Defensins identified through molecular de-extinction

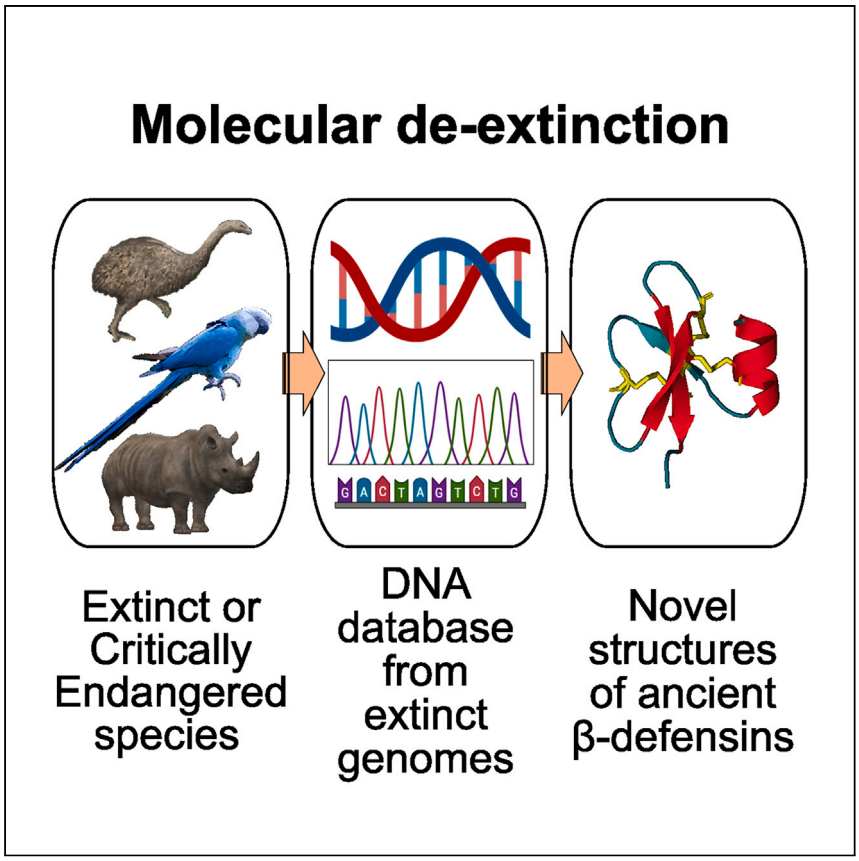
Adryan F.L. Ferreira, Karen O Osiro, Kamila B.S. de Oliveira, Marlon H Cardoso, Lucas R de Lima, Harry M Duque, Maria L.R. Macedo, Céline Landon, Cesar de la Fuente-Nunez, Octavio L Franco

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Defensins identified through molecular de-extinction



Molecular de-extinction aims to discover useful molecules throughout evolution. Through genome mining, using a combination of computational tools, Ferreira et al. identify six β-defensins derived from the extinct New Zealand moa, the extinct-in-the-wild Spix's macaw, and the critically endangered western black rhino.

Adryan F.L. Ferreira, Karen O. Osiro, Kamila B.S. de Oliveira, ..., Céline Landon, Cesar de la Fuente-Nunez, Octavio L. Franco

cfuente@upenn.edu (C.d.l.F.-N.)
ocfranco@gmail.com (O.L.F.)

Highlights

Molecular de-extinction identified β-defensins in extinct genomes

β-defensins are mined from extinct avians and a critically endangered mammal

This study provides new insights into the molecular evolution of defensins

Ferreira et al., Cell Reports Physical Science 5, 102193
September 18, 2024
Published by Elsevier Inc.
https://doi.org/10.1016/j.xcrp.2024.102193

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Defensins identified through molecular de-extinction

Adryan F.L. Ferreira,[1,9] Karen O. Osiro,[2,9] Kamila B.S. de Oliveira,[1,9] Marlon H. Cardoso,[1,2,3] Lucas R. de Lima,[1] Harry M. Duque,[2] Maria L.R. Macedo,[3] Céline Landon,[4] Cesar de la Fuente-Nunez,[5,6,7,8,*] and Octavio L. Franco[1,2,10,*]

SUMMARY

Molecular de-extinction is an emerging field that identifies potentially useful molecules throughout evolution. Here, we computationally mine genomes, searching for molecules called defensins, which play a role in host immunity. Our approach leads to the discovery of six undescribed β-defensins, five of which are derived from two different extinct bird species and one from a mammalian species. These organisms included an extinct moa species (*Anomalopteryx didiformis*) that inhabited New Zealand and the extinct Spix's macaw (*Cyanopsitta spixii*), which was endemic to Brazil, as well as the black rhino (*Diceros bicornis minor*). Evolutionary and structural analyses of the β-defensins are performed to further characterize these molecules. This study identifies molecules from extinct organisms, revealing defensins and opening new avenues for antibiotic discovery.

INTRODUCTION

Molecular de-extinction has emerged as a new field seeking to discover molecules through evolutionary history and use them to solve present-day problems.[1–4] Defensins are small disulfide-rich cationic proteins that play a key role in the defense mechanisms of living organisms and have been proposed as potential new antibiotics.[5] They are naturally produced by microorganisms, plants, vertebrates, and invertebrate animals.[6] Defensins are highly conserved molecules among species, with specific structural folds and highly conserved disulfide patterns, which make them amenable for genome mining efforts.[7] Most defensins have 2–6 disulfide bonds, which helps maintain their functional conformation.[8] Furthermore, defensins are classified into two superfamilies, *cis*- and *trans*-defensins, based on their structural similarity and topology. Each superfamily has its independent evolutionary origin, which underwent extensive divergent evolution in terms of sequence, structure, and function.[5] In vertebrates, defensins belong to the *trans*-defensins superfamily and are classified as α-, β-, and θ-defensins.[9] β-defensins are present in all vertebrates and represent the largest family of vertebrate defensins except for double-size AvBD11 proteins.[10] They generally comprise 35–50 residues with a core of three anti-parallel β-strands stabilized by three disulfide bonds (Cys1–Cys5, Cys2–Cys4, and Cys3–Cys6).[11] These disulfide bonds provide stability to the peptide's three-dimensional structure. Functionally, defensins act as effectors of innate immunity and enhancers of antigen-specific cellular and humoral immune responses, protecting the skin and mucous membranes of the respiratory, genitourinary, and gastrointestinal systems.[12]

Here, we computationally mined eight extinct vertebrate genomes, searching for defensin molecules and examining their evolution and structure. Studying the

[1]S-Inova Biotech, Pós-graduação Em Biotecnologia, Universidade Católica Dom Bosco, Campo Grande, Mato Grosso do Sul, Brazil

[2]Centro de Análises Proteômicas e Bioquímicas, Pós-graduação Em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília, Brazil

[3]Laboratório de Purificação de Proteínas e Suas Funções Biológicas, Universidade Federal de Mato Grosso do Sul, Cidade Universitária, Campo Grande 79070900, Mato Grosso do Sul, Brazil

[4]Centre for Molecular Biophysics, National Center for Scientific Research (CNRS), Orléans, France

[5]Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[6]Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA

[7]Department of Chemistry, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA

[8]Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA, USA

[9]These authors contributed equally

[10]Lead contact

*Correspondence: cfuente@upenn.edu (C.d.l.F.-N.), ocfranco@gmail.com (O.L.F.)

https://doi.org/10.1016/j.xcrp.2024.102193

Table 1. Summary of results during the identification process of defensins through molecular de-extinction

Species	Type of genome	No. of genes identified by AUGUSTUS	Potential defensins identified by HMMER	Confirmed defensin by INTERPRO	Defensin type
T. cynocephalus	complete	48,424	12	–	–
D. bicornis minor	complete	42,479	13	1	β
H. giga	complete	94,556	3	–	–
A. didiformis	complete	18,369	3	2	β
C. spixii	complete	30,273	6	3	β
M. giganteus	partial	37	–	–	–
B. anhydra	partial	1,013	–	–	–
C. campestris	partial	4,861	–	–	–

relationship between defensins from extinct and extant animals may shed light on their biological role throughout evolution. By examining structural information, including secondary structure orientation, cysteine motifs, disulfide bond connectivity, tertiary structure similarities, and precursor gene sequence, we can gain a better understanding of their evolutionary relationships. This work employs different *in silico* tools to prospect and evaluate the structural and functional connections of β-defensins from extinct animals. We also perform phylogenetic and molecular dynamics (MD) analyses, exploring their evolutionary pathways and characterizing their structures.

