RESEARCH PAPER

Intrinsically Disordered Proteins 1, e27450; 2013; © 2013 Landes Bioscience

Evolutionary journey of the Gc protein (vitamin D-binding protein) across vertebrates

Shaheena Anwar*[1], Mohammed Perwaiz Iqbal[2], Shamshad Zarina[3], and Zulfiqar A Bhutta[1]

[1]Division of Women and Child Health; Aga Khan University; Karachi, Pakistan; [2]Department of Biological and Biomedical Sciences; Aga Khan University; Karachi, Pakistan; [3] National Center for Proteomics; University of Karachi; Karachi, Pakistan

Keywords: Gc, vitamin D-binding protein, albumin superfamily, Gc evolution, phylogenetic tree, conserved regions, animal kingdom

Abbreviations: Gc, group specific component of serum; kDa, kilo Dalton; kb, kilo bases; Myrs, million years; ALB, albumin; AFP/ α-fetoprotein, alpha feto protein; AFM, afamin; mRNA, messenger ribonucleic acid

With so many diverse functions such as transporter of vitamin D metabolites and fatty acids, actin scavenger and macrophage activating factor, Gc must have been one of the most conserved proteins in animal kingdom. Our objective was to investigate the evolution of Gc (vitamin D binding protein) by analyzing its differences at protein level across vertebrates. Using BLAST (Basic Local Alignment Search Tool) searches, Gc amino acid sequences were analyzed for homology. Clustal W2 and Jalview were used for multiple sequence alignment analysis, PhyML 3.0 for phylogenetic tree, while Batch Web CD-Search Tool was used for identification of conserved domains within protein sequences. Gc protein percent identity between human and rabbit was 83%, which decreased to 81% with cow, 78% with mouse, 76% with rat, 51% with chicken, 41% with frog, and 28% with zebrafish. Phylogram showed that rat Gc was the most diverged, while chicken Gc was the most conserved protein. Analysis also indicated high homology among mammals (human, rabbit, cow, rat, and mouse). Gc appeared highly conserved in chicken and zebrafish. However, the distance from ancestral protein gradually increased in amphibian (frog) and mammals (human, rabbit, cow, rat, and mouse). Human Gc and rabbit Gc appear to be recently evolved proteins. There appears to be an interesting evolutionary pattern- chicken Gc has the least distance from the ancestral protein, while rat Gc is the most diverged. There is no vertebrate devoid of Gc which is suggestive of its important role in vitamin D metabolism in vertebrates.

Introduction

Gc protein was discovered in human as Gc-globulin (Group-specific component of serum) in 1959 by Hirschfeld. This protein was used in studies pertaining to population genetics and forensic medicine before its function was eventually discovered in 1975.[1,2] Later, it was categorized as vitamin D-binding protein (DBP). It is a polymorphic plasma protein of molecular mass 52–59 kDa.[3,4] It is expressed in liver, kidney, gonads, fat, and neutrophils in humans.[2,5] Its most important function is transportation of vitamin D metabolites. However, it also functions as an actin scavenger, macrophage activating factor and fatty acid transporter.[2,3]

The human Gc gene is present on chromosome 4 (4q11–13).[6] The gene is 42.5 kb long and belongs to albumin super family which also includes albumin (ALB), α-fetoprotein (AFP) and afamin (AFM) genes.[3,6] The family shares a 3-domain structure. Gc, ALB, AFP, and AFM are homologous in structure but vary considerably in functions.[4]

Gc is separated by a distance of 1500 kb from other members of the family.[7] Studies have shown that Gc is present in rat, mouse, rabbit, turtle, chicken, guinea pig, horse, cow, dog, rhesus monkey, and chimpanzee.[1,8] Gc in turtle (*Trachemysscripta)* has a unique ability to bind with thyroxine as well. There is no Gc homolog in *Drosophila melanogaster*.[1]

While Gc gene is located on chromosome 4 in humans, its location in cow, mouse, chicken, zebrafish, rat, and rabbit is on chromosome 6, 5, 4, 5, 14, and 15, respectively (http://www.ncbi.nlm.nih.gov/gene/).The chromosomal location of Gc gene in Western clawed frog has not been established although the whole genome of this amphibian has been sequenced.[9]

The objective of this study was to examine the sequence similarities and degree of conservation at amino acid level in order to comprehend the evolution of Gc protein from aquatic fish to mammals.

