# Evolution of Matrix and Bone γ-Carboxyglutamic Acid Proteins in Vertebrates*

### Vincent Laizé‡§, Paulo Martel¶‖, Carla S. B. Viegas‡, Paul A. Price**, and M. Leonor Cancela‡

*From the ‡Centro de Ciências do Mar (CCMAR), Universidade do Algarve, 8005-139 Faro, Portugal, ¶Instituto de Tecnologia Química e Biológica (ITQB), Universidade Nova de Lisboa, 2781-901 Oeiras, Portugal, ‖Centro de Biomedicina Molecular e Estrutural (CBME), Universidade do Algarve, 8005-139 Faro, Portugal, and **Division of Biology, University of California San Diego, La Jolla, California 92093-0368*

The evolution of calcified tissues is a defining feature in vertebrate evolution. Investigating the evolution of proteins involved in tissue calcification should help elucidate how calcified tissues have evolved. The purpose of this study was to collect and compare sequences of matrix and bone γ-carboxyglutamic acid proteins (MGP and BGP, respectively) to identify common features and determine the evolutionary relationship between MGP and BGP. Thirteen cDNAs and genes were cloned using standard methods or reconstructed through the use of comparative genomics and data mining. These sequences were compared with available annotated sequences (a total of 48 complete or nearly complete sequences, 28 BGPs and 20 MGPs) have been identified across 32 different species (representing most classes of vertebrates), and evolutionarily conserved features in both MGP and BGP were analyzed using bioinformatic tools and the Tree-Puzzle software. We propose that: 1) MGP and BGP genes originated from two genome duplications that occurred around 500 and 400 million years ago before jawless and jawed fish evolved, respectively; 2) MGP appeared first concomitantly with the emergence of cartilaginous structures, and BGP appeared thereafter along with bony structures; and 3) BGP derives from MGP. We also propose a highly specific pattern definition for the Gla domain of BGP and MGP.

BGP[1] (bone Gla protein or osteocalcin) and MGP (matrix Gla protein) belong to the growing family of vitamin K-dependent (VKD) proteins, the members of which are involved in a broad range of biological functions such as skeletogenesis and bone maintenance (BGP and MGP), hemostasis (prothrombin, clotting factors VII, IX, and X, and proteins C, S, and Z), growth control (gas6), and potentially signal transduction (proline-rich Gla proteins 1 and 2). VKD proteins are characterized by the presence of several Gla residues resulting from the post-translational vitamin K-dependent γ-carboxylation of specific glutamates, through which they can bind to calcium-containing mineral such as hydroxyapatite. To date, VKD proteins have only been clearly identified in vertebrates (1) although the presence of a γ-glutamyl carboxylase has been reported in the fruit fly *Drosophila melanogaster* (2) and in marine snails belonging to the genus *Conus* (3). Gla residues have also been found in neuropeptides from *Conus* venoms (4), suggesting a wider prevalence of γ-carboxylation.

Bone and matrix Gla proteins have essential roles in controlling tissue mineralization and form a subgroup within the VKD family. BGP and MGP have low similarity with blood coagulation factors, although they are believed to have diverged from a common ancestor (5–7). Although a number of questions still remain concerning their mode of action at the molecular level, a large number of studies (Refs. 15–18, 26, 27, 31, and references therein) have clearly shown that BGP and MGP are central to the control of tissue mineralization mechanisms.

MGP is a 10-kDa protein produced and secreted by vascular smooth muscle cells (8) and chondrocytes (9) and significantly accumulated in bone, cartilage, and dentin from mammals, amphibians, and cartilaginous and bony fishes (10–14). MGP gene expression is confined to proliferative and late hyperthrophic chondrocytes within the mammalian growth plate during endochondral bone formation and has been therefore described as a marker of the chondrogenesis cell lineage (15). A signal peptide, a phosphorylation domain, a γ-carboxylase recognition site, and 4–5 Gla residues are the conserved features among MGPs. The spontaneous calcification of arteries and cartilage in mouse lacking MGP or the abnormal cartilage and artery calcification in patients affected by the Keutel syndrome (autosomal recessive disorder caused by mutations in the MGP gene) indicates that this protein is a physiological inhibitor of mineralization (16–18). Additional data suggest that MGP is involved in protecting tissues from ectopic calcification in mammals (17, 19) and controlling cartilage cell differentiation and mineralization during early development of chicken limbs (9). *In vitro* cell culture experiments have shown that MGP gene expression can be regulated by $1,25\text{-(OH)}_2$ vitamin $D_3$ (20), retinoic acid (21, 22), extracellular calcium (23, 24), growth factors, and cell proliferation events (25), but the mechanisms responsible for the transcriptional regulation of MGP still remain largely unknown.

[1] The abbreviations used are: BGP, bone Gla protein; MGP, matrix Gla protein; VKD, vitamin K-dependent; EST, expressed sequence tag; aa, amino acid; Myr, million years; PAM, point accepted mutations; GGCX, γ-glutamyl carboxylase.
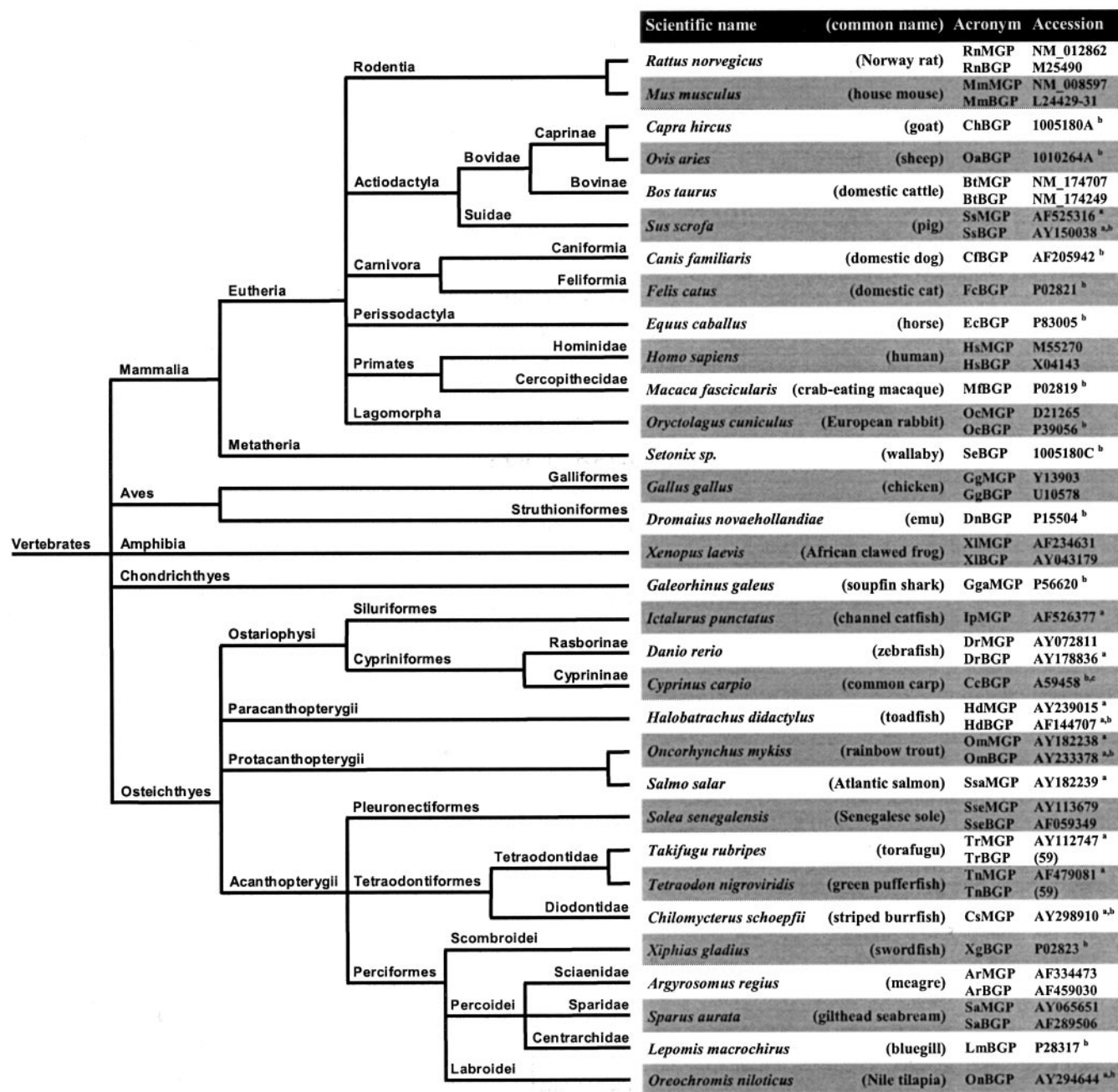
Fig. 1. **BGP and MGP sequences used in this study and taxonomy of represented species.** Taxonomic data were retrieved September 14, 2004 from the Integrated Taxonomic Information System at www.itis.usda.gov. *a*, this study; *b*, partial sequence; *c*, Protein Information Resource accession number.