Collectively, we find dozens of potential defensins, out of which six are identified as authentic defensins. The structures for the six β-defensins are predicted, two of which are derived from the extinct New Zealand moa (*Anomalopteryx didiformis*), three from extinct-in-the-wild Spix's macaw (*Cyanopsitta spixii*), and one from the western black rhino (*Diceros bicornis minor*), which is classified as a critically endangered species. By integrating comprehensive phylogenetic analyses and MD simulations, we not only reveal structures of ancient β-defensins but also gain insights into their evolutionary trajectories. Molecular de-extinction efforts open new avenues for understanding molecular structure and function throughout evolution.

RESULTS

Non-mitochondrial genomic analysis uncovers protein-coding genes in eight extinct and endangered species

Initially, 154 extinct or endangered animals were found through our mining approach. However, our analysis subsequently excluded all mitochondrial-based sequences since mitochondrial genomes are not capable of expressing defensins, narrowing down the list to eight distinct animals, all of which possess non-mitochondrial genomic data accessible in the National Center for Biotechnology Information (NCBI). The selected species were the Tasmanian tiger *Thylacinus cynocephalus*, western black rhino *D. bicornis minor*, Steller's sea cow *Hydrodamalis gigas*, New Zealand moa *A. didiformis*, Spix's macaw *C. spixii*, the giant deer *Megaloceros giganteus*, desert bettong *Bettongia anhydra*, and the desert rat-kangaroo *Caloprymnus campestris*. Genomic analysis in search of protein-coding genes by the AUGUSTUS program resulted in 48,424 genes found in *T. cynocephalus*, 42,479 in *D. bicornis minor*, 94,556 in *H. gigas*, 18,369 in *A. didiformis*, 30,273 in *C. spixii*, 37 in *M. giganteus*, 1,013 in *B. anhydra*, and 4,861 in *C. campestris* (Table 1). The lack of information in the partial genomes added a limitation in the computational analysis, making a contrast between the number of genes and the peptide identification.

Table 2. Physicochemical data of prospective defensins

Species	β-Defensins	De-extinction nomenclature	No. of amino acid residues	GRAVY[a]	Molecular mass (Da)	Negative residues	Positive residues	Charges
A. didiformis	*Ad*-AvBD5	AvBD5 *didiformin*	40	−0.422	4,428.11	4	7	+3
A. didiformis	*Ad*-AvBD10	AvBD10 *didiformin*	45	−0.320	4,844.49	4	5	+1
C. spixii	*Cs*-AvBD1	AvBD1 *spixiin*	38	−0.161	4,285.17	1	8	+7
C. spixii	*Cs*-AvBD10	AvBD10 *spixiin*	46	−0.276	4,849.54	3	5	+2
C. spixii	*Cs*-AvBD9	AvBD9 *spixiin*	43	−0.232	4,545.17	3	6	+3
D. bicornis minor	*Db*-BD4	Db-BD4 *bicornin*	34	0.079	3,647.3	0	5	+5

[a]The GRAVY (Grand Average of Hydropathy) values are defined as the average hydropathy value of a peptide or protein. Proteins with negative GRAVY values are hydrophilic, and positive values mean they are hydrophobic.

Molecular de-extinction reveals six β-defensins

The HMMER data analysis revealed significant variations in the number of potential defensins. The mammals *D. bicornis minor* and *T. cynocephalus* exhibited the highest count, with 13 and 12 possible defensins, respectively. The mammal *H. gigas* showed a more limited number of potential defensins, predicting only three. No potential defensins were found in *M. giganteus*, *B. anhydra*, or *C. campestris* animals. In avian species, three potential defensins were identified in *A. didiformis* and six in *C. spixii*. Among the possible defensins found, only six were validated by the INTERPRO according to inherent defensin's disulfide binding patterns (Table 1): two β-defensins from *A. didiformis*, three from *C. spixii*, and one from *D. bicornis minor*. Several potential defensins were excluded during the validation process by the InterProScan program because they were not classified as authentic defensins due to the display of different disulfide binding patterns, the cysteine patterns, and the identified function. The defensins from *A. didiformis* were named *Ad*-AvBD5 and *Ad*-AvBD10. The defensins from *C. spixii* were designated *Cs*-AvBD1, *Cs*-AvBD9, and *Cs*-AvBD10. Likewise, the *D. bicornis minor* defensin was named *Db*-BD4 (Table 2).

Sequence alignment and phylogenetic tree analysis of de-extinct β-defensins

Avian species exhibited the highest number of potential defensins validated by INTERPRO, which identified and confirmed 9 possible defensins. Notably, all identified defensins belonged to the β-type, presenting a theoretical core of three antiparallel β-strands stabilized by three disulfide bonds (Figures 1 and 2). Analyzing the multiple alignments of avian β-defensins carried out by the T-COFFEE server, we noticed the similarity between the β-defensin sequences of prospected birds and their parent defensin. The β-defensins were arranged according to their similarities in their primary structure, aligning all the primary structures and the cysteine pairing (Cys1–Cys5, Cys2–Cys4, and Cys3–Cys6) (Figures 1 and 2). In the multiple alignments performed by T-COFFEE to analyze mammalian β-defensins, we focused on detection similarities between the prospected defensin and its parents. Our results revealed partial similarity of their sequences, particularly in the positions of cysteines forming the disulfide bonds, β-strands, and α-helix structures, which aligns with the expected pattern observed in β-defensins (Figures 1 and 2).

The avian β-defensin phylogenetic tree revealed that *Ad*-AvBD10 and *Cs*-AvBD9 were highly conserved, indicating their proximity to the root ancestor. Conversely, β-defensins *Ad*-AvBd5, *Cs*-AvBd1, and *Cs*-AvBd10 appeared to represent more recent evolutionary developments (Figure 3).

For birds, the common chicken *Gallus gallus* possesses 14 defensin genes known as AvBD1-14 (also referred to as GLL1-14).[13] In this study, the two β-defensins