Results

The Gc protein precursor sequence lengths and their accession numbers are summarized in **Table 1**. Percent identities were calculated using pairwise approach by Clustal W2 (**Table 2**).

*Correspondence to: Shaheena Anwar; Email: shaheena.anwar@aku.edu
Submitted: 12/06/2013; Accepted: 12/06/2013; Published Online: 01/01/2014
http://dx.doi.org/10.4161/idp.27450

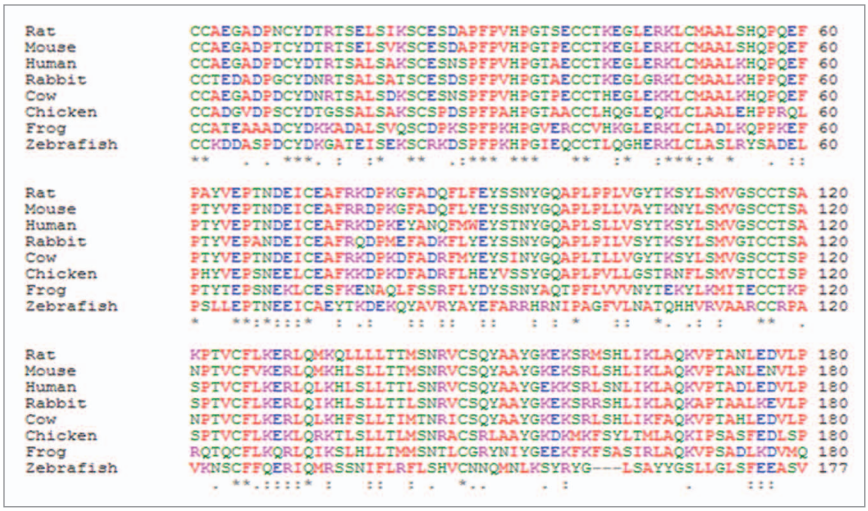

Figure 1. Clustal W2 sequence alignment of vertebrate Gc amino acid. Single fully conserved residues are denoted by asterisk (*). Conservative change in amino acid is denoted by semi colon (:). A neutral change is denoted by dot (.). Dash (-) represent non homologous segments. The numbers on the right represent the amino acid positions.

Table 1. Gc protein precursor accession numbers and lengths of the eukaryotic organisms

Organisms	Accession number	Protein precursor (aa)*
Human	NP_000574.2	474
Cow	NP_001030457.1	474
Mouse	NP_032122.1	476
Chicken	NP_990213.1	484
Frog	NP_001015745.1	482
Zebrafish	NP_001002568.1	464
Rat	AAA41080.1	476
Rabbit	BAA06137.1	476

*amino acids

Discussion

A comprehensive view of the alignment results in Clustal W2 and Jalview signifies Gc as highly conserved protein across the studied eukaryotic animals. The sequence alignment results from Clustal W2 (**Fig. 1**) and the consensus sequences in Jalview are shown in **Figure 2A and B** (complete results are available in the **Supplemental Material**).

Results of phylogenetic analysis revealed Mammalian Gc proteins (human, rabbit, mouse, rat) were distantly related to ancestral Gc, while chicken, zebrafish, and frog Gc proteins appeared to be closely related to it (**Fig. 3**). Rabbit and human Gc proteins appear to have recently evolved. The position of zebrafish Gc in the phylogenetic tree reinforced the results of multiple sequence alignment, which showed the lowest percent identity with all other animals (**Table 2**). In our analysis, human and mouse Gc proteins are clustered with rabbit and rat Gc proteins, respectively reflecting their greater degree of sequence similarity. Interestingly, chicken Gc has the smallest distance from ancestral protein although it holds a higher position in the evolution hierarchy when compared with zebrafish and frog.