BGP is a 5.6-kDa secreted protein that, unlike MGP, has proven to be specific for vertebrate calcified tissues, with the exception of cartilaginous fishes, and represents 1–2% of the total bone protein. BGP is synthesized by osteoblasts and odontoblasts. All known BGPs are synthesized as pre-pro-proteins and contain 3 Gla residues located within a conserved domain in the central part of the mature protein. Previous studies have failed to reveal a critical role for BGP in mammalian bone formation. BGP-deficient mice exhibit a higher bone mass when compared with control (26), and IR microspectroscopy has shown that BGP is required for the correct maturation of the hydroxyapatite in mammalian bone (27). The tertiary structures of bovine (28), porcine (29), and meagre (30) BGP have recently been determined by x-ray crystallography (porcine and meagre) and NMR (bovine). Despite some differences, the three structures are very similar, with a novel fold com-

prising a C-terminal region with the 3 helices and a small hydrophobic core and a short, unstructured N-terminal sequence. The 3 Gla residues (4 in meagre) are on the outer face of helix 2, coordinating the $Ca^{2+}$ ions present in the structures and suggesting a mechanism for attachment to the surface of the hydroxyapatite crystals in bone (29). In the absence of $Ca^{2+}$, the apoprotein form of BGP is in a disorganized state, as indicated by NMR studies (28). Interestingly, re-expression of MGP in arteries of MGP−/− mice can reverse the arterial mineralization, a result that cannot be mimicked by BGP indicating that these evolutionarily related proteins do not work in a similar way (31).

Understanding more clearly the evolution of BGP and MGP will contribute to better evaluation of the various hypotheses about their role and function in tissue mineralization. It should also help illuminate how calcified tissues have evolved and

TABLE I
*List of PCR primers used in this study*
Hd, *Halobatrachus didactylus*; Cs, *Chilomycterus schoepfii*; On, *Oreochromis niloticus*; Dr, *Danio rerio*.

| DNA amplified | | Primers (5' to 3') | Tm |
|---|---|---|---|
| **cDNA** | | | |
| HdBGP | 3'end [a] | SBG5F – TGCGAGCACATGATGGACACTGAGGGAATC (7) | 70°C |
| | | SBG4R – GTTAGGGGAAATGATCGAATCACAGTGGG (7) | 65°C |
| | 5'end | AP 1 – GTAATACGACTCACTATAGGGC (Clontech) | 45°C |
| | | HdBGP1R – GCGGTCTAAAAGGGAATCGGCCCGT | 68°C |
| HdMGP | 3'end | ArMGP3F – AGGCGTGCAGAGACCTGCGAG (12) | 61°C |
| | | Universal adapter – ACGCGTCGACCTCGAGATCGATG | 64°C |
| | Internal | HdMGP1F – GGTCTCGGTCTGCCTCTGCTATGA | 61°C |
| | | ArMGP2R – GCNAAYWSNTTYATGACNCCNCA [b] (12) | 58°C |
| | 5'end | AP 1 – GTAATACGACTCACTATAGGGC (Clontech) | 45°C |
| | | ArMGP1R – CGCTGGAGACTTCACCGTCCTCATC (12) | 64°C |
| CsMGP | 3'end | DsMGP-03F – TTGTGTCTGCTATGATGACTCTGATGA | 57°C |
| | | DsMGP-04R – TGGAAAGTAATCCTCACACGTCTCTGC | 61°C |
| OnBGP | 3'end | TilBGP-01F – GCTTGTGAGCACATGATGGACATAAA | 59°C |
| | | Universal adapter – ACGCGTCGACCTCGAGATCGATG | 64°C |
| | Internal | SBG13F – CCAGCCTGCCAGTGACAACCC (7) | 61°C |
| | | TilBGP-03R – TTTATGTCCATCATGTGCTCACAAGC | 59°C |
| **Gene** | | | |
| DrBGP | Exons 1 to 3 | zBGP8F – ATCTTCTGCTGCCTGATGACT | 51°C |
| | | zBGP2R – CCACGTCACGCTTCACAAACA | 57°C |

[a] J. P. Pinto, personal communication.
[b] Y, C or T; W, A or T; S, G or C.

FIG. 2. **Structural organization of MGP (*A*) and BGP (*B*) coding sequences at the gene level.** *Black boxes* indicate exons (or parts of exons) representing the coding sequence, starting from the translation initiation codon and ending at the translation termination codon. *Numbers above* the *boxes* indicate their length. *Dashed line* in zebrafish MGP sequence indicates incomplete data on intron size, although total length of the genomic fragment is known. Data on each gene of the mouse BGP gene cluster are also presented.
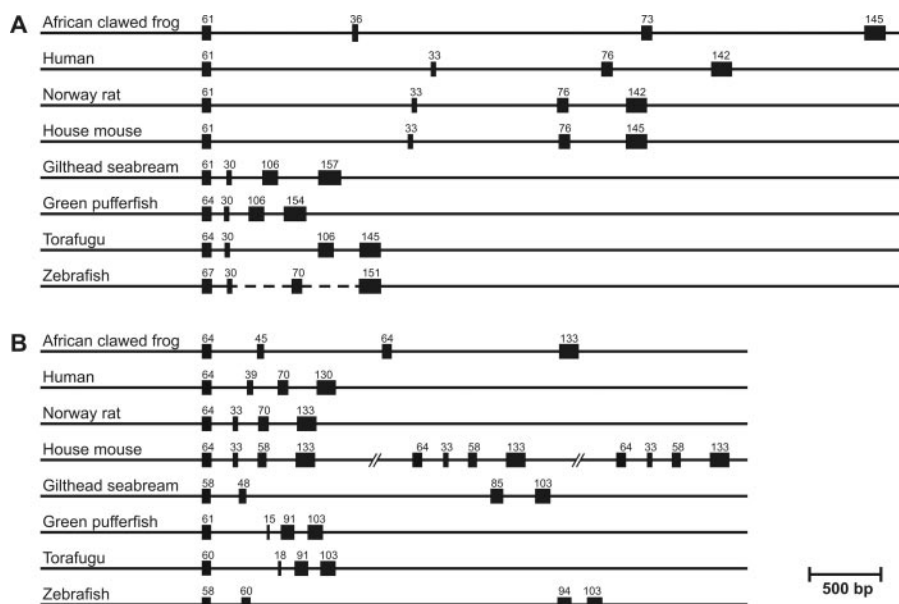
provide key insights into vertebrate evolution. Specific goals of this work were: (i) to identify new MGP and BGP homologues using standard cloning methods or comparative genomics and data mining of available genomic and EST (expressed sequence tag) libraries, (ii) to provide insight into MGP and BGP common features by looking at similarities among gene and protein sequences, and (iii) to infer particular trends in BGP and MGP evolutionary history through a phylogenetic analysis of all available amino acid sequences. Finally, we propose a model for the early origin and evolution of BGP and MGP on the basis of results presented here and in previous studies (Refs. 7, 12–14, and references therein).

### EXPERIMENTAL PROCEDURES

*BGP and MGP Sequence Collection*—Previously characterized sequences were obtained from the GenBank™ (www.ncbi.nlm.nih.gov) and Swiss-Prot (us.expasy.org) data bases. Additional sequences were either cloned using standard PCR techniques or reconstructed from EST and whole genome shotgun sequences obtained from public sequence data bases. dbEST and Trace Archive were extensively searched using the BLAST facilities at NCBI for sequences showing similarities to known BGP and MGP transcripts or genes. Species-specific sequences were first clustered, and elements of each cluster were assembled using Clustal X (32) to generate, after manual correction, highly accurate consensus sequences. Virtual transcripts and genes were deduced from joined consensus sequences using stringent overlap criteria. Virtual gene structure was predicted using comparative methods (homology to previously annotated genes) and electronic splicing as predicted by GenScan. Genomic DNA (prepared with Qiagen DNeasy tissue kit), cDNA libraries (prepared with Clontech Marathon cDNA amplification kit), or reverse transcribed RNA (prepared using Invitrogen Moloney murine leukemia virus reverse transcriptase) were used for PCR amplification. Genomic and cDNA fragments were amplified in a GeneAmp 2400 thermal cycler (PerkinElmer Life Sciences) using *Taq* DNA polymerase (Invitrogen) and the primer sets described in Table I. Following an initial denaturation step of 5 min at 95 °C, specific DNAs were amplified with 30 cycles (one cycle is 30 s at 95 °C, 45 s at $T_m$ (see Table I), and 1 min at 72 °C) and a final elongation step of 10 min at 72 °C. The resulting PCR products were size-fractionated by agarose gel electrophoresis, purified, and ligated into pGEM-T Easy vector (Promega). Final identification was achieved by DNA sequence analysis.

*Determination of Sparus aurata Genome Size*—Blood was collected

from a healthy specimen of gilthead seabream (*S. aurata*), smeared on a microscope slide, and subjected to genome sizing according to Hardie *et al.* (33).

*Sequence Alignment and Analysis*—Separate alignments for MGP and BGP sequences were created using Clustal X and the Gonnet250 mutation matrix, a gap-opening penalty of 10, and a gap penalty of 0.2. Clustal X alignments were fed to the more rigorous T-Coffee multiple sequence alignment software (34) with parameters set to the default. Manual adjustments were made in a few cases to improve alignments. An alignment of the whole set of MGP and BGP sequences was produced using a similar procedure. Because there was no recognizable similarity between the MGP and the BGP sequences outside the conserved Gla region, only the latter part of the alignment was kept. Sequence logos presented in Fig. 3 were created from the T-Coffee multiple sequence alignments using the WebLogo facilities (35). The sequence logos are presented as a graphical display where the height of each letter is made proportional to its frequency. This shows the conserved residues as larger characters. Putative signal peptide and phosphorylation sites were identified in protein sequences using the SignalP and NetPhos facilities, respectively (29, 30).