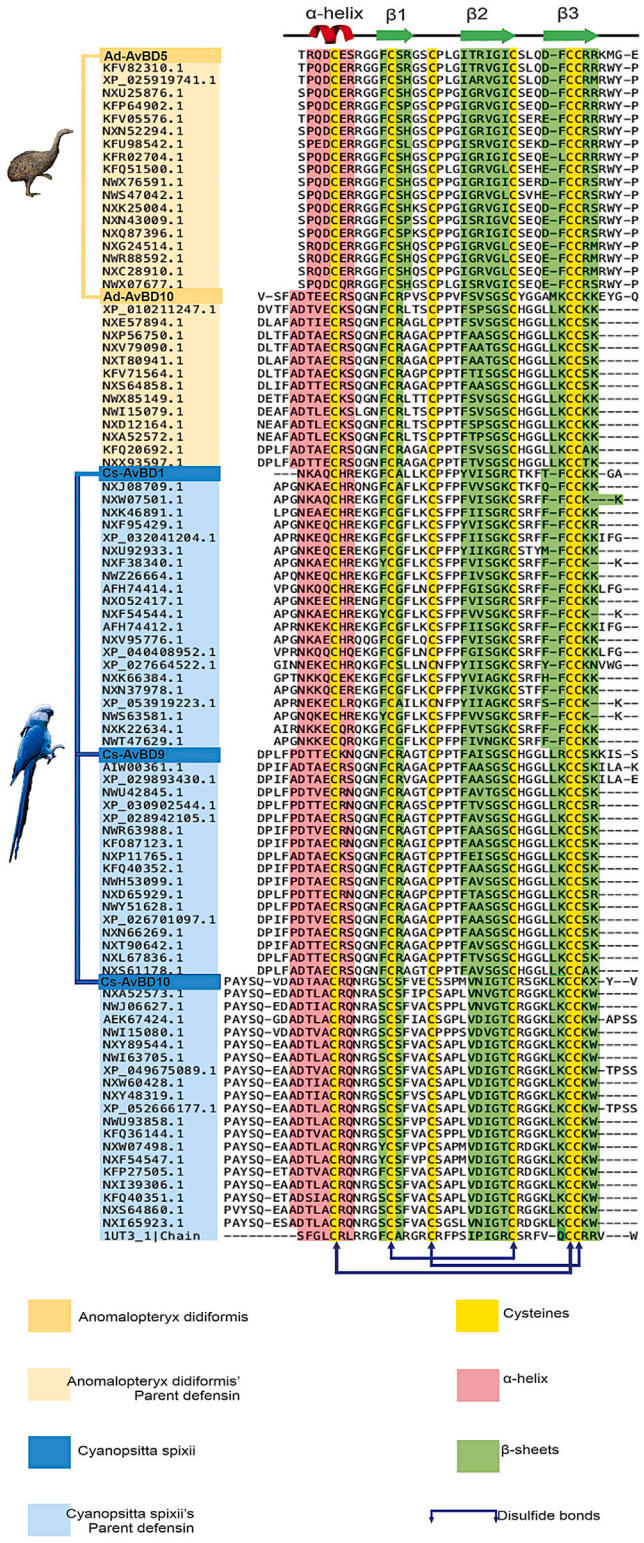


Figure 1. Sequence alignments of avian β-defensins

A protein blast was performed with each avian defensin prospected, and the sequences of different species with more than 80% of query cover were selected in a total of 19 samples for each prospected defensin. Multiple alignments were carried out on the T-COFFEE server with the

Figure 1. *Continued*
prospected defensin sequences from *C. spixii* (dark blue), the *C. spixii*'s parent defensin (light blue), *A. didiformis* (dark orange), and *A. didiformis*'s parent defensin (orange) birds and 19 more parent sequences of each prospected defensin, for a total of 100 aligned sequences. Predicted secondary structure elements (α-helix, β-strands) are displayed in orange and green, respectively. Cys residues are displayed in yellow, with arrows showing the disulfide connectivity. The animals depicted were obtained from Wikipedia, licensed under CC BY 4.0 International (https://creativecommons.org/licenses/by/4.0/). Figure editing was performed using Paint 3D by Microsoft Corporation, Inkscape 1.2, and CorelDRAW.

identified in *A. didiformis*, *Ad*-AvBD5 and *Ad*-AvBD10, shared close evolutionary relationships with GLL5 and GLL10, respectively (Figure 3). Likewise, the evolutionary trajectory of *Cs*-AvBD10 was closely related to *Ad*-AvBD10 in terms of its similarity to GLL10. However, despite the shared homology with the common avian β-defensin GLL10, it is noteworthy that GLL10 from the Dinornithiformes (*Ad*-AvBD10) branch revealed an ancient lineage, in contrast to the GLL10 variant within the Psittaciformes (*Cs*-AvBD10) order (Figure 3). Both *Ad*-AvBD10 and *Cs*-AvBD9 defensins were settled in a distinct branch, positioned closer to the phylogenetic tree root, and are the most ancient β-defensins among the prospected bird β-defensins described in this study (Figure 3). Although *Cs*-AvBD9 stands out as a solitary and conserved branch, it seems to be related to GLL9. Moreover, the placement of *Cs*-AvBD1 in the tree suggests a relationship with the avian AMP1 β-defensin family, a lineage that appears to have significantly diverged from the other defensins examined here, forming an isolated root. Conversely, *Ad*-AvBD5 displays a more recent evolutionary divergence, finding itself situated alongside a common ancestor shared with GAL5 from the *Struthio camelus australis*, which belongs to the Palaeognathae infraclass. *A. didiformis* also belongs to this same infraclass (Figure 3).

Mammalian β-defensins were also analyzed. We found that black rhino *Db*-BD4 was classified as one of the most recent in its tree and remained close to β-defensin-4 from *Mastomys coucha* (Figure 4). According to the phylogenetic tree results, the identified β-defensin from *D. bicornis minor, Db*-BD4, presented the highest homology with β-defensin 4 of Rodentia (e.g., *M. coucha*), Primates (e.g., *Sapajus apella*, *Cebus imitator*), and Perissodactyla (e.g., *Equus caballus*) orders (Figure 4). In general, all defensins mined in this study demonstrated genetic similarity to their closest parent, as determined by high bootstrap and Bayesian values (Figure 4).

Cationicity and amphipathy features confirm the β-defensin structures

The structures of all prospected β-defensins described here were predicted with AlphaFold2 and demonstrated a three-stranded anti-parallel β-sheet, a right-handed α-helix, and three disulfide bonds formed between Cys1–Cys5, Cys2–Cys4, and Cys3–Cys6 cysteines. The sequences showed clear physicochemical similarities except for cationic properties, which can be variable (Table 1; Figure 5).

Moreover, to assess structural stability, MD was performed during a period of 100 ns in high-salt content. Firstly, we evaluated the moa defensins. The *Ad*-AvBD5 defensin showed clear stability of its β-sheet structures (Figure 5A). The fluctuations, gyration, and deviation did not affect the secondary structure. The *Ad*-AvBD10 defensin maintained its structural conformation until 20 ns (Figure 5B). However, β-sheets were lost in some moments, as well as its α-helices, demonstrating that the molecule shows lower structural stability than *Ad*-AvBD5 (Figures 5A and 5B). The root-mean-square fluctuation (RMSF) graph (Figure S1) showed large fluctuations in residues 1–5 (Val-Ser-Phe-Ala-Asp) and 42–44 (Tyr-Gly-Gln), but there were large fluctuations in the RMSD and radius of gyration (RG) graphs. By contrast, Spix's macaw