Our results reveal that a high degree of Gc homology exists among mammals; however, the protein varies considerably from Pisces to Mammals. All eight Gc protein sequences have conserved domains for albumin superfamily and vitamin D binding site III (**Fig. 4A**) except for zebrafish Gc (**Fig. 4B**). It has been reported that zebrafish Gc does not have a vitamin D binding motif but has a number of albumin binding motifs. This observation merits some discussion because zebrafish genome doesn't have ALB, AFP and AFM genes even though ALB is present in salmon, brown trout and rainbow fish. There is a possibility that other plasma proteins in zebrafish might be performing the same function as albumin in other species. It has also been reported that zebrafish diverged from salmon around 170–310 Myrs (million years) ago. Most members of the albumin superfamily appeared at a later stage of vertebrate evolution during the appearance of amphibians and reptiles.[13]

Previous studies indicated that the ancestral gene for ALB and AFP was 300–500 Myrs old.[14,15] It was also suggested that Gc is older than ALB and AFP.[7] However, these estimates were questioned by other investigators. Rooted and time calibrated phylogenetic analysis by Noel et al.[13] indicated that Gc and a precursor for ALB, AFP, AFM appeared for the very first time around 570–880 Myrs ago. ALB and a precursor for AFP, AFM appeared 360–410 Myrs ago. This process probably occurred

Pairwise percent identity was maximum between mouse–rat pair (91%), followed by human–rabbit pair (83%), and minimum for mouse–zebrafish pair (26%), and intermediate (49–51%) between chicken and mammals (human, rabbit, cow, rat, and mouse).

Multiple sequence alignment analysis indicated numerous fully conserved residues in Gc protein among the 8 eukaryotes as shown in **Figure 1** (Clutal W2 output) and in **Figure 2** (Jalview output). (Complete results are available in the **Supplemental Material**).

Figure 3 shows phylogenetic tree constructed using downloaded sequences. The tree indicates that rabbit and human share a more recent common ancestor, while rat Gc is evolutionarily most distant than other animals. Chicken and zebrafish Gc are least distant from the ancestor protein.

Figure 4 shows conserved domains within Gc proteins. **Figure 4A** represents all the studied animals except zebrafish Gc, which lacks vitamin D binding site (**Fig. 4B**).

