an initial step toward a complete character analysis with the goal to determine the exact nature of the evolution of osteocalcin.

BIBLIOGRAPHY

Barnes, I., P. Matheus, B. Shapiro, D. Jensen, and A. Cooper. 2002. Dynamics of Pleistocene population extinctions in Beringian brown bears. Science 295:2267-2270.

Cancela ML, Williamson MK, and P. PA. 1995. Amino-acid sequence of bone Gla protein from the African clawed toad *Xenopus laevis* and the fish Sparus aurata. International Journal of Peptide and Protein Research 46:419-423.

Collins, M. J., A. M. Gernaey, C. M. Nielsen-Marsh, C. Vermeer, and P. Westbroek. 2000. Slow rates of degradation of osteocalcin: Green light for fossil bone protein? Geology 28:1139-1142.

Dowd, T. L., J. F. Rosen, L. Li, and C. M. Gundberg. 2003. The three-dimensional structure of bovine calcium ion-bound osteocalcin using HNMR spectroscopy. Biochemistry 42:7769-7779.

Frazao, C., D. C. Simes, R. Coelho, D. Alves, M. K. Williamson, P. A. Price, M. L. Cancela, and M. A. Carrondo. 2005. Structural Evidence of a Fourth Gla Residue in Fish Osteocalcin: Biological Implications. 44:1234-1242.

Geiger, T., and S. Clarke. 1987. Deamidation, Isomerization, and Racemization at Asparaginyl and Aspartyl Residues in Peptides - Succinimide-Linked Reactions That Contribute to Protein-Degradation. Journal of Biological Chemistry 262:785-794.

Harris, A. H., and J. S. Findley. 1964. Pleistocene-Recent Fauna of Isleta Caves Bernalillo County New Mexico. American Journal of Science 262:114-120.

Hauschka, P. V., J. B. Lian, D. E. C. Cole, and C. M. Gundberg. 1989. Osteocalcin and Matrix Gla Protein - Vitamin K-Dependent Proteins in Bone. Physiological Reviews 69:990-1047.

Hoang, Q. Q., F. Sicheri, A. J. Howard, and D. S. C. Yang. 2003. Bone recognition mechanism of porcine osteocalcin from crystal structure. Nature 425:977-980.

Hofreiter, M., D. Serre, H. N. Poinar, M. Kuch, and S. Paabo. 2001. Ancient DNA. Nature Reviews Genetics 2:353-359.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The Rapid Generation of Mutation Data Matrices from Protein Sequences. Computer Applications in the Biosciences 8:275-282.

Kurten, B., and E. Anderson. 1980. Pleistocene mammals of North America. Columbia University Press, New York.

Laize, V., C. S. B. Viegas, P. A. Price, and M. L. Cancela. 2006. Identification of an Osteocalcin Isoform in Fish with a Large Acidic Prodomain. Pages 15037-15043.

Lange, I. M. 2002. Ice Age mammals of North America : a guide to the big, the hairy, and the bizarre. Mountain Press Pub. Co., Missoula, Montana.

Lian, J., C. Stewart, E. Puchacz, S. Mackowiak, V. Shalhoub, D. Collart, G. Zambetti, and G. Stein. 1989. Structure of the Rat Osteocalcin Gene and Regulation of Vitamin D-Dependent Expression. Pages 1143-1147.

Maddison, D. R., and W. P. Maddison. 2000. MacClade version 4 Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Massachusetts.

McNulty, T., A. Calkins, P. Ostrom, H. Gandhi, M. Gottfried, L. Martin, and D. Gage. 2002. Stable isotope values of bone organic matter: Artificial diagenesis experiments and paleoecology of Natural Trap Cave, Wyoming. Palaios 17:36-49.

Meyer, A. 1994. Shortcomings of the cytochrome b gene as a molecular marker Trends in Ecology and Evolution 9:278-280.

Mundy, G. R., and J. W. Poser. 1983. Chemotactic Activity of the Gamma-Carboxyglutamic Acid Containing Protein in Bone. Calcified Tissue International 35:164-168.

Nielsen-Marsh, C. M., P. H. Ostrom, H. Gandhi, B. Shapiro, A. Cooper, P. V. Hauschka, and M. J. Collins. 2002. Sequence preservation of osteocalcin protein and mitochondrial DNA in bison bones older than 55 ka. Pages 1099-1102.

Nielsen-Marsh, C. M., M. P. Richards, P. V. Hauschka, J. E. Thomas-Oates, E. Trinkaus, P. B. Pettitt, I. Karavanic, H. Poinar, and M. J. Collins. 2005. Osteocalcin protein sequences of Neanderthals and modern primates. Proceedings of the National Academy of Sciences of the United States of America 102:4409-4413.

Ostrom, P. H., M. Schall, H. Gandhi, T. L. Shen, P. V. Hauschka, J. R. Strahler, and D. A. Gage. 2000. New strategies for characterizing ancient proteins using matrix-assisted laser desorption ionization mass spectrometry. Geochimica Et Cosmochimica Acta 64:1043-1050.

Ostrom, P. H., H. Gandhi, J. R. Strahler, A. K. Walker, P. C. Andrews, J. Leykam, T. W. Stafford, R. L. Kelly, D. N. Walker, M. Buckley, and J. Humpula. 2006. Unraveling the sequence and structure of the protein osteocalcin from a 42 ka fossil horse. Geochimica Et Cosmochimica Acta 70:2034-2044.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Robbins, L. L., G. Muyzer, and K. Brew. 1993. Macromolecules form living and fossil biominerals; Implications for the establishment of molecuar phyolgenies. In

Organic Geochemistry M. H. Engle and S. A. Macko, pp. 799 - 816. Plenum. Pages 799-816 *in* Organic Geochemistry: Principles and Applications (M. H. Engle, and S. A. Macko, eds.). Plenum Press, New York.

Savage, D. E. 1951. Late Cenozoic Vertebrates of the San Francisco Bay Region. Univ. Calif. Public. Geol. Sci. 18:215-314.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504.

Stark, G. R. 1965. Reactions of cyanate with functional groups of proteins. Inertness of aliphatic hydroxyl groups. Formation of carbamyl- and acylhydantoins. Biochemistry 4:2363-2367.

Swofford, D. L. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods). Version 4.0 Beta. Sinauer Associates, Sunderland, Massachusetts.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25:4876-4882.

Vogt, W. 1995. Oxidation of Methionyl Residues in Proteins - Tools, Targets, and Reversal. Free Radical Biology and Medicine 18:93-105.

Wayne, R. K., J. A. Leonard, and A. Cooper. 1999. Full of sound and fury: The recent history of ancient DNA. Annual Review of Ecology and Systematics 30:457-477.

Webb, S. D. 1974. Pleistocene llamas of Florida, with a brief review of the Lamini Pages 170-214 *in* Pleistocene mammals of Florida (S. D. Webb, ed.) University Press of Florida, Gainsville, Florida.

Zardoya, R., and A. Meyer. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Molecular Biology and Evolution 13:933-942.