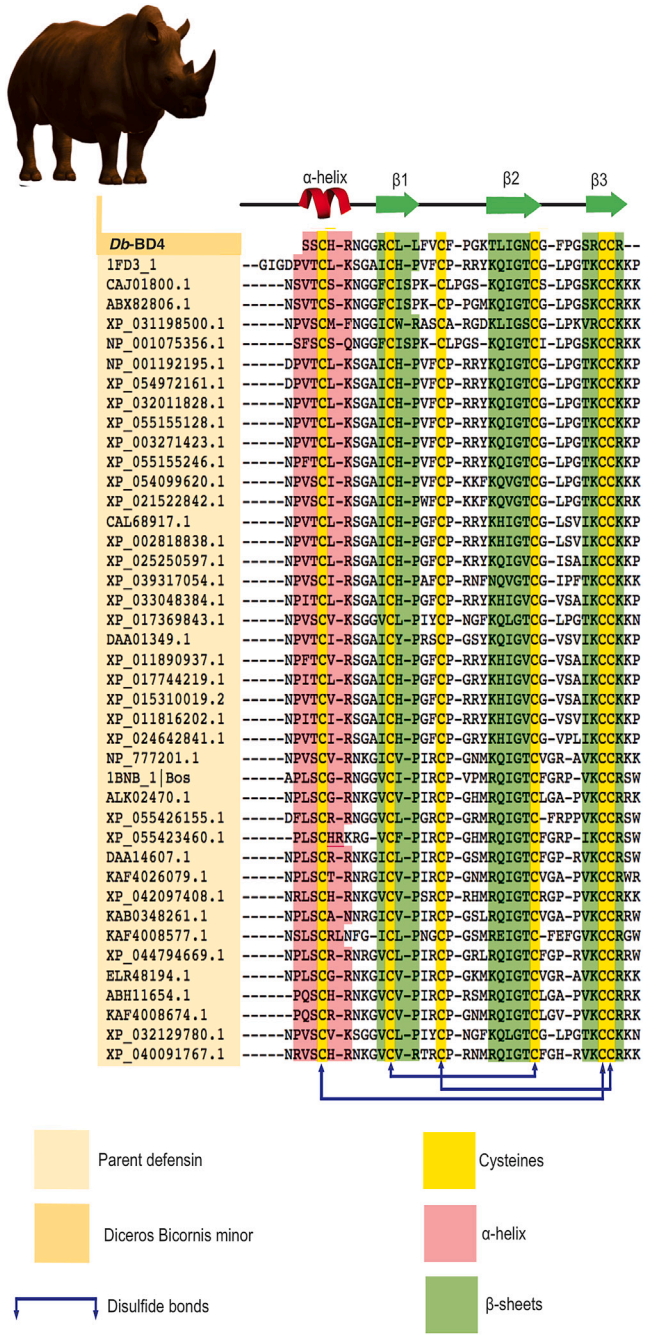


Figure 2. Sequence alignments among mammalian β-defensins

A protein blast was performed with the defensin sequence of *D. bicornis minor*. Due to the limited sampling of species, a BLAST search was conducted with the most representative species, selecting sequences with over 80% query coverage from a total of 41 samples. Multiple alignments were carried out on the T-COFFEE server, including one defensin sequence prospected from *D. bicornis minor* (dark orange) and another 41 parent sequences (orange) from each defensin prospected, totaling 42 aligned sequences. Predicted secondary structure elements (α-helix, β-strands) are displayed in red and green, respectively. Cys residues are displayed in yellow, with arrows showing the disulfide connectivity. The animals depicted were obtained from Wikipedia, licensed under CC BY 4.0 International (https://creativecommons.org/licenses/by/4.0/). Figure editing was performed using Paint 3D by Microsoft Corporation, Inkscape 1.2, and CorelDRAW.

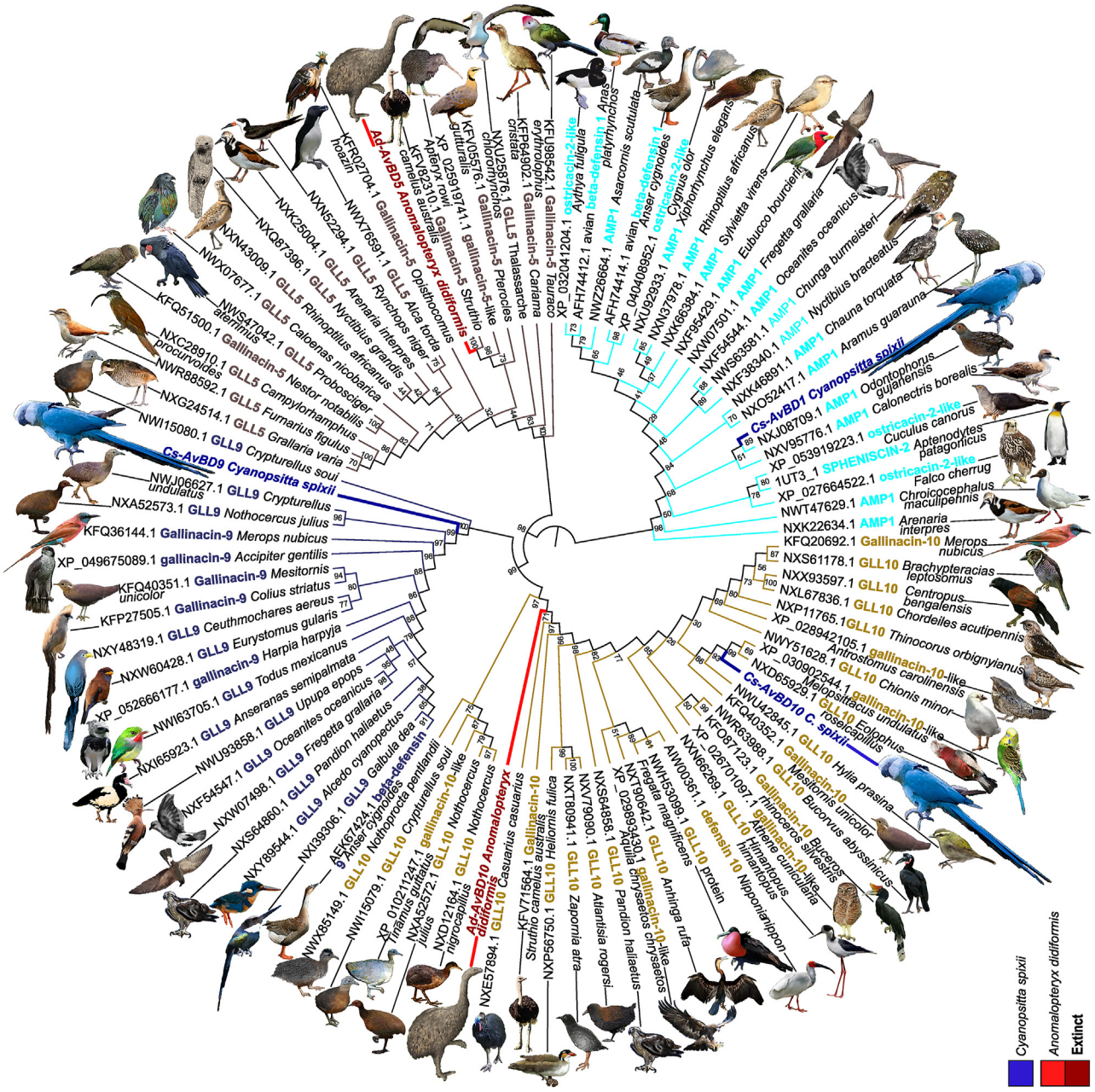


Figure 3. Phylogenetic tree of avian β-defensins

The phylogenetic tree was generated by the IQ-TREE server and organized with the FigTree program using the prospected defensin sequences from *C. spixii* and *A. didiformis* birds and another 19 parent sequences of each prospected defensin, for a total of 100 aligned sequences. The animals depicted were obtained from Wikipedia, licensed under CC BY 4.0 International (https://creativecommons.org/licenses/by/4.0/). Figure editing was performed using Paint 3D by Microsoft Corporation, Inkscape 1.2, and CorelDRAW.

β-defensins also showed a variation in structural stability when evaluated by MD. High structural stability was observed throughout the MD simulations for *Cs*-AvBD1 and *Cs*-AvBD10, where secondary structures were maintained until the end of the simulations (Figures 5C and 5E). *Cs*-AvBD9 had its secondary structures affected after 20 ns, and one of the β-strands was unfolded (Figure 5D). The *Cs*-AvBD9 average fluctuation varied between 0.05 and 0.63 nm simulation with a high RG modification between 0.9 and 1.3 nm (Figure S1). Finally, for the black rhino