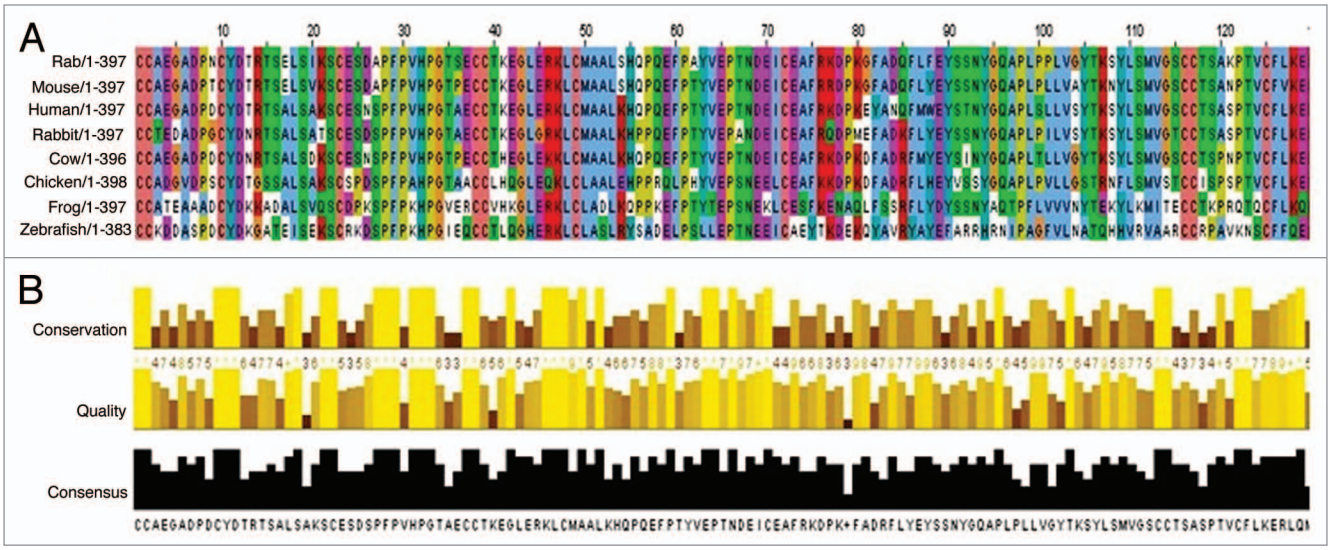

Figure 2. (**A**) Jalview sequence alignment of Gc amino acids.The residues are colored by default settings of Clustal X where a minimum percentage of single residues or a combination of residues must be achieved for a color to be applied. The position of amino acids is specified in the top. The vertebrates are indicated on the left. (**B**) Protein consensus and conserved sequences in Jalview. The amino acid property conservation defines the measurement of conservation of physiochemical properties in a column. Any change in amino acid homology is calculated as observed substitution by BLOSUM62 program in Jalview and denoted by golden bars. The consensus column defines the most common residues for each column of the alignment. The conserved amino acids are indicated in the bottom line. The black boxes are a visual summary of the degree of conservation in Gc protein which appears to be quite high for some regions.

after the amphibians and reptiles separated because AFP and AFM are absent in amphibians and fish but present in reptiles and other higher vertebrates.[13] AFP and AFM appeared around 250–330 Myrs ago after Mammalia emerged. This is supported by the fact that AFM is present only in mammals but absent in amphibians and fish.[13] Their analysis also revealed that ALB and Gc probably evolved at the same rate. Hence, Gc may not be considered older than ALB.[13]

The findings of this study must be viewed in the light of certain limitations. Only 8 vertebrates were selected for this study. Moreover, we couldn't study Gc in Reptilia because only predicted sequences were available, though Gc has been reported to be present in turtle.[1] Other members of albumin superfamily were not compared with Gc and with each other. This would have been helpful in verifying the extent of evolution regarding this family of proteins. Our results conform well to those reported by Noel et al. who have found nearly the same protein identity (29.5%) between human and zebrafish Gc.[13]

We selected reference sequences of Gc protein from each class of phylum Chordata except for Reptilia (since only predicted sequences were available) for constructing phylogenetic tree in order to evaluate the pattern of Gc protein evolution. Noel et al. used reference and predicted mRNA sequences, not protein sequences, for phylogenetic analysis.[13] They used fewer protein sequences (fish, amphibian, and 2 mammals) for multiple alignment analysis only. Predicted mRNA sequences are weak evidences for studying protein evolution as they are not subjected to wet lab experiments to verify whether these predicted sequences translate into functional proteins. Hence, they are not strong candidates for studying protein evolution. Our phylogenetic tree is based on protein sequences which provide a better model to study evolution as sequences are more variable at mRNA level as compared with protein due to degeneracy of genetic code—which ultimately code for same amino acid. Thus protein sequences provide a true picture of actual evolution taking place within any protein.

Our primary objective was to gain a better understanding of the evolution of Gc protein precursors in various classes of phylum Chordata. Further studies would be needed to compare all members of albumin superfamily in various classes of phylum Chordata. Moreover, in-depth analysis of domain and helices homology in primates would be helpful in enhancing our knowledge of Gc evolution in terms of structure-function relationship.

Conclusion

Chicken Gc was found to be closely related to ancestral Gc, while rat Gc was the most distant. Human Gc appeared to have recently evolved. Structurally, Gc was closely related among mammals (human, rabbit, cow, mouse, and rat) but was evolutionally distinct from Gc in chicken, zebrafish, and frog indicating possibility of functional versatility of this protein. Protein analysis provided a clear picture of evolutionary trend in vertebrates.

Methods

Selection of sequences

Human Gc protein sequence from NCBI database (http://www.ncbi.nlm.nih.gov/protein/) was used to search for other protein sequences using Protein BLAST

Table 2. Pairwise comparison of protein precursors in the selected vertebrate species

Vertebrate 1	Vertebrate2	Protein identity (%)
Human	Rabbit	83
Human	Cow	81
Human	Mouse	78
Human	Rat	76
Human	Chicken	51
Human	Frog	41
Human	Zebrafish	28
Cow	Rabbit	77
Cow	Mouse	74
Cow	Rat	72
Cow	Chicken	51
Cow	Frog	41
Cow	Zebrafish	28
Mouse	Rat	91
Mouse	Rabbit	76
Mouse	Chicken	50
Mouse	Frog	42
Mouse	Zebrafish	26
Chicken	Rat	51
Chicken	Rabbit	49
Chicken	Frog	41
Chicken	Zebrafish	30
Frog	Rabbit	42
Frog	Rat	41
Frog	Zebrafish	27
Rat	Rabbit	73
Rat	Zebrafish	26
Rabbit	Zebrafish	28