*Sequence Identity*—Pairwise sequence identity values were computed as percent of identical residues over the total number of aligned residues using alignments generated with T-Coffee.

*Phylogenetic Analysis*—All maximum likelihood phylogenetic trees were built from T-Coffee amino acid alignments using the Quartet puzzling algorithm of the Tree-Puzzle software (36). The PAM250 mutation data matrix was chosen because it produced the smallest number of unresolved quartets in all cases. A total number of 100,000 puzzling steps was used, and the rate of change was taken as site-independent (the use of a $\gamma$ distributed variable rate of change among sites was tried and produced worse results in all cases). Test runs were also performed with Proml from the Phylip suite of programs (37). Constraint phylogenetic trees were generated using TreeView (38), where the internal branch labels, which are an estimate of branch assignment reliability, are Tree-Puzzle branch support values and must not be confused with the bootstrap values produced by other programs, *e.g.* Phylip.

*Calculation of Evolutionary Rates*—If a clock-like behavior for evolution is assumed, approximate rates of evolution for molecular sequences can be calculated from sequence distances, provided that divergence times between taxa are known from the fossil record (39). Distances between groups were calculated from the T-Coffee BGP and MGP multiple sequence alignments using the Grishin formula to correct for multiple substitutions and rate variation among sites (40). The mammalian-avian split (310 Myr ago), tetrapod-fish split (405 Myr ago), mammalian-amphibian split (200 Myr ago), radiation of teleosts (430 Myr ago), and radiation of mammals (100 Myr ago) (41–43) were used to convert evolutionary distances in PAM (point accepted mutations per 100 aa) per Myr. The final rates were averages of the values obtained with the above times. The clock-like assumption was checked with the Tajima test (44) and accepted at the 95% level for all triplets of sequences.

## RESULTS

*Cloning and Identification of New BGP and MGP Sequences*—To increase the likelihood of identifying all available BGP and MGP sequences from the vast amount of sequence data, various homologues were used as query sequences to search public sequence data bases. Positive search results were considered to be members of the BGP/MGP family if they showed significant sequence similarity to any of the previously identified members and if this similarity extended throughout the protein. Numerous sequences previously identified as BGP or MGP were collected, and their accession numbers or reference number (59) in the literature is indicated in Fig. 1. Our search within public data bases also identified many ESTs and whole genome shotgun sequences with similarities to previously characterized sequences from which we could reconstruct 6 cDNAs (pig BGP and MGP, rainbow trout BGP and MGP, Atlantic salmon MGP, and channel catfish MGP) and 2 genes (Torafugu MGP and green pufferfish MGP). Finally, 4 cDNAs (toadfish BGP and MGP, striped burrfish MGP, and Nile tilapia BGP) and 1 gene (zebrafish BGP) were amplified by reverse transcription-PCR from RNAs prepared from calcified tissues or by

TABLE II
*Phase and length of introns present in BGP and MGP coding region*

Phase of intron is defined according to Patthy (46) and is indicated in parentheses (ii) intron lying between the first and second nucleotides of a codon (phase 1 intron); (iii) intron lying between the second and third nucleotides of a codon (phase 2 intron). Intron sizes are in base pairs.

| BGP | Intron 1 | Intron 2 | Intron 3 |
|---|---|---|---|
| African clawed frog | 318 (ii) | 827 (ii) | 1165 (iii) |
| Human | 257 (ii) | 170 (ii) | 201 (iii) |
| Norway rat | 148 (ii) | 143 (ii) | 200 (iii) |
| House mouse | 144 (ii) | 141–142 (ii) | 206 (iii) |
| Gilthead seabream | 197 (ii) | 1713 (ii) | 221 (iii) |
| Green pufferfish | 390 (ii) | 77 (ii) | 96 (iii) |
| Torafugu | 467 (ii) | 96 (ii) | 88 (iii) |
| Zebrafish | 257 (ii) | 2152 (ii) | 105 (iii) |
| MGP | Intron 1 | Intron 2 | Intron 3 |
| African clawed frog | 986 (ii) | 1985 (ii) | 1490 (iii) |
| Human | 1485 (ii) | 1165 (ii) | 691 (iii) |
| Norway rat | 1409 (ii) | 984 (ii) | 406 (iii) |
| House mouse | 1375 (ii) | 1023 (ii) | 390 (iii) |
| Gilthead seabream | 114 (ii) | 219 (ii) | 283 (iii) |
| Green pufferfish | 91 (ii) | 133 (ii) | 142 (iii) |
| Torafugu | 98 (ii) | 617 (ii) | 179 (iii) |
| Zebrafish | 112 (ii) | >194 (ii) | >677 (iii) |

TABLE III
*Comparison of genome and total intron size*

| Species | Haploid C-value (pg) | Intron total size [e] | |
|---|---|---|---|
| | | BGP | MGP |
| African clawed frog | 3.30 ± 0.31 [a,b] | 2310 | 4461 |
| Human | 3.50 [b] | 628 | 3341 |
| Norway rat | 3.36 ± 0.40 [a,b] | 491 | 2799 |
| House mouse | 3.28 ± 0.38 [a,b] | 491–492 | 2788 |
| Gilthead seabream | 1.36 [c] | 2131 | 616 |
| Green pufferfish | 0.35 [d] | 563 | 366 |
| Torafugu | 0.40 [b] | 651 | 894 |
| Zebrafish | 1.86 ± 0.24 [a,b] | 2514 | >983 |

[a] Mean value ± S.D.
[b] From www.genomesize.com.
[c] This study.
[d] From Lamatsch *et al.* (58).
[e] In base pairs.

PCR from genomic DNA using the oligonucleotides listed in Table I. The identity of the amplified fragments was confirmed by sequence comparison with previously annotated sequences. Rhesus macaque MGP (*Macaca mulatta*, GenBank™ accession number AF162477) and steppe bison BGP (*Bison priscus*, Swiss-Prot accession number P83489) were not included in this study because the sequences were too small or identical to domestic cattle BGP, respectively. No evidence for the presence of any pseudogenes or alternatively spliced transcripts was found. The GenBank™/Swiss-Prot accession numbers of all osteocalcin and matrix Gla protein sequences used in this study are listed in Fig. 1.

*Taxa Represented*—A total of 48 complete or nearly complete sequences (28 BGPs and 20 MGPs) have been identified from 32 different species representing most classes of vertebrates (and only in vertebrates). These include mammals, birds, amphibians, bony fish, and cartilaginous fish (Fig. 1). Only one homologue of MGP and BGP was identified per species with one exception (the mouse, with a cluster of three BGP-related genes (45)). No homologues were found in archaea, bacteria, viruses, fungi, plants, arthropods, nematodes, and chordates, especially in the complete genomes of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Ciona intestinalis*, the closest non-vertebrate ancestor of vertebrates with its genome completely sequenced.

TABLE IV
*Pairwise percent identities among BGP (lower triangle) and MGP (upper triangle) protein sequences*

Diagonal bold values are sequence lengths, and shaded areas indicate identities within different groups of organisms (black, birds and amphibians; dark gray, mammals; light gray, fish). The identity values are calculated from alignments described in the text.