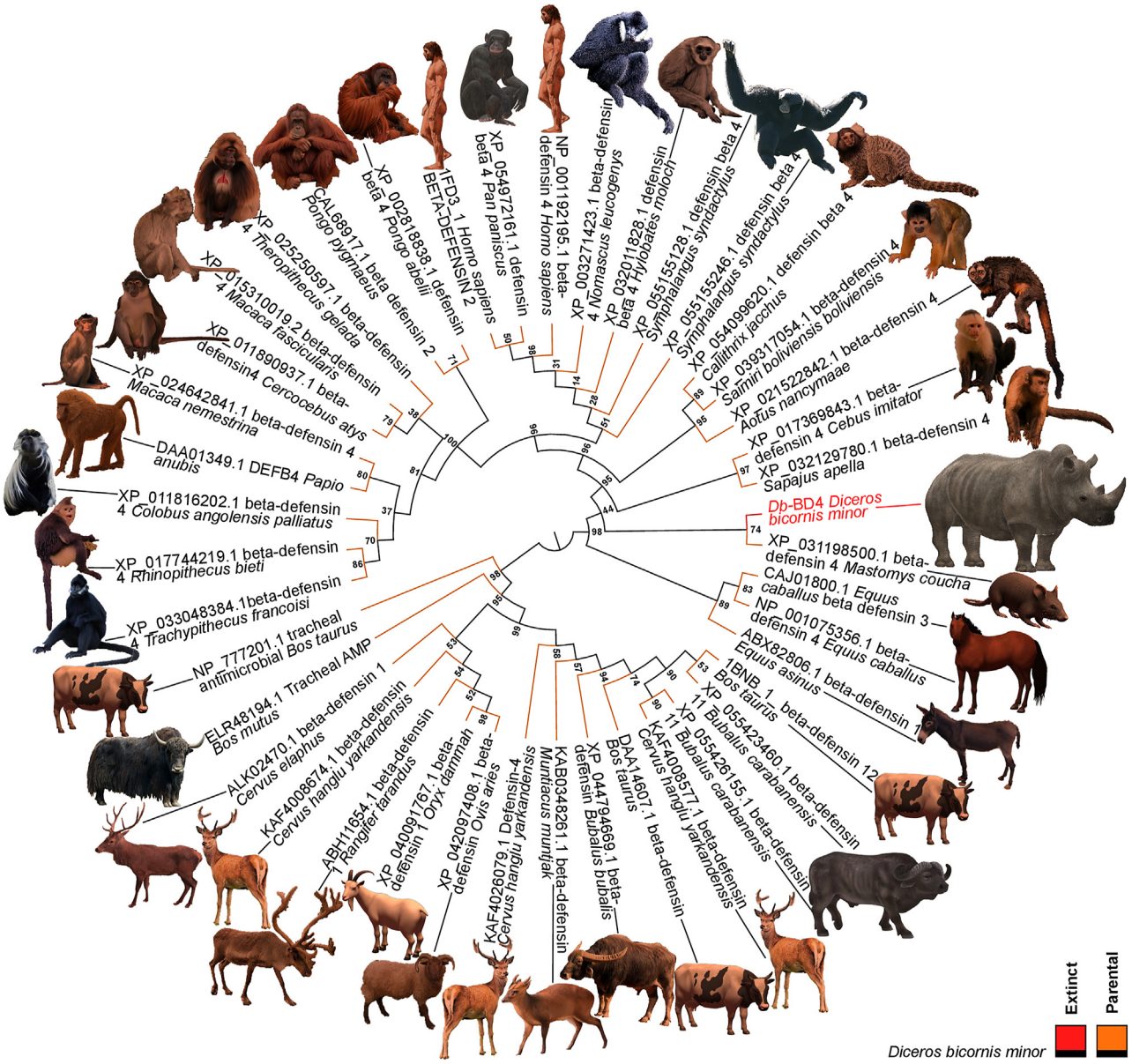


Figure 4. Phylogenetic tree of mammalian β-defensins

Firstly, a blast was made with the defensin sequence of *D. bicornis minor*. However, due to the small sampling of species, the blast was performed with the most representative species, and the sequences with more than 80% of the query cover were selected. The sequences were aligned by the T-COFFEE server, and with the alignment file, the phylogenetic tree was made by the IQ-TREE server and organized with the FigTree program using the defensin of *D. bicornis* minor and 41 other parental sequences of each prospected defensin. The animals depicted were obtained from Wikipedia, licensed under CC BY 4.0 International (https://creativecommons.org/licenses/by/4.0/). Figure editing was performed using Paint 3D by Microsoft Corporation, Inkscape 1.2, and CorelDRAW.

β-defensin *Db*-BD4, there was a destabilization of the α-helix at 50 ns, becoming a random coil and recovering the α-helical structure at 100 ns (Figure 5F). However, the RMSD for *Db*-BD4 demonstrated that the deviations did not undergo large oscillations (Figure S1). Therefore, the secondary structures of defensins *Ad*-AvBD5, *Cs*-AvBD1, and *Cs*-AvBD10 presented high stability, whereas *Ad*-AvBD10, *Cs*-AvBD9, and *Db*-BD4 significantly lost their structures, according to the MD simulations.