(http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=). We selected sequences of human (*Homo sapiens*), cow (*Bos taurus*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), rabbit (*Oryctolagus cuniculus*), chicken (*Gallus gallus*), Western clawed frog (*Xenopus tropicalis*) and zebrafish (*Danio rerio*). These species are representative of various classes of Phylum Chordata: Mammalia (human, cow, rabbit, mouse, and rat), Aves (chicken), Amphibian (frog) and Pisces (zebrafish).

Multiple sequence alignment

Multiple sequence alignment of the downloaded sequences was conducted using Clustal W2 online tool at European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Default parameters were used for the analysis.

Phylogenetic analysis

A Maximum-Likelihood tree was constructed using PhyML 3.0, with following parameters: Tree topology: NNIs; Initial tree: BioNJ; Model of amino acids substitution: LG; Log-likelihood: -4713.91023; Unconstrained Likelihood: -2261.14381; Discrete gamma model: Yes, and Gamma shape parameter: 1.720.[10,11]

Domain analysis

Conserved domains within the protein sequences were analyzed by using Batch Web CD-Search tool[12] (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi)

Disclosure of Potential Conflicts of Interest

The authors declare that they do not have any conflicting interests.

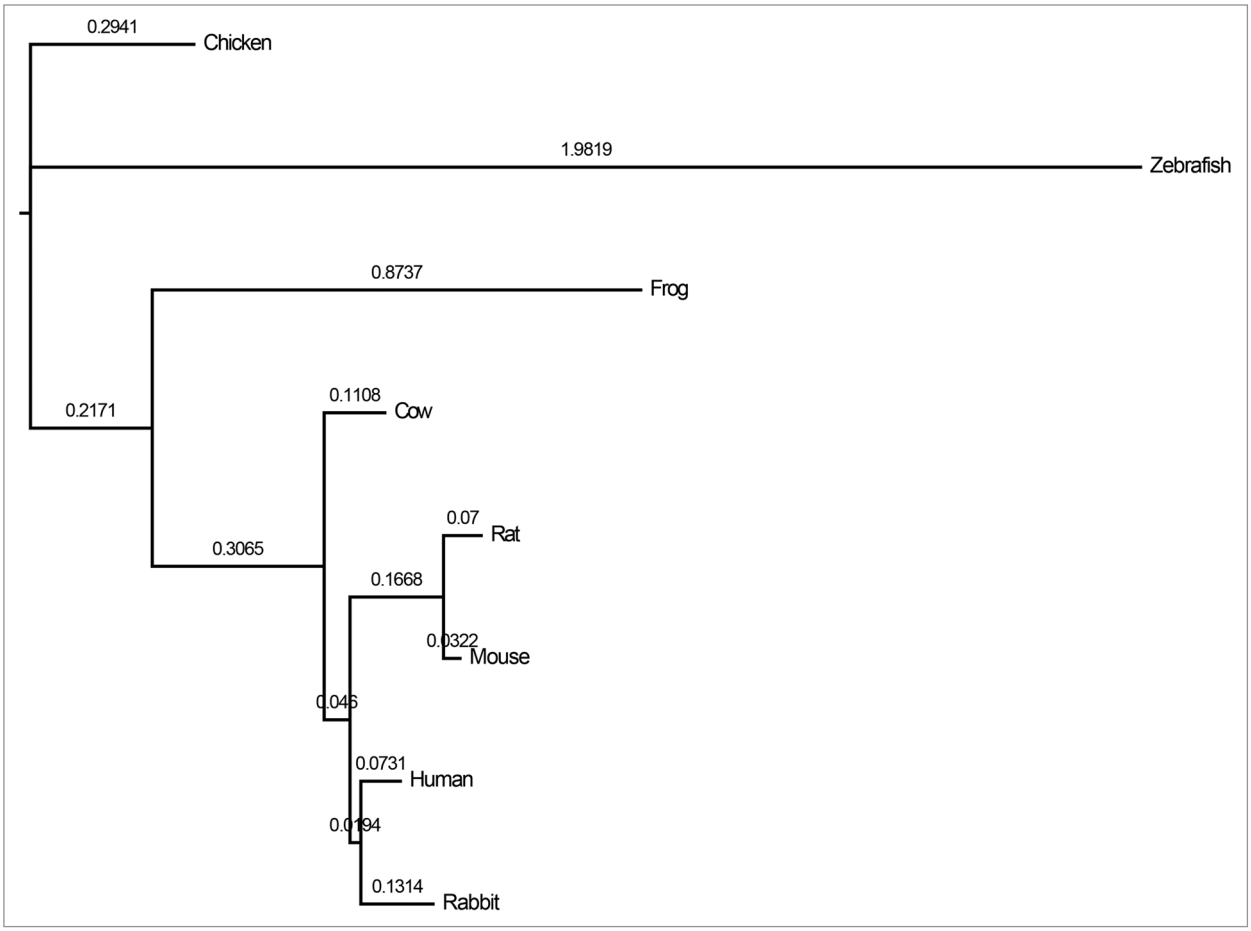

Figure 3. Phylogenetic tree of Gc amino acids by PhyML 3.0

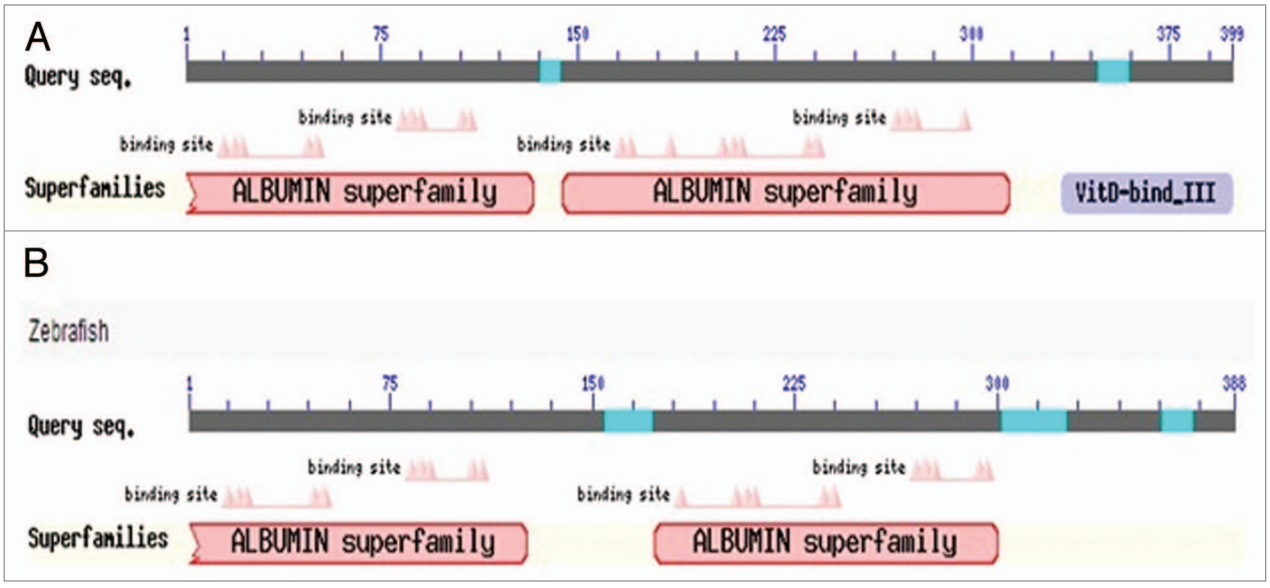

Figure 4. (**A**) Conserved domains of Gc protein precursor of human, rabbit, cow, rat, mouse, chicken, and frog. The triangles represent the amino acids of conserved domains in albumin superfamily. The vitamin D binding motif is denoted by a box at the end of the sequence. The positions of the binding sites are specified numerically at the top. (**B**) Conserved domains of Gc protein of zebrafish. The vitamin D binding site is absent in zebrafish. The albumin superfamily binding sites are indicated by triangles. The blue colored region was not included in database search because it was recognized as biased region by the software.

Funding sources

We were provided with a separate space along with computational support by Department of Biological and Biomedical Sciences, Aga Khan University for conducting this study.