| | | | | | | | | | Gga | Dr | Ip | Tr | Tn | Cs | Ar | Sa | Sse | Hd | Om | Ssa | Xl | Gg | Hs | Bt | Ss | Oc | Mm | Rn | MGP* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dr | 104 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Cc | 79 | 48 | | | | | | | 102 | 30 | 32 | 26 | 30 | 29 | 33 | 32 | 35 | 33 | 30 | 31 | 37 | 32 | 28 | 26 | 27 | 27 | 27 | 26 | Gga |
| Tr | 56 | 58 | 90 | | | | | | | 105 | 48 | 39 | 40 | 38 | 47 | 44 | 48 | 44 | 49 | 50 | 40 | 35 | 28 | 29 | 29 | 27 | 28 | 27 | Dr |
| Tn | 52 | 54 | 87 | 89 | | | | | | | 105 | 37 | 41 | 38 | 47 | 44 | 49 | 44 | 46 | 45 | 39 | 37 | 32 | 33 | 35 | 33 | 34 | 33 | Ip |
| Ar | 49 | 58 | 49 | 54 | 97 | | | | | | | 110 | 60 | 50 | 55 | 54 | 53 | 54 | 49 | 48 | 36 | 31 | 32 | 32 | 36 | 35 | 32 | 32 | Tr |
| Sa | 59 | 73 | 48 | 52 | 69 | 97 | | | | | | | 117 | 56 | 56 | 55 | 54 | 51 | 50 | 53 | 38 | 36 | 33 | 32 | 36 | 35 | 31 | 32 | Tn |
| Lm | 73 | 80 | 67 | 69 | 73 | 87 | 45 | | | | | | | 90 | 62 | 71 | 68 | 70 | 50 | 50 | 31 | 29 | 28 | 27 | 29 | 27 | 27 | 28 | Cs |
| On | 69 | 73 | 49 | 51 | 63 | 86 | 84 | 65 | | | | | | | 117 | 83 | 76 | 79 | 56 | 59 | 44 | 36 | 36 | 37 | 38 | 39 | 38 | 37 | Ar |
| Xg | 70 | 60 | 74 | 72 | 66 | 64 | 64 | 62 | 47 | | | | | | | 117 | 81 | 91 | 56 | 59 | 43 | 37 | 39 | 39 | 40 | 40 | 38 | 39 | Sa |
| Sse | 60 | 71 | 48 | 47 | 63 | 73 | 80 | 78 | 60 | 97 | | | | | | | 144 | 79 | 61 | 64 | 38 | 38 | 37 | 38 | 39 | 38 | 37 | 38 | Sse |
| Hd | 54 | 60 | 44 | 47 | 63 | 65 | 71 | 77 | 62 | 71 | 97 | | | | | | | 117 | 58 | 59 | 42 | 39 | 41 | 41 | 42 | 42 | 39 | 41 | Hd |
| Om | 61 | 69 | 46 | 45 | 49 | 59 | 71 | 63 | 60 | 59 | 57 | 74 | | | | | | | 114 | 96 | 41 | 38 | 35 | 36 | 37 | 36 | 34 | 36 | Om |
| Gg | 29 | 29 | 28 | 29 | 32 | 28 | 31 | 28 | 38 | 28 | 28 | 27 | 98 | | | | | | | 112 | 42 | 38 | 35 | 36 | 37 | 36 | 34 | 36 | Ssa |
| Dn | 38 | 33 | 44 | 42 | 40 | 35 | 36 | 33 | 45 | 35 | 40 | 35 | 71 | 48 | | | | | | | 104 | 50 | 50 | 51 | 52 | 47 | 49 | 50 | Xl |
| Xl | 27 | 33 | 32 | 34 | 31 | 31 | 40 | 32 | 47 | 34 | 31 | 28 | 48 | 75 | 101 | | | | | | | 103 | 55 | 55 | 55 | 52 | 50 | 53 | Gga |
| Hs | 28 | 31 | 37 | 37 | 37 | 34 | 38 | 35 | 47 | 34 | 34 | 32 | 49 | 73 | 49 | 98 | | | | | | | 103 | 84 | 84 | 81 | 84 | 81 | Hs |
| Mf | 35 | 31 | 45 | 45 | 41 | 37 | 38 | 35 | 47 | 35 | 41 | 37 | 63 | 73 | 71 | 96 | 49 | | | | | | | 103 | 93 | 83 | 83 | 83 | Bt |
| Ec | 35 | 31 | 45 | 45 | 41 | 37 | 38 | 35 | 47 | 35 | 41 | 37 | 61 | 73 | 71 | 94 | 94 | 49 | | | | | | | 103 | 85 | 85 | 84 | Ss |
| Ss | 35 | 31 | 43 | 43 | 39 | 35 | 36 | 33 | 45 | 33 | 39 | 35 | 57 | 71 | 67 | 88 | 88 | 94 | 49 | | | | | | | 103 | 85 | 85 | Oc |
| Bt | 27 | 31 | 34 | 35 | 37 | 34 | 38 | 35 | 47 | 35 | 35 | 32 | 47 | 73 | 51 | 82 | 96 | 98 | 92 | 100 | | | | | | | 104 | 88 | Mm |
| Ch | 29 | 25 | 37 | 37 | 33 | 29 | 29 | 27 | 38 | 27 | 33 | 29 | 51 | 65 | 61 | 78 | 82 | 82 | 88 | 84 | 49 | | | | | | | 103 | Rn |
| Oa | 31 | 27 | 39 | 39 | 35 | 31 | 31 | 29 | 40 | 29 | 35 | 31 | 55 | 69 | 65 | 84 | 84 | 88 | 86 | 86 | 94 | 49 | | | | | | | |
| Fc | 33 | 29 | 41 | 41 | 37 | 33 | 33 | 31 | 43 | 31 | 37 | 35 | 57 | 71 | 65 | 84 | 84 | 86 | 86 | 88 | 82 | 82 | 49 | | | | | | |
| Cf | 37 | 33 | 45 | 45 | 43 | 37 | 38 | 35 | 47 | 35 | 41 | 37 | 57 | 71 | 65 | 88 | 88 | 88 | 86 | 84 | 80 | 82 | 84 | 49 | | | | | |
| Se | 27 | 23 | 37 | 35 | 33 | 27 | 27 | 24 | 36 | 27 | 31 | 29 | 55 | 63 | 61 | 78 | 82 | 76 | 78 | 78 | 84 | 80 | 76 | 71 | 49 | | | | |
| Oc | 37 | 33 | 45 | 45 | 43 | 37 | 38 | 35 | 47 | 35 | 41 | 37 | 55 | 65 | 63 | 76 | 80 | 80 | 76 | 80 | 69 | 71 | 80 | 80 | 63 | 49 | | | |
| Mm | 32 | 29 | 36 | 36 | 37 | 34 | 36 | 34 | 40 | 34 | 34 | 36 | 39 | 46 | 40 | 60 | 59 | 59 | 61 | 55 | 55 | 53 | 59 | 61 | 49 | 61 | 95 | | |
| Rn | 30 | 31 | 37 | 37 | 36 | 31 | 36 | 32 | 45 | 31 | 32 | 27 | 43 | 63 | 45 | 66 | 59 | 80 | 82 | 67 | 73 | 73 | 84 | 82 | 69 | 76 | 76 | 99 | |
| BGP* | Dr | Cc | Tr | Tn | Ar | Sa | Lm | On | Xg | Sse | Hd | Om | Gg | Dn | Xl | Hs | Mf | Ec | Ss | Bt | Ch | Oa | Fc | Cf | Se | Oc | Mm | Rn | |

* Ar, *Argyrosomus regius* (meagre); Bt, *Bos taurus* (domestic cattle); Cc, *Cyprinus carpio* (common carp); Cf, *Canis famillaris* (domestic dog); Ch, *Capra hircus* (goat); Cs, *Chilomycterus schoepfi* (stripped burrfish); Dn, *Dromaius novaeholiandiae* (emu); Dr, *Danio rerio* (zebrafish); Ec, *Equus caballus* (horse); Fc, *Felis catus* (domestic cat); Gg, *Gallus gallus* (chicken); Gga, *Galeorhinus galeus* (soupfin shark); Hd, *Halobatrachus didactylus* (toadfish); Hs, *Homo sapiens* (human); Ip, *Ictalurus punctatus* (channel catfish); Lm, *Lepomis macrochirus* (bluegill); Mf, *Macaca fascicularis* (crab-eating macaque); Mm, *Mus musculus* (house mouse); Oa, *Ovis aries* (sheep); Oc, *Oryctolagus cuniculus* (European rabbit); Om, *Oncorhynchus mykiss* (rainbow trout); On, *Oreochromis niloticus* (Nile tilapia); Rn, *Rattus norvegicus* (Norway rat); Sa, *Sparus aurata* (gilthead seabream); Se, *Setonix sp.* (wallaby); Ss, *Sus scrofa* (pig); Ssa, *Salmo salar* (Atlantic salmon); Sse, *Solea senegalensis* (Senegalese sole); Tn, *Tetraodon nigroviridis* (green pufferfish); Tr, *Takifugu rubripes* (torafugu); Xg, *Xiphias gladius* (swordfish); Xl, *Xenopus laevis* (African clawed frog).

*Structure of MGP and BGP Genes*—BGP and MGP gene structure organization was determined from 16 genes representing 8 different species (non-fish species: African clawed frog, human, Norway rat, and house mouse; fish species: gilthead seabream, green pufferfish, torafugu, and zebrafish). For each new gene, putative splice sites and potential coding regions were predicted using GenScan. Predictions were verified by aligning the predicted amino acid sequences with those already characterized in other species and further confirmed by comparison with known gene structures. Results presented in Fig. 2 showed that BGP and MGP genes shared the same simple gene organization with four coding exons and three introns splitting the coding region. The phase of introns positioned in the coding sequence is also rigorously conserved not only among all BGP and all MGP genes but also between BGP and MGP genes (Table II). The intron phases, as defined by Patthy (46), in both BGP and MGP genes are: phase 1 for the first intron, phase 1 for the second intron, and phase 2 for the third intron. Although the lengths of exons are well conserved within as well as between BGP and MGP genes (exon 4 > exon 3 > exon 1 > exon 2), the total size of introns varies from 491 to 2514 bp in BGP genes and from 366 to 4461 bp in MGP genes (Table II). These variations are not related to the size of the genome (species with long introns are not necessarily those with big genomes; Table III), to a subset of sequences (fish *versus* non-fish sequences; Fig. 2) or to the compactness of the chromosome region to which they belong. Altogether, these results suggest an overall conservation of gene structure between MGP and BGP.

*Conserved Features of MGP and BGP*—In pairwise comparisons of all BGP or MGP protein sequences (Table IV), overall sequence identity was 63 ± 11 (fish BGP), 65 ± 15 (amphibian and bird BGP), 79 ± 11 (mammalian BGP), 52 ± 15 (fish MGP), 50 ± 0 (amphibian and bird MGP), and 84 ± 3 (mammalian MGP) indicating more constrained sequences in mammals, especially for MGP. Protein size was found to be significantly different between BGPs and MGPs, the latter being the larger, for which it contributed the absence of a propeptide. Accordingly, BGPs were 97 ± 4 aa (unprocessed protein) and 48 ± 2 aa (mature protein), whereas MGPs were 110 ± 10 aa (unprocessed protein) and 92 ± 10 aa (mature protein). Higher variation in length within MGPs can be attributed to a larger protein in fish (116 ± 11 aa (unprocessed protein) and 97 ± 10 aa (mature protein) in fish *versus* 103 ± 0 aa (unprocessed protein) and 84 ± 0 aa (mature protein) in non-fish sequences) resulting mainly from an extended C terminus.