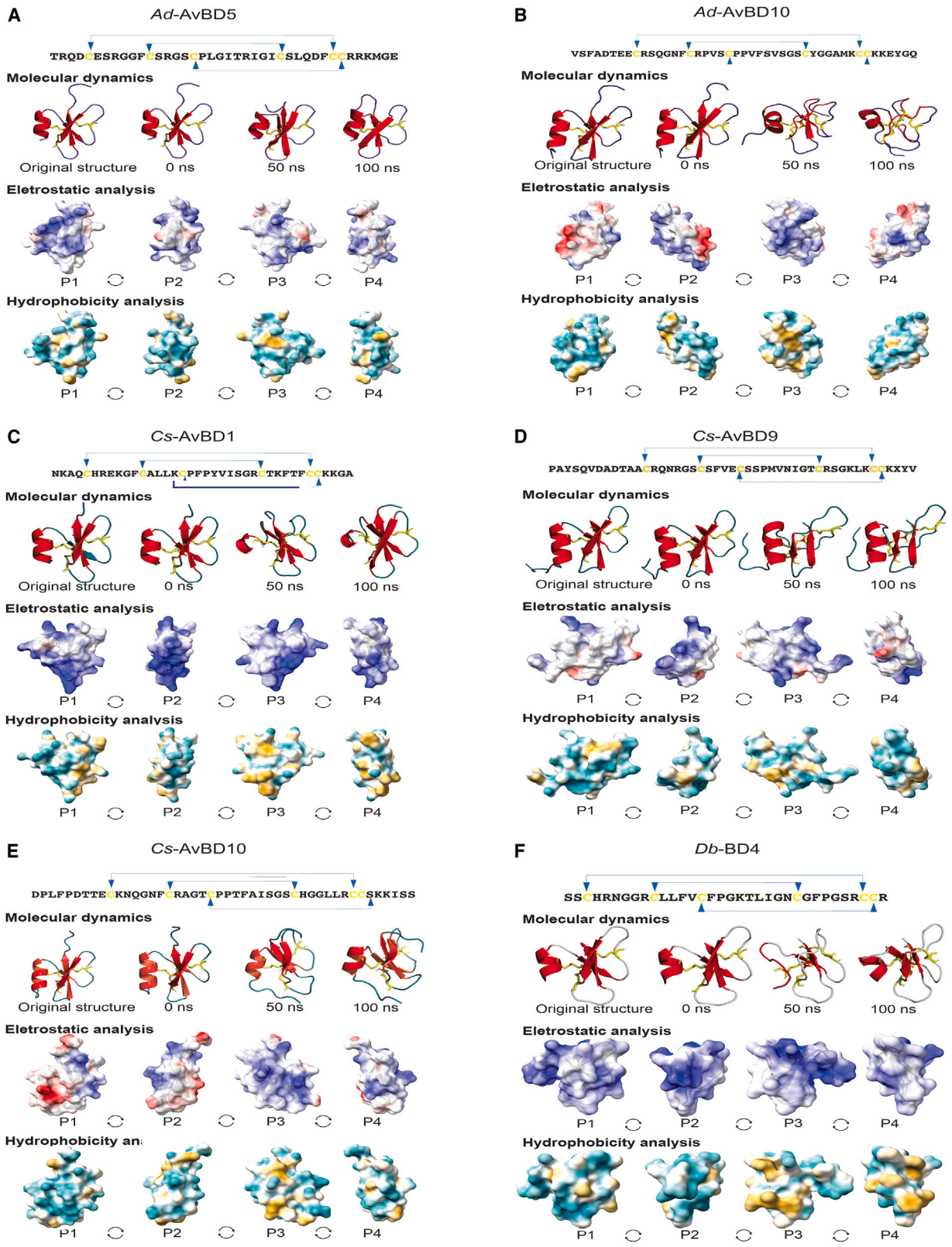

A
Ad-AvBD5
TRQDCESRGGFCSRGSCPLGITRIGICSLQDFCCRRKMGE
Molecular dynamics
Original structure
0 ns
50 ns
100 ns
Eletrostatic analysis
P1
P2
P3
P4
Hydrophobicity analysis
P1
P2
P3
P4
B
Ad-AvBD10
VSFADTEECRSQGNFCRPVSCPPVFSVSGSCYGGAMKCCKKEYGQ
Molecular dynamics
Original structure
0 ns
50 ns
100 ns
Eletrostatic analysis
P1
P2
P3
P4
Hydrophobicity analysis
P1
P2
P3
P4
C
Cs-AvBD1
NKAQCHREKGFCALLKCPFPYVISGRCTKFTFCCKKGA
Molecular dynamics
Original structure
0 ns
50 ns
100 ns
Eletrostatic analysis
P1
P2
P3
P4
Hydrophobicity analysis
P1
P2
P3
P4
D
Cs-AvBD9
PAYSQVDADTAACRQNRGSCSFVECSSPMVNIGTCRSGKLKCCKXYV
Molecular dynamics
Original structure
0 ns
50 ns
100 ns
Eletrostatic analysis
P1
P2
P3
P4
Hydrophobicity analysis
P1
P2
P3
P4
E
Cs-AvBD10
DPLFPDTTECKNQGNFCRAGTCPPTFAISGSCHGGLLRCCSKKISS
Molecular dynamics
Original structure
0 ns
50 ns
100 ns
Eletrostatic analysis
P1
P2
P3
P4
Hydrophobicity an
P1
P2
P3
P4
F
Db-BD4
SSCHRNGGRCLLFVCFPGKTLIGNCGFPGSRCCR
Molecular dynamics
Original structure
0 ns
50 ns
100 ns
Eletrostatic analysis
P1
P2
P3
P4
Hydrophobicity analysis
P1
P2
P3
P4

Figure 5. Predicted 3D structures of β-defensins

The interconnected arrows at the peptides' sequences represent the pairing between the cysteines, forming disulfide bonds (cysteine—letter C— is highlighted in yellow). Conceptual structure snapshots derived from molecular dynamic simulation (secondary structure elements in red and disulfide bridges in yellow). Visualization of 0, 50, and 100 ns intervals of each prospected defensin during the molecular dynamic simulation, showing the stability of all the secondary structures in simulation. Electrostatic and lipophilic potentials were determined using chimera X[14] on defensin surfaces (90° rotation between each face). The electrostatic potential is represented by coloring ranges. Red corresponds to the negative potential, white corresponds to the neutral surfaces, and blue corresponds to the positive potential. Lipophilic potentials are also represented by coloring ranges. Dark gold corresponds to the most hydrophobic zones, white corresponds the neutral surfaces, and cyan corresponds to the most hydrophilic potentials. The predicted 3D structures of the β-defensins were generated using PyMOL, v.3.0 (https://pymol.org).

Structural and physicochemical properties analysis revealed that the *Ad*-AvBD10 sequence showed a global charge of +1, although the local charge distribution was uneven, with 4 negative and 5 positive residues. In different orientations, the surface exhibited variation, being more negative in the first orientation and globally positive in the third orientation (Figure 5B). Similarly, *Cs*-AvBD10, with global charge +2, presented an uneven distribution, with 3 negative and 5 positive residues (Figure 5E). By contrast, *Cs*-AvBD1 exhibited a notable overall charge of +7, characterized by eight cationic residues (Figure 5C). Its surface was predominantly positive, highlighting an absence of the negative regions observed in previous defensins (Figure 5C). This pattern was also evident in *Db*-BD4 (Figure 5F).

Regarding hydrophobicity, our analysis revealed that *Ad*-AvBD10 has a face with a notable cluster of hydrophobic residues. Furthermore, we observed that in specific orientations, *Db*-BD4 showed hydrophobic surfaces, opposite to the positively charged faces in other orientations.

DISCUSSION

Molecular de-extinction has enabled the identification of molecules from extinct organisms.[1] This emerging field, enabled by advances in computational biology and machine learning, has already mined archaic hominids as well as the entire "extinctome."[1,15] Here, we explored the genomes of animals regarded as extinct, including *A. didiformis*, which became extinct approximately 600 years ago,[16] as well as those extinct in the wild, and critically endangered animals.

We explored more than 150 datasets of extinct or endangered animals, looking for complete or partially complete non-mitochondrial genomic data. Six β-defensins were unambiguously highlighted in this study, five from two avian species: the extinct New Zealand moa *A. didiformis* and the extinct-in-the-wild Spix's macaw *C. spixii*, and one mammal β-defensin from the critically endangered western black rhino *D. bicornis minor*.

It has been hypothesized that β-defensins originally emerged over 520 million years ago from big ancestral defensins akin to defensins identified in urochordates, cephalochordates, and vertebrates.[12] β-defensin genes have been identified in all vertebrate species, whereas α- and θ-defensins could be regarded as more contemporary variations, with α-defensins appearing in mammals and θ-defensins being exclusive to some great primates.[17] Phylogenetic analyses have revealed frequent changes in the number of β-defensin genes, with species-specific gene gains and losses throughout vertebrate evolution. The number of β-defensin genes in a species has been linked to changing environmental microbial challenges.[18] The avian immune system rose from a span of over 310 million years of divergent evolution between birds and mammals, being more compact than that of primates yet resembling its structure and function.[17] Following the Cretaceous-Paleogene (K-Pg) mass extinction event around 66 million years ago, avian lineages

underwent a rapid evolutionary process, in contrast to the gradual evolution of primates.[12,17,19,20]

Vertebrate genomes contain a series of genes for defensins exhibiting an extensive range of tissue expression patterns. They are situated together in a singular cluster positioned at the terminal region of chromosome 3, acting as functional equivalents of mammalian β-defensins.[21] Among them, ten genes (AvBD2, -4, -5, -7, -8, -9, -10, -11, -12, and -13) point toward high conservation spanning over 100 million years, maintaining a 1:1 orthologous relationship across the majority of species studied.[22] In this study, two avian β-defensins, *Ad*-AvBD10 and *Cs*-AvBD10, seem to be closely related to GLL10 (Figure 3). However, despite these β-defensins showing homology to the same β-defensin, AvBD10 demonstrates a significant distance between the Dinornithiformes (*Ad*-AvBD10) and Psittaciformes (*Cs*-AvBD10) orders (Figure 3). These findings indicate that AvBD10 from *A. didiformis* has a more ancient position in the evolutionary chronological sequence compared to AvBD10 from *C. spixii*.

Indeed, members of the β-defensin family display poor sequence similarity but present a notable degree of similarity in their tertiary structures.[23] Variation in homology exists across sequences from various species within orthologous pairs, typically ranging from 75% to 95%. For instance, the amino acid identity between duck AvBD2 and its chicken ortholog reaches 76.5%, whereas quail and chicken AvBD10 show 84% similarity.[20,24,25] Besides this, *Ad*-AvBD10 traces its lineage back to *A. didiformis*, an extinct ratite moa that was an important species in the New Zealand ecosystem until its extinction ≈600 years ago.[16] As mentioned before, AvBD10 is a highly conserved gene often observed on a distinct and isolated branch, closely associated with an ancestral lineage, indicating its status as an ancient β-defensin.[21,26–29] The extensive branch length observed in the *Ad*-AvBD10 tree (Figure 3) is corroborated by other phylogenetic analyses of AvBD10,[21,22,29] as well as from the deep-rooted history of moa evolution.[30] Additionally, an analysis of 12S and ND6 sequences suggests a notably limited level of genetic diversity within these ratites,[31] further reinforcing *Ad*-AvBD10's ancient history. Conversely, *Ad*-AvBD5 demonstrates a closer association with gallinacin-5 (Figure 3), a β-defensin that is known to induce significantly elevated immunoglobulin (Ig)A and IgY antibody titers in chickens.[32]