Acknowledgments

We are very thankful to Dr Syed Hani Abidi (Research Fellow, Aga Khan University) for constructing the phylogenetic tree using PhyML 3.0.

Authors' contributions

S.A. participated in the study design, methodology, data analysis and write up. Iqbal MP conceived the idea and supervised the whole study. S.Z. performed phylogenetic analysis, helped with the write up, and provided expert advice. Z.A.B. read the manuscript and gave important input. All authors reviewed and approved the final manuscript.

Supplemental Material

Supplemental material may be found here: https://www.landesbioscience.com/journals/idp/article/27450/

References

1. White P, Cooke N. The multifunctional properties and characteristics of vitamin D-binding protein. Trends Endocrinol Metab 2000; 11:320-7; PMID:10996527; http://dx.doi.org/10.1016/S1043-2760(00)00317-9
2. Chishimba L, Thickett DR, Stockley RA, Wood AM. The vitamin D axis in the lung: a key role for vitamin D-binding protein. Thorax 2010; 65:456-62; PMID:20435872; http://dx.doi.org/10.1136/thx.2009.128793
3. Speeckaert M, Huang G, Delanghe JR, Taes YE. Biological and clinical aspects of the vitamin D binding protein (Gc-globulin) and its polymorphism. Clin Chim Acta 2006; 372:33-42; PMID:16697362; http://dx.doi.org/10.1016/j.cca.2006.03.011
4. Verboven C, Rabijns A, De Maeyer M, Van Baelen H, Bouillon R, De Ranter C. A structural basis for the unique binding features of the human vitamin D-binding protein. Nat Struct Biol 2002; 9:131-6; PMID:11799400; http://dx.doi.org/10.1038/nsb754
5. Cooke NE, David EV. Serum vitamin D-binding protein is a third member of the albumin and alpha fetoprotein gene family. J Clin Invest 1985; 76:2420-4; PMID:2416779; http://dx.doi.org/10.1172/JCI112256
6. Gozdzik A, Zhu J, Wong BY, Fu L, Cole DE, Parra EJ. Association of vitamin D binding protein (VDBP) polymorphisms and serum 25(OH)D concentrations in a sample of young Canadian adults of different ancestry. J Steroid Biochem Mol Biol 2011; 127:405-12; PMID:21684333; http://dx.doi.org/10.1016/j.jsbmb.2011.05.009
7. Song YH, Naumova AK, Liebhaber SA, Cooke NE. Physical and meiotic mapping of the region of human chromosome 4q11-q13 encompassing the vitamin D binding protein DBP/Gc-globulin and albumin multigene cluster. Genome Res 1999; 9:581-7; PMID:10400926
8. Weitkamp LR, Allen PZ. Evolutionary conservation of equine gc alleles and of Mammalian gc/albumin linkage. Genetics 1979; 92:1347-54; PMID:17248956
9. Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L, et al. The genome of the Western clawed frog Xenopus tropicalis. Science 2010; 328:633-6; PMID:20431018; http://dx.doi.org/10.1126/science.118367
10. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003; 52:696-704; PMID:14530136; http://dx.doi.org/10.1080/10635150390235520
11. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010; 59:307-21; PMID:20525638; http://dx.doi.org/10.1093/sysbio/syq010
12. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011; 39:D225-9; PMID:21109532; http://dx.doi.org/10.1093/nar/gkq1189
13. Noël ES, Reis MD, Arain Z, Ober EA. Analysis of the Albumin/alpha-Fetoprotein/Afamin/Group specific component gene family in the context of zebrafish liver differentiation. Gene Expr Patterns 2010; 10:237-43; PMID:20471496; http://dx.doi.org/10.1016/j.gep.2010.05.002
14. Yang F, Brune JL, Naylor SL, Cupples RL, Naberhaus KH, Bowman BH. Human group-specific component (Gc) is a member of the albumin family. Proc Natl Acad Sci U S A 1985; 82:7994-8; PMID:2415977; http://dx.doi.org/10.1073/pnas.82.23.7994
15. Cooke NE, Haddad JG. Vitamin D binding protein (Gc-globulin). Endocr Rev 1989; 10:294-307; PMID:2476303; http://dx.doi.org/10.1210/edrv-10-3-294