Analysis of MGP and BGP logos (Fig. 3) generated from T-Coffee multiple sequence alignments identified highly conserved residues (23 in MGP and 20 in BGP) that have been further analyzed using bibliographical data and tools available through the Internet. The following conserved domains were identified. 1) A transmembrane signal peptide was found in both BGP and MGP to control protein entry into the secretory pathway. Signal peptides were predicted to be 16–21 and 18–25 aa long in MGP and BGP, respectively. 2) A propeptide was found in BGP but not in MGP. The propeptide is cleaved by furin at the R$X$(K/R)R polybasic cleavage site that is present in all BGPs with one exception (meagre BGP, where the cleavage site is R$XX$R). This is a common feature for proteins known to require post-translational proteolytic processing such as clotting factors (47). 3) A phosphorylation domain was found in MGP (48) as well as 4) a γ-glutamyl carboxylation recognition
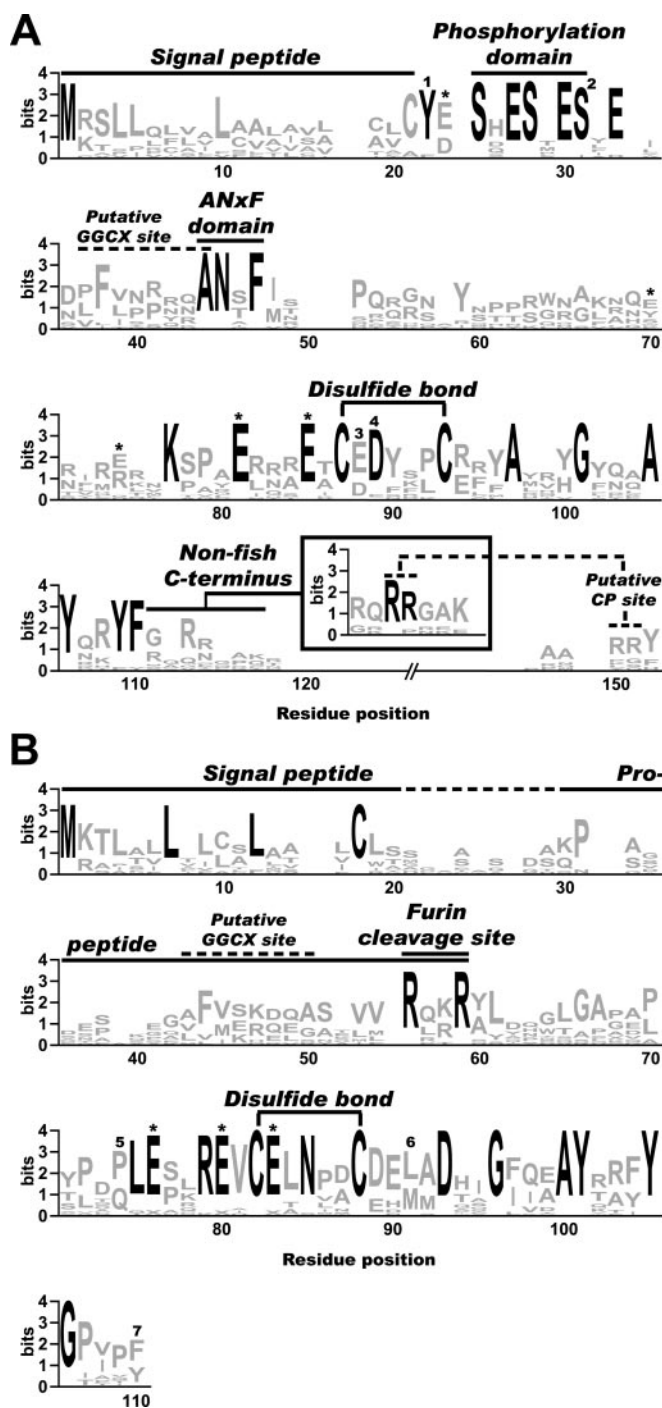
FIG. 3. **Conserved features among vertebrate MGP (*A*) and BGP (*B*) proteins.** Sequence logos reveal the conservation of residues at particular positions in the protein sequence. Highly conserved residues are in *black*. *, Gla residues; *1*, tyrosine at position 22 in MGPs, except in *Ictalurus punctatus* MGP; *2*, serine at position 32 in MGPs, except in *G. galeus* MGP; *3*, glutamic acid at position 88 in MGPs, except in mammals; *4*, aspartic acid at position 89 in MGPs, except in *G. galeus* MGP; *5*, proline at position 74 in BGPs, except in fish; *6*, leucine at position 91 in BGPs, except in fish; *7*, only in fish. *CP*, carboxypeptidase.

site, which targets the VKD proteins to the γ-glutamyl carboxylase (GGC*X*). It is most likely that this region docks with the membrane-bound GGC*X*, bringing the active site of the enzyme in close proximity to the substrate glutamate residues on the precursor form of VKD proteins. 5) An AN*X*F domain was found in MGP, which is likely to be a proteolytic cleavage site involved in post-translational processing. 6) Two invariable

cysteine residues required for intramolecular disulfide bridge were also identified. 7) A C-terminal Gla domain was found in both BGP and MGP, where the majority of the Glu residues are carboxylated. This domain is responsible for the high affinity binding of calcium ions. A conserved motif is found in the middle of the domain, which seems to be important for substrate recognition by the carboxylase. 8) A C-terminal carboxypeptidase processing site was found removing the exposed C-terminal basic residues. Interestingly, when rating the gene structure with the protein domains of MGP/BGP it is clear that each coding exon always encodes the same protein domain(s) (Fig. 4). Exon 1 encodes the complete transmembrane signal peptide and its cleavage site; exon 3 encodes the carboxylase recognition site, a proteolytic cleavage site (AN*X*F in MGP and R*XX*R in BGP) and part of the Gla domain; and exon 4 encodes the major part of the Gla domain including the canonical cysteine residues. Exon 2, the smallest exon in both genes, encodes a domain exhibiting less conservation; it is a highly conserved phosphorylation domain in MGP and a poorly conserved region of the BGP propeptide with no function assigned yet (Figs. 3 and 4). Altogether, these results suggest an overall conservation of protein structure between MGP and BGP.

Putative serine, tyrosine, and threonine phosphorylation sites were identified for each sequence using NetPhos. No conserved phosphorylation site was identified in BGP, whereas three highly conserved serine phosphorylation sites (positions 25, 28, and 31 in the consensus sequence) were found in MGP (Fig. 3*A*). Interestingly, a serine phosphorylation site (position 78) and two tyrosine phosphorylation sites (positions 90 and 96) were also highly conserved but only in fish and amphibians (Ser-78), in mammals (Tyr-90) or in mammals and amphibians (Tyr-96).

Domains identified through the T-Coffee/Logo analysis were converted to PROSITE-like domain definition (Table V) and used to scan Swiss-Prot (release 45.0) and TrEMBL (release 28.0) vertebrate entries using the ScanPROSITE facilities at us.expasy.org. MGP phosphorylation domain definition identified most MGPs (only one exception, *Galeorhinus galeus* MGP), other matrix- or bone-related proteins (dentin matrix protein 1 and osteopontin-like protein), and proteins apparently unrelated to mineralization mechanisms. The AN*X*F domain definition identified all MGPs, the VKD TMG4 protein, osteomodulin (extracellular matrix protein produced by osteoblasts), and many other proteins including protein kinases and membrane receptors. GlaMGP and GlaBGP domain definitions identified all and only MGPs and BGPs, respectively. Similarly the Gla*X*GP domain definition identified all and only *X*GPs (bone and matrix Gla proteins). Therefore the BGP/MGP Gla domain can be considered the core family domain. The efficiency of the Gla*X*GP pattern in identifying BGPs and MGPs was evaluated using Swiss-Prot release 45.0 and TrEMBL release 28.0 (containing 25 BGPs and 16 MGPs) and compared with similar published pattern definitions including (i) PROSITE PS00011, (ii) Pfam PF00594, (iii) Smart SM00069, and (iv) InterPro IPR000294 and IPR002384. Published pattern definitions never identified the whole set of BGP/MGP sequences and always picked up other sequences, mostly other members of the VKD protein family (Table VI). On the contrary, the Gla*X*GP pattern definition identified all and only reference sequences demonstrating its specificity and its suitability to identify bone and matrix Gla proteins.