Defensin *Cs*-AvBD1 forms a distinct clade positioned closer to AMP1 from other bird species. Its position implies its comparatively recent placement along the evolutionary timeline (Figure 3). AMP1 shares a close relation with AvBD1, GAL-1, and CHP-1 across chickens and other avian species, as well as THP-1 in turkeys and Osp-2 in ostriches.[19] Regarding *Cs*-AvBD9, this particular β-defensin is connected within the same branch as other GLL9 (Figure 3), which is also considered an ancient avian β-defensin.[22]

In structural analysis, all β-defensins had three anti-parallel β-sheets and a right-handed α-helix (Figure 5), with a conserved pattern of disulfide bridges that provides stability and structural maintenance.[12] MD analyses carried out using GROMACS revealed different behaviors of the molecules.[33] β-Defensins *Ad*-AvBD5, *Cs*-AvBD1, *Cs*-AvBD10, and *Db*-BD4 maintained their structures at the end of the 100 ns simulation, demonstrating greater stability and preserving their compaction. The disulfide bonds ensured a compact structure resistant to all torsional forces (visible until the end of the simulations), which, in particular, enabled the maintenance of intermolecular hydrogen interactions.[34] However, the *Ad*-AvBD10 and *Cs*-AvBD9 defensins lost part of their secondary structures, indicating some structural instability. In

the case of *Ad*-AvBD10, the gradual loss of the three strands within the β-sheet demonstrates a lower ability to stabilize its structures despite its disulfide bonds remaining until the end of the simulations. This was probably due to the deviation rate and the rotation of the molecule.[35] In *Cs*-AvBD9 defensin, only one of the β-strands was denatured, which could directly affect its partial stability.

An analysis of ancestral α-defensin sequences revealed coevolution between the acidic and basic amino acid domains in the mature defensin.[36] In addition to serving as a targeting signal, the prodomain plays a crucial role in folding and protection against auto-cytotoxicity. This protective role aims to preserve the extremely positive charge of mature defensins, avoiding deleterious interactions with lipids or other cellular proteins.[37] Therefore, it is reasonable to infer that the observed positive charge of β-defensins from extinct animals, as in the case of *Dd*-BD4, may have evolved as an adaptive strategy to preserve their anti-microbial efficacy over time.

The evolution of defensins in animals has resulted in several subfamilies of DLPs (defensin-like peptides) with neurotoxic functions. These DLPs share a similar structure to defensins, but their intercysteine loop sequences have been selected for specific interactions with ion channels.[38] These structural and functional adaptations may indicate an evolutionary diversification beyond anti-microbial functions, leading to more specific and selective interactions. Therefore, charge and hydrophobicity may have been shaped throughout evolution to optimize these new functions. Thus, the emergence of features such as α-toxins and additional cysteines may explain the variations in surface charges and amphiphilicity observed in defensins from extinct animals.

In conclusion, we computationally mined eight genomes and identified and revealed structures of ancient β-defensins from extinct genomes. We further gained insights into their evolutionary trajectories by integrating comprehensive phylogenetic analyses and MD simulations. This work may unlock new avenues for discovering defensins, enhancing our understanding of ancient biodiversity. Molecular de-extinction offers a promising avenue for the reconstruction and analysis of genetic and molecular data across evolutionary history.

EXPERIMENTAL PROCEDURES

Gene selection and prediction

We mined non-mitochondrial genomic data from extinct and critically endangered animals by leveraging GenBank (GenBank, https://www.ncbi.nlm.nih.gov/genbank/).[39] A total of eight extinct organisms with complete or partial non-mitochondrial genomic data were found[39,40] and selected for further evaluation. Each protein-coding sequence was then subsequently predicted using the human genome as a base for the AUGUSTUS web server (AUGUSTUS, v.3.5.0, https://bioinf.uni-greifswald.de/augustus/).[41,42] As a gene selection criterion in the target genomes, the parameters file "Human," available in the AUGUSTUS server, was used as a model to identify the intronless protein-coding genes. The predicted genes were then translated and exported from the program into protein sequences in FASTA format.[41,42]

In silico prospection and classification of extinct defensins

The predicted gene files from all six extinct, one extinct in the wild, and one critically endangered animal were filtered using the HMMER program (HMMER, v.3.4, http://hmmer.org),[43] which applies user-defined filters for analysis and homology detection enabled by hidden Markov models (HMMs).[44] β-Defensins were sought according to the animal that expresses them. Nine sequence filters obtained from the InterProScan

databank (InterProScan, v.5.67-99.0, https://www.ebi.ac.uk/interpro/)[45] were predefined. These filters were classified as AlphaBeta82 (82 α- and β-defensin sequences reviewed), Beta116 (116 β-defensin sequences reviewed and not reviewed), MDefensinFull338 (338 mammalian defensin sequences reviewed and not reviewed), MdefensinPfam196 (196 mammalian α-defensin sequences mammals), MReviewed69 (69 defensins found in mammals reviewed), ReviewedC63 (63 defensin sequences reviewed at the C terminus), and ReviewedN67 (67 defensin sequences reviewed at the N terminus). The 37 prospective defensin sequences were then submitted to the InterProScan program after filtering to perform sequence analysis and identify their function and classification.

Phylogenetic analysis

The β-defensin sequences identified were submitted to blastP to search for similar defensins and then integrated into a single FASTA file and submitted for alignment on the T-COFFEE server (T-COFFEE, v.13.46.0.919e8c6, https://tcoffee.org).[46] Two sequences were classified as extinct from *A. didiformis* species, three sequences as extinct in the wild from *C. spixii*, and one sequence as critically endangered from *D. bicornis minor*. These sequences were submitted for alignment refinement using the TCS server (https://tcoffee.crg.eu/apps/tcoffee/do:core)[47] with their respective parent sequences. The aligned and refined sequences were used as input on the IQ-TREE server (IQ-Tree, v.1.6.12, http://www.iqtree.org),[48,49] where the maximum likelihood phylogenetic tree was generated. The default values used by the server were 1,000 bootstraps alignment number, 1,000 repetitions, 0.99 correlation minimum, and 1,000 maximum interactions. There was now an outgroup determined by the IQ-TREE server, in addition to providing bootstrap metrics and Bayesian support. The phylogenetic tree was visualized and edited using FigTree (FigTree, v.1.4.4, http://tree.bio.ed.ac.uk/software/figtree/).

β-defensin structural prediction

Protein sequences classified as defensins were processed through AlphaFold2[50–52] to predict their three-dimensional structures using the highest-rank classification. The predicted structures were validated by the PROCHECK server (PROCHECK, v.3.5.4, https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/)[53–55] and analyzed according to the bond and angle geometry, secondary structure, and protein packaging. The defensin sequences were also submitted to the Expasy Server to analyze their physicochemical features by the ProtParam tool (https://web.expasy.org/protparam/).[56] The predicted and validated defensins were analyzed and modeled for three-dimensional visualization using PyMOL (PyMOL, v.3.0, https://pymol.org). The cysteines, disulfide bonds, and secondary structures (such as α-helices and β-sheets) were highlighted.

MD analyses

The GROMACS program (GROMACS, v.2023.1, https://www.gromacs.org/index.html)[57–59] was used to perform MD analyses of the six β-defensins in a simulation box containing a saline medium (ionic strength: 0.15 M NaCl). This analysis focused on determining electrostatic interactions, the stabilization of charges, physiological conditions, and the structure and dynamics analysis, lasting 100 ns. The dynamic data were interpreted using RMSD (root-mean-square deviation), RMSF, and RG graphs, enabling the analysis and visualization of the molecular stability and the dynamics of the structures during the simulations. Snapshots of the molecules were also taken during the interaction at 0, 20, 40, 60, 80, and 100 ns, allowing the visualization of changes in the structural conformation of each molecule.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Octavio L. Franco (ocfranco@gmail.com).