*Evolutionary Aspects*—Maximum likelihood Quartet puzzling phylogenetic trees were generated from BGP, MGP, and Gla domain T-Coffee multiple sequence alignments (28, 20, and 48 sequences, respectively) excluding poorly conserved regions (Fig. 5). Phylogenetic trees were rooted with either the se-
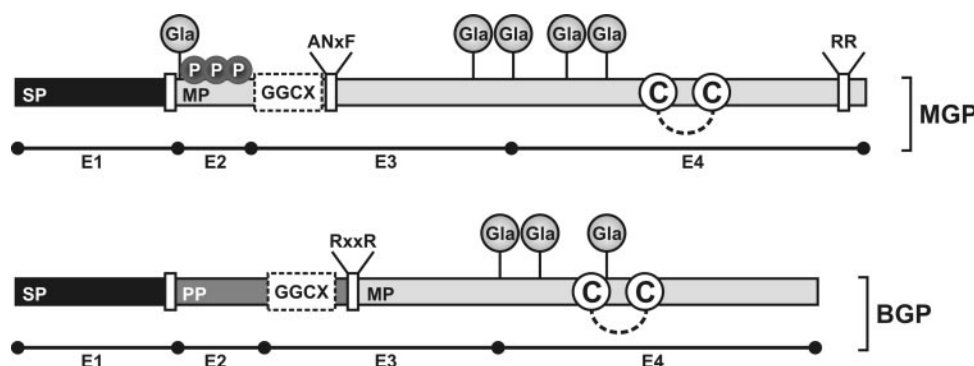
Fig. 4. **Structural organization of MGP and BGP proteins.** Exonic structure of MGP and BGP coding sequence is indicated *below* each protein structure. *Dashed line* represents intramolecular disulfide bond. *SP*, signal peptide; *PP*, propeptide; *MP*, mature peptide; *GGCX*, γ-glutamyl carboxylase recognition site; *P*, phosphorylation; *C*, conserved cysteine residues; *ANXF*, *RXXR*, and *RR* indicate proteolytic cleavage sites; *E*, exon.

TABLE V

*List of conserved motifs identified in BGP and MGP*

| Domain name | PROSITE-like definition |
|---|---|
| *Matrix Gla protein* | |
| Phosphorylation | [ED]-S-x-E-S-x-E-S-x-[ED] |
| ANxF cleavage site | A-N-[TSGAI]-F-[MI] |
| GlaMGP | E-x(3)-E-x-C-[ED]-[ED]-x(3)-C-x(3)-A-x(3)-G-x(3)-A-Y |
| *Bone Gla protein* | |
| Propeptide cleavage site | R-X-[KR]-R [a] or R-X-X-R |
| GlaBGP | E-x(3)-E-x-C-E-x(4)-C-[ED]-x(3)-D-x(2)-G-x(3)-A-Y |
| *Bone and matrix Gla proteins* | |
| GlaXGP | E-x(3)-E-x-C-[ED]-x(4)-C-x(7)-G-x(3)-A-Y |
| GGCX [b] recognition site | [APVL]-[FVL]-[MVL]-x(4)-A [c] |

[a] Preferred site.
[b] γ-Glutamyl carboxylase.
[c] Poorly conserved site.

TABLE VI

*Efficiency of selected pattern definitions to identify MGP and BGP reference sequences (containing at least the mature protein) searching Swiss-Prot (release 45.0) and TrEMBL (release 28.0) data bases*

| | Motif | BGP [a] | MGP [b] | Other hits [c] |
|---|---|---|---|---|
| PROSITE | PS00011 | 25/25 | 11/16 | 142 |
| PFAM | PF00594 | 24/25 | 15/16 | 144 |
| SMART | SM00069 | 17/25 | 10/16 | 173 |
| INTERPRO | IPR000294 | 22/25 | 11/16 | 131 |
| INTERPRO | IPR002384 | 20/25 | 9/16 | 34 |
| This study | GlaXGP | 25/25 | 16/16 | 0 |

[a] 25 BGP reference sequences.
[b] 16 MGP reference sequences.
[c] Vertebrate entries, including other members of the VKD family.

quence of the more ancient species (*G. galeus* MGP (GgaMGP)) for the MGP and Gla domain trees; Fig. 5, *A* and *C*) or the more ancient taxa (bony fish for the BGP tree; Fig. 5*B*). In agreement with the generally accepted phylogeny of vertebrates, our analysis clustered amphibian, avian, and mammalian sequences separately from that of fish. The unexpected position of the toadfish MGP (*Halobatrachus didactylus* MGP) in the MGP tree (toadfish is part of the Paracanthopterygii superorder, and *H. didactylus* MGP is therefore not expected to cluster together with proteins from Acanthopterygii) was confirmed by the cloning and the sequencing of a 194-bp MGP cDNA fragment from a closely related species, the oyster toadfish *Opsanus tau* (GenBank™ accession number AY383483).[2] The peptide encoded (including 65% of the mature protein and most of the Gla domain) was 100% identical to *H. didactylus* MGP. This result

may be due to the lack of a sufficient phylogenetic signal, caused by both short sequence length and high conservation of fish MGP sequences. Phylogenetic analysis of the Gla domain supports the notion that MGP and BGP represent distinct evolutionary groups (Fig. 5*C*), as indicated by the relatively high branching support values.

Based on the calibration points described under "Experimental Procedures," both BGP and MGP averaged evolutionary rates that were close to $1.00 \times 10^{-3}$ changes/site/Myr. These rates place BGP and MGP in the class of slowly evolving proteins (49), indicating a strong evolutionary pressure to conserve the sequence and likewise the structure.

*Time Scale for BGP and MGP Origin*—To estimate approximate times for the emergence of the MGP and BGP sequences, average intragroup evolutionary distances were estimated for the BGP and MGP sequences using the Grishin formula and the Gla region alignment of the 48 sequences. These average distances were then converted in time units using the previously estimated evolutionary rate of 0.1 PAM/Myr. Based on this estimate, the origin of MGP is placed at $480 \pm 133$ Myr ago and that of BGP at $381 \pm 102$ Myr ago. Although these values are only indicative, they agree well with our hypothesis for the origin of MGP and BGP in relation to the phylogeny of animal species (see "Discussion" and Fig. 7).

DISCUSSION

This work identified 13 new sequences (8 MGPs and 5 BGPs) through cDNA cloning or sequence reconstruction, increasing the number of available sequences to 48 (20 MGPs and 28 BGPs). Taking advantage of the resulting large amount of data, we have compared MGP and BGP genes and proteins, created multiple sequence alignments to identify conserved features likely to be important for protein structure, function, and regulation, and performed a phylogenetic analysis of both proteins.

*MGP and BGP Are Vertebrate-specific Proteins*—This work identified sequences across 32 different species representing most classes of vertebrates (including mammals, birds, amphibians, bony fish, and cartilaginous fish) but none in other organisms (including archaea, bacteria, viruses, fungi, plants, arthropods, nematodes, and chordates). An extensive *in silico* analysis of public data bases available through the Internet failed to identify a Gla domain within non-vertebrate genomes already complete. However, the fruit fly *D. melanogaster* protein Msp-300 (muscle-specific protein 300) (50) has shown some homology with the MGP Gla domain.[3] The homologous region

---

[2] C. S. B. Viegas, V. Laizé, and M. L. Cancela, unpublished results.

[3] P. A. Price and M. K. Williamson, unpublished data.
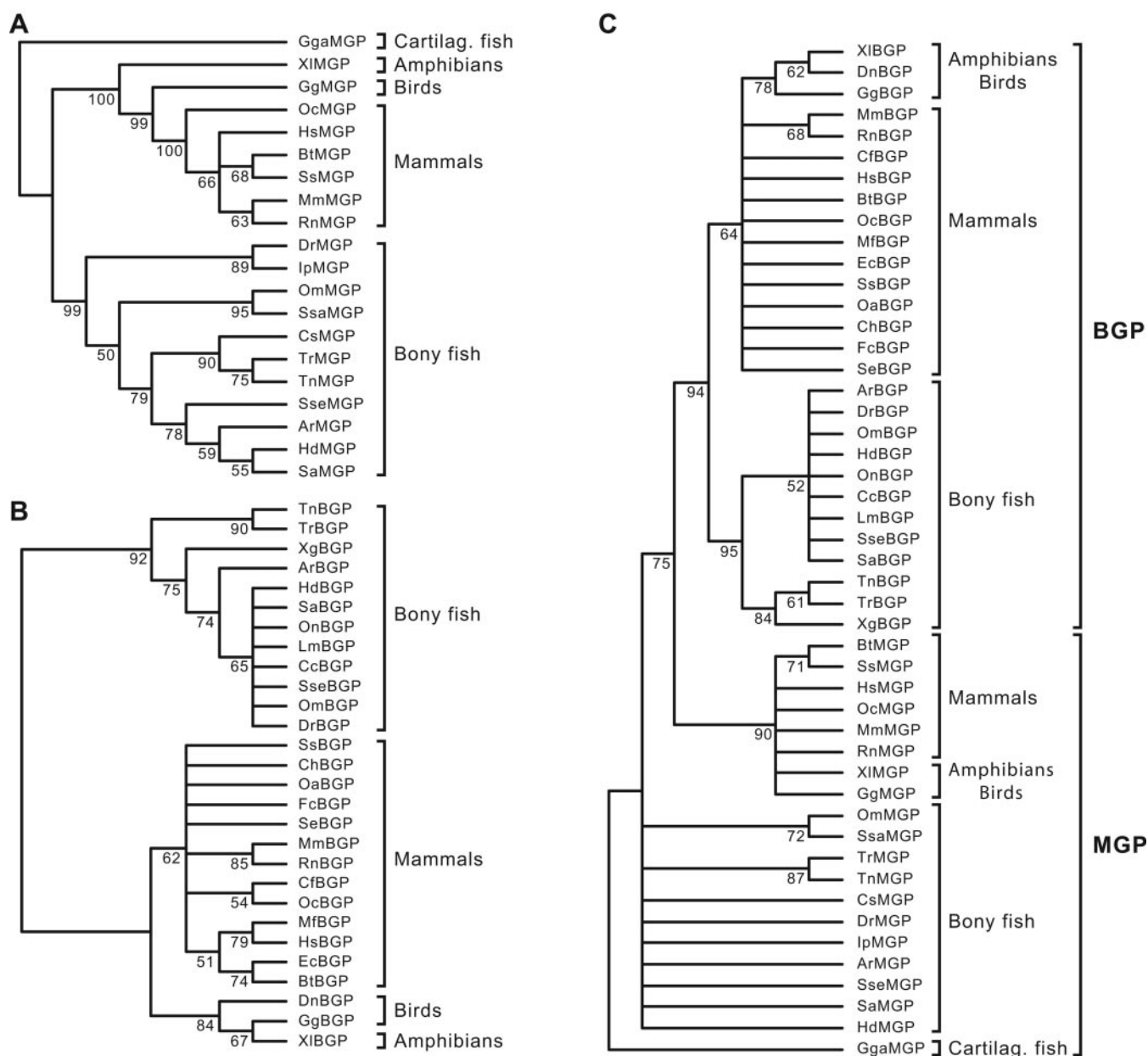
*Evolution of MGP and BGP in Vertebrates*





FIG. 5. **Quartet puzzling tree of MGPs (*A*), BGPs (*B*), and Gla domain (*C*) sequences, with maximum likelihood distances computed with the Dayhoff PAM250 matrix.** *Internal numbers* are supporting values for the corresponding branch points (support values below 50% were omitted, producing multifurcated branching). The BGP tree is rooted with the bony fish group of sequences, and both the MGP and Gla domain trees are rooted with the soupfin shark sequence. *Cartilag.*, cartilaginous.



FIG. 6. **Msp-300 pseudo-Gla domain.** MGP sequence is presented as sequence logos, where highly conserved residues are displayed as *large black characters* and an *asterisk* indicates Gla residues found in vertebrate proteins. Disulfide bond linking invariant cysteine residues is also indicated.

is located at the C terminus of Msp-300 and is only 18 aa long (Fig. 6). Despite being short, this region contains important MGP-conserved features like the invariant cysteine residues involved in intramolecular disulfide bridge and 3 Glu residues shown to be γ-carboxylated in vertebrates. A vitamin K-dependent γ-glutamyl carboxylase has also been described

in *D. melanogaster* (51) suggesting that Glu residues present within the pseudo-Gla domain of Msp-300 could be γ-carboxylated and play a role in calcium binding. Despite similarities in both sequence and tissue distribution (*i.e.* both proteins are expressed in muscle cells), Msp-300 is clearly not a matrix Gla protein, and similarities observed in protein sequences may only indicate that MGP and Msp-300 ancestors have shared at some point the same Gla-like domain that evolved differently, but have retained some characteristic features.

*MGP- and BGP-conserved Features*—Many invariant residues (shared or not between MGP and BGP) have been identified in consensus sequences obtained from multiple alignments. Their conservation over more than 450 million years of evolution strongly suggests that these residues are required for correct protein structure or to preserve a critical function or interaction (*e.g.* MGP with BMP2). It has been possible to

FIG. 7. **MGP and BGP emergence during vertebrate evolution as a result of genome duplication events.** Evolution time is in millions of years (*Myr*). *Breaks* in *arrows* indicate that lines of evidence of the existence of MGP and BGP are from preliminary unpublished results, *i.e.* MGP- and BGP-like immunoreactive proteins have been observed in lamprey and in shark, respectively (Footnote 7). *Cartilag.*, cartilaginous; *N.D.*, not detected.
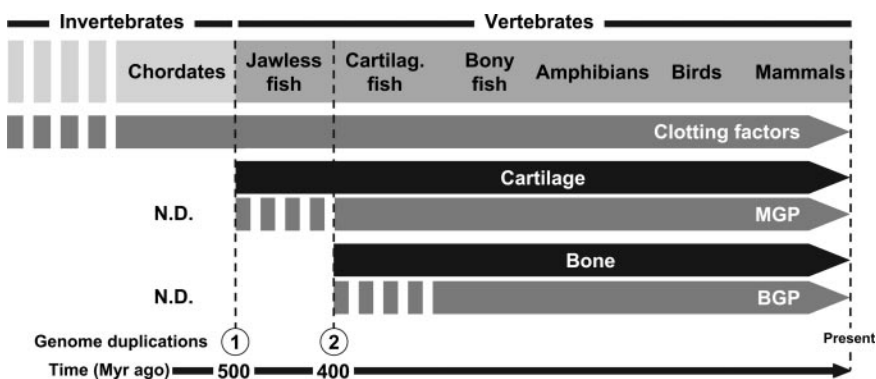


cluster these conserved residues into various functional domains and post-translational processing sites. Some of them have been identified in previous studies (48, 52, 53), but one is described for the first time in this work (the AN*X*F putative peptide cleavage site) although it has already been observed in the past.[4] This site is present in MGP and would cleave what is considered the core of the protein (the C terminus containing the Gla domain) from the N terminus containing the phosphorylation and the GGC*X* domains. The purification of low molecular weight MGP fragments identifying a cleavage site after the asparagine (N) residue of the AN*X*F domain further confirmed the post-translational proteolytic processing of MGP at this site (12).[5]

*In silico* analyses of available sequences have predicted three highly conserved serine phosphorylation sites within the N-terminal domain of MGP and none in BGP in agreement with previous *in vitro* studies (48).[6] It has been proposed recently that the Golgi casein kinase could be responsible for MGP phosphorylation at these sites (54), a post-translational modification that would serve as a sorting signal for MGP.

*MGP and BGP Phylogeny*—Vertebrate phylogeny estimated from the BGP and MGP set of sequences is consistent with the generally accepted phylogeny of vertebrates estimated from morphological and palaeontological data and is also in agreement with previous studies (7, 13, 55) even though a rather more extensive set of sequences is used. A number of branches display low support values, indicating poor agreement between the multiple solutions produced by the Quartet puzzling algorithm (36). This is more clearly seen in the case of BGP, where most branches have values below 90% (support values below 50% were omitted, producing multifurcated branching). The shorter length of the BGP sequences and the greater degree of conservation may contribute to the low support values because they result in a weak phylogenetic signal. Despite these results, separation between BGP sequences of fish and land animals is strongly supported. The Quartet puzzling tree for MGP is much better defined, with only one support value lower than 50% and many greater than 90%. Clustering of the various species agrees exactly with the accepted phylogeny of vertebrates except for the clustering of *H. didactylus* MGP and *S. aurata* MGP. As described previously, the *H. didactylus* MGP sequence is known with a high degree of certitude, so the unexpected high similarity between the two sequences must come from random mutation noise obliterating the phylogenetic signal. Molecular trees obtained from the analysis of the BGP and MGP set of sequences are not species trees, and a number of evolutionary processes can introduce differences between a correctly estimated gene phylogeny and the correct species phylogeny. These processes are horizontal transfer, duplication and loss, and deep coalescence (56).

Overall phylogenetic analysis based exclusively on the Gla motif region of both MGP and BGP is able to differentiate the two proteins and the land/sea taxon but beyond that, all branching is very poorly defined. This is obviously because of the small number of aligned residues and high amount of sequence conservation. However, it must be stressed that the Gla motif region contains enough information to tell BGP and MGP apart and their sea/land taxa. When more structural information is available, it may be possible to correlate these differences with adaptation to the different environments.

*Origin and Evolution of BGP and MGP*—It is quite clear that BGP and MGP have a common origin and that they are more closely related to each other than to any other VKD protein. This is well supported by the data base searching, the exon patterns, and protein sequence alignments showing that 1) BGP and MGP genes share the same simple gene organization and protein structure and 2) the C-terminal domain signature is similar in MGP and BGP. It is also clear that BGP and MGP are nearly if not completely absent in most non-vertebrate taxa. Levels of sequence similarity between BGP and MGP are much higher than one would expect to occur by convergence, suggesting that this gene group originated through gene duplication followed by subsequent sequence divergence. This duplication (gene, chromosome, or genome duplication) probably occurred very early in vertebrate evolution, being almost certainly an ancient event. Strong evidence for whole genome duplication has been shown for vertebrates (57). After the divergence of the cephalochordates from the chordate line, a genomic duplication occurred before jawless fish evolved around 500 Myr ago (Fig. 7). Another genomic duplication event probably led to the evolution of jawed fish around 400 Myr ago (Fig. 7). These two duplications led to the development of many distinctive vertebrate features, such as cartilage and bone. We hypothesize that the first genome duplication (before the branching of jawless fish) originated the ancestor gene of MGP, and the second genome duplication (before the branching of cartilaginous fish) would have produced the BGP ancestor gene. The BGP (380 Myr ago) and MGP (480 Myr ago) emergence times estimated from the evolutionary rate agree well with this conjecture. The appearance of MGP would be followed by cartilage formation and that of BGP by bone formation. After duplication, gene duplicates (BGP in this case) often experience relaxed evolutionary constraints. This promotes functional diversification of duplicates and biochemical innovation through mutations and recombination. In other words, duplicates evolve to acquire new functions. Several more speculative lines of evidence for BGP being a duplicate of MGP have been collected: 1) the presence of a MGP-like immunoreactive protein has been observed in lamprey (the more ancient species

---

[4] P. A. Price and M. K. Williamson, unpublished data.
[5] P. A. Price personal communication.
[6] P. A. Price, unpublished laboratory data.

tested), whereas BGP was not detected[7]; 2) MGP is associated with cartilage, which appeared with the first vertebrates, whereas BGP is only associated with bone, a structure that appeared later in evolution; and 3) BGP seems to be better conserved than MGP.

Although the existence of BGP and MGP paralogues and/or pseudogenes cannot be ruled out, the absence of genes identified during both our survey and our various cloning events in different species that could have resulted from vertebrate genome duplication tends to confirm this hypothesis. The only exception identified so far, the mouse BGP gene cluster (45), is likely to be a recent duplication event after the rodent branching.

*Conclusions*—Our results support the hypothesis that: 1) bone and matrix Gla protein genes originated from two genome duplications that occurred around 500 and 400 Myr ago before jawless and jawed fish evolved, respectively; 2) cartilage and bone tissues arose concomitantly with the appearance of MGP and BGP, respectively; and 3) BGP is a duplicate of MGP. Finally we propose highly specific pattern definitions for the BGP and MGP Gla domain that should help to reliably identify bone and matrix Gla proteins in the growing family of non-annotated sequences originating from the numerous sequencing projects.

## REFERENCES

1. King, K. (1978) *Biochim. Biophys. Acta* **542,** 542–546
2. Walker, C. S., Shetty, R. P., Clark, K., Kazuko, S. G., Letsou, A., Olivera, B. M., and Bandyopadhyay, P. K. (2001) *J. Biol. Chem.* **276,** 7769–7774
3. Stanley, T., Stafford, D., Olivera, B., and Bandyopadhyay, P. (1997) *FEBS Lett.* **407,** 85–88
4. Craig, A. G., Bandyopadhyay, P., and Olivera, B. M. (1999) *Eur. J. Biochem.* **264,** 271–275
5. Cancela, M. L., Williamson, M. K., and Price, P. A. (1995) *Int. J. Pept. Protein Res.* **46,** 419–423
6. Rice, J. S., Williamson, M. K., and Price, P. A. (1994) *J. Bone Miner. Res.* **9,** 567–576
7. Pinto, J. P., Ohresser, M. C. P., and Cancela, M. L. (2001) *Gene* **270,** 77–91
8. Proudfoot, D., Skepper, J. N., Shanahan, C. M., and Weissberg, P. L. (1998) *Arterioscler. Thromb. Vasc. Biol.* **18,** 379–388
9. Yagami, K., Suh, J. Y., Enomoto-Iwamoto, M., Koyama, E., Abrams, W. R., Shapiro, I. M., Pacifici, M., and Iwamoto, M. (1999) *J. Cell Biol.* **147,** 1097–1108
10. Fraser, J. D., and Price, P. A. (1988) *J. Biol. Chem.* **263,** 11033–11036
11. Pinto, J. P., Conceicao, N., Gavaia, P. J., and Cancela, M. L. (2003) *Bone* **32,** 201–210
12. Simes, D. C., Williamson, M. K., Ortiz-Delgado, J. B., Viegas, C. S. B., Price, P. A., and Cancela, M. L. (2003) *J. Bone Miner. Res.* **18,** 244–259
13. Cancela, M. L., Ohresser, M. C. P., Reia, J. P., Viegas, C. S. B., Williamson, M. K., and Price, P. A. (2001) *J. Bone Miner. Res.* **16,** 1611–1621
14. Ortiz-Delgado, J. B., Simes, D. C., Gavaia, P. J., Sarasquete, C., and Cancela, M. L. (2005) *Histochem. Cell Biol.*, in press
15. Luo, G., D'Souza, R., Hogue, D., and Karsenty, G. (1995) *J. Bone Miner. Res.* **10,** 325–334
16. Luo, G., Ducy, P., McKee, M. D., Pinero, G. J., Loyer, E., Behringer, R. R., and Karsenty, G. (1997) *Nature* **386,** 78–81
17. Munroe, P. B., Olgunturk, R. O., Fryns, J. P., Van Maldergem, L., Ziereisen, F., Yuksel, B., Gardiner, R. M., and Chung, E. (1999) *Nat. Genet.* **21,** 142–144
18. Meier, M., Weng, L. P., Alexandrakis, E., Ruschoff, J., and Goeckenjan, G. (2001) *Eur. Respir. J.* **17,** 566–569
19. Schinke, T., and Karsenty, G. (2000) *Nephrol. Dial. Transplant.* **15,** 1272–1274
20. Fraser, J. D., and Price, P. A. (1990) *Calcif. Tissue Int.* **46,** 270–279
21. Kirfel, J., Kelter, M., Cancela, L. M., Price, P. A., and Schüle, R. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94,** 2227–2232
22. Cancela, M. L., and Price, P. A. (1992) *Endocrinology* **130,** 102–108
23. Farzaneh-Far, A., Proudfoot, D., Weissberg, P. L., and Shanahan, C. M. (2000) *Biochem. Biophys. Res. Commun.* **277,** 736–740
24. Conceição, N., Henriques, N. M., Ohresser, M. C. P., Hublitz, P., Schüle, R., and Cancela, M. L. (2002) *Eur. J. Biochem.* **269,** 1947–1956
25. Cancela, M. L., Hu, B., and Price, P. A. (1997) *J. Cell. Physiol.* **171,** 125–134
26. Ducy, P., Desbois, C., Boyce, B., Pinero, G., Story, B., Dunstan, C., Smith, E., Bonadio, J., Goldstein, S., Gundberg, C., Bradley, A., and Karsenty, G. (1996) *Nature* **382,** 448–452
27. Boskey, A. L., Gadaleta, S., Gundberg, C., Doty, S. B., Ducy, P., and Karsenty, G. (1998) *Bone* **23,** 187–196
28. Dowd, T., Rosen, J., Li, L., and Gundberg, C. (2003) *Biochemistry* **42,** 7769–7779
29. Hoang, Q., Sicheri, F., Howard, A., and Yang, D. (2003) *Nature* **425,** 977–980
30. Frazão, C., Simes, D. C., Coelho, R., Alves, D., Williamson, M. K., Price, P. A., Cancela, M. L., and Carrondo, M. A. (2005) *Biochemistry* **44,** 1234–1242
31. Murshed, M., Schinke, T., McKee, M. D., and Karsenty, G. (2004) *J. Cell Biol.* **165,** 625–630
32. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) *Nucleic Acids Res.* **25,** 4876–4882
33. Hardie, D. C., Gregory, T. R., and Hebert, P. D. N. (2002) *J. Histochem. Cytochem.* **50,** 735–749
34. Notredame, C., Higgins, D. G., and Heringa, J. (2000) *J. Mol. Biol.* **302,** 205–217
35. Schneider, T. D., and Stephens, R. M. (1990) *Nucleic Acids Res.* **18,** 6097–6100
36. Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002) *Bioinformatics* **18,** 502–504
37. Felsenstein, J. (1989) *Cladistics* **5,** 164–166
38. Page, R. D. (1996) *Comput. Appl. Biosci.* **12,** 357–358
39. Kumar, S., and Nei, M. (2000) *Molecular Evolution and Phylogenetics*, pp. 24–25, Oxford University Press, New York
40. Grishin, N. V. (1995) *J. Mol. Evol.* **41,** 675–679
41. Feng, D. F., Cho, G., and Doolittle, R. F. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94,** 13028–13033
42. Gu, X., Wang, Y. F., and Gu, J. Y. (2002) *Nat. Genet.* **31,** 205–209
43. Kumar, S., and Hedges, S. B. (1998) *Nature* **392,** 917–920
44. Tajima, F. (1993) *Genetics* **135,** 599–607
45. Desbois, C., Hogue, D. A., and Karsenty, G. (1994) *J. Biol. Chem.* **269,** 1183–1190
46. Patthy, L. (1987) *FEBS Lett.* **214,** 1–7
47. Zhou, A., Webb, G., Zhu, X., and Steiner, D. F. (1999) *J. Biol. Chem.* **274,** 20745–20748
48. Price, P. A., Rice, J. S., and Williamson, M. K. (1994) *Protein Sci.* **3,** 822–830
49. Creighton, T. E. (1993) *Proteins: Structures and Molecular Properties*, pp. 121–123, W. H. Freeman, New York
50. Volk, T. (1992) *Development* **116,** 721–730
51. Li, T., Yang, C. T., Jin, D. Y., and Stafford, D. W. (2000) *J. Biol. Chem.* **275,** 18291–18296
52. Hale, J., Williamson, M., and Price, P. (1991) *J. Biol. Chem.* **266,** 21145–21149
53. Price, P. A., Fraser, J. D., and Metz-Virca, G. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84,** 8335–8339
54. Wajih, N., Borras, T., Xue, W., Hutson, S. M., and Wallin, R. (2004) *J. Biol. Chem.* **279,** 43052–43060
55. Viegas, C. S. B., Pinto, J. P., Conceição, N., Simes, D. C., and Cancela, M. L. (2002) *Gene* **289,** 97–107
56. Cotton, J., and Page, R. (2002) *Proc. R. Soc. Lond. B Biol. Sci.* **269,** 1555–1561
57. Postlethwait, J., Yan, Y., Gates, M., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson, M., Abduljabbar, T., Yelick, P., Beier, D., Joly, J., Larhammar, D., and Rosa, F. (1998) *Nat. Genet.* **18,** 345–349
58. Lamatsch, D. K., Steinlein, C., Schmid, M., and Schartl, M. (2000) *Cytometry* **39,** 91–95
59. Nishimoto, S. K., Waite, J. H., Nishimoto, M., and Kriwacki, R. W. (2003) *J. Biol. Chem.* **278,** 11843–11848

[7] D. Simes, personal communication.