Materials availability

Materials generated in this study have been deposited to Zenodo: https://doi.org/10.5281/zenodo.11042500.

Data and code availability

The data and code generated during this study are available at Zenodo: https://doi.org/10.5281/zenodo.11042500.

ACKNOWLEDGMENTS

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF), Fundação de Apoio ao Desenvolvimento do Ensino, and Ciência e Tecnologia do Estado de Mato Grosso do Sul (Fundect). C.d.l.F.-N. holds a Presidential Professorship at the University of Pennsylvania and acknowledges funding from the Procter & Gamble Company, United Therapeutics, a BBRF Young Investigator Grant, the Nemirovsky Prize, Penn Health-Tech Accelerator Award, Defense Threat Reduction Agency grants HDTRA11810041 and HDTRA1-23-1-0001, and the Dean's Innovation Fund from the Perelman School of Medicine at the University of Pennsylvania. Research reported in this publication was supported by the Langer Prize (AIChE Foundation), the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM138201, and DTRA HDTRA1-21-1-0014. All figures were prepared in BioRender or Corel Draw, and some organisms were acquired and edited in Paint 3D from Microsoft Corporation, whose digital products from MC are licensed for use. Other organisms were extracted from the free encyclopedia Wikipedia under license CC BY 4.0 DEED, attribution 4.0 International: https://creativecommons.org/licenses/by/4.0/. Organisms were created from modifications and editions made using Inkscape 1.2.

AUTHOR CONTRIBUTIONS

Conceptualization, A.F.L.F. and H.M.D.; methodology, A.F.L.F., K.O.O., and K.B.S.d.O.; software, A.F.L.F.; writing – original draft, A.F.L.F., K.O.O., K.B.S.d.O., M.H.C., L.R.d.L., M.L.R.M., C.L., and C.d.l.F.-N.; writing – review & editing, A.F.L.F., K.O.O., K.B.S.d.O., M.H.C., L.R.d.L., C.L., C.d.l.F.-N., and O.L.F.; formal analysis, K.O.O. and K.B.S.d.O.; investigation, K.O.O., K.B.S.d.O., and L.R.d.L.; visualization, K.O.O., K.B.S.d.O., and H.M.D.; supervision, M.H.C., C.L., C.d.l.F.-N., and O.L.F.; resources, O.L.F.; funding acquisition, O.L.F.; project administration, O.L.F.

DECLARATION OF INTERESTS

C.d.l.F.-N. is a member of the advisory board for *Cell Reports Physical Science*.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2024.102193.

Received: June 4, 2024
Revised: August 1, 2024
Accepted: August 12, 2024
Published: September 11, 2024

REFERENCES

1. Maasch, J.R.M.A., Torres, M.D.T., Melo, M.C.R., and de la Fuente-Nunez, C. (2023). Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning. Cell Host Microbe *31*, 1260–1274.e6. https://doi.org/10.1016/j.chom.2023.07.001.

2. Wan, F., Torres, M.D.T., Peng, J., and de la Fuente-Nunez, C. (2023). Molecular de-extinction of antibiotics enabled by deep learning. Preprint at bioRxiv. https://doi.org/10.1101/2023.10.01.560353.

3. Wan, F., Wong, F., Collins, J.J., and de la Fuente-Nunez, C. (2024). Machine learning for antimicrobial peptide identification and design. Nat. Rev. Bioeng. *2*, 392–407. https://doi.org/10.1038/s44222-024-00152-x.

4. Wan, F., Torres, M.D.T., Peng, J., and de la Fuente-Nunez, C. (2024). Deep-learning-enabled antibiotic discovery through molecular de-extinction. Nat. Biomed. Eng. *8*, 854–871. https://doi.org/10.1038/s41551-024-01201-x.

5. Shafee, T.M.A., Lay, F.T., Phan, T.K., Anderson, M.A., and Hulett, M.D. (2017). Convergent evolution of sequence, structure and function. Cell. Mol. Life Sci. *74*, 663–682. https://doi.org/10.1007/s00018-016-2344-5.

6. Gao, X., Ding, J., Liao, C., Xu, J., Liu, X., and Lu, W. (2021). Defensins: The natural

peptide antibiotic. Adv. Drug Deliv. Rev. *179*, 114008. https://doi.org/10.1016/j.addr.2021.114008.

7. Islam, S., Akhand, M.R.N., and Hasan, M. (2023). Evolutionary trend of bovine β-defensin proteins toward functionality prediction: A domain-based bioinformatics study. Heliyon *9*, e14158. https://doi.org/10.1016/j.heliyon.2023.e14158.

8. Contreras, G., Shirdel, I., Braun, M.S., and Wink, M. (2020). Defensins: Transcriptional regulation and function beyond antimicrobial activity. Dev. Comp. Immunol. *104*, 103556. https://doi.org/10.1016/j.dci.2019.103556.

9. Ganz, T. (2003). Defensins: Antimicrobial peptides of innate immunity. Nat. Rev. Immunol. *3*, 710–720. https://doi.org/10.1038/nri1180.

10. Guyot, N., Meudal, H., Trapp, S., Iochmann, S., Silvestre, A., Jousset, G., Labas, V., Reverdiau, P., Loth, K., Hervé, V., et al. (2020). Structure, function, and evolution of Gga -AvBD11, the archetype of the structural avian-double-β-defensin family. Proc. Natl. Acad. Sci. USA *117*, 337–345. https://doi.org/10.1073/pnas.1912941117.

11. Lehrer, R.I., and Ganz, T. (2002). Defensins of vertebrate animals. Curr. Opin. Immunol. *14*, 96–102. https://doi.org/10.1016/S0952-7915(01)00303-X.

12. Zhu, S., and Gao, B. (2013). Evolutionary origin of β-defensins. Dev. Comp. Immunol. *39*, 79–84. https://doi.org/10.1016/j.dci.2012.02.011.

13. Lynn, D.J., Higgs, R., Lloyd, A.T., O'Farrelly, C., Hervé-Grépinet, V., Nys, Y., Brinkman, F.S.L., Yu, P.-L., Soulier, A., Kaiser, P., et al. (2007). Avian beta-defensin nomenclature: A community proposed update. Immunol. Lett. *110*, 86–89. https://doi.org/10.1016/j.imlet.2007.03.007.

14. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., and Ferrin, T.E. (2023). UCSF ChimeraX: Tools for structure building and analysis. Protein Sci. *32*, e4792. https://doi.org/10.1002/pro.4792.

15. Wong, F., de la Fuente-Nunez, C., and Collins, J.J. (2023). Leveraging artificial intelligence in the fight against infectious diseases. Science *381*, 164–170. https://doi.org/10.1126/science.adh1114.

16. Bunce, M., Worthy, T.H., Phillips, M.J., Holdaway, R.N., Willerslev, E., Haile, J., Shapiro, B., Scofield, R.P., Drummond, A., Kamp, P.J.J., and Cooper, A. (2009). The evolutionary history of the extinct ratite moa and New Zealand Neogene paleogeography. Proc. Natl. Acad. Sci. USA *106*, 20646–20651. https://doi.org/10.1073/pnas.0906660106.

17. van Dijk, A., Guabiraba, R., Bailleul, G., Schouler, C., Haagsman, H.P., and Lalmanach, A.C. (2023). Evolutionary diversification of defensins and cathelicidins in birds and primates. Mol. Immunol. *157*, 53–69. https://doi.org/10.1016/j.molimm.2023.03.011.

18. Tu, J., Li, D., Li, Q., Zhang, L., Zhu, Q., Gaur, U., Fan, X., Xu, H., Yao, Y., Zhao, X., and Yang, M. (2015). Molecular evolutionary analysis of β-defensin peptides in vertebrates. Evol. Bioinform. Online *11*, 105–114. https://doi.org/10.4137/EBO.S25580.

19. Wong, J.H., Xia, L., and Ng, T.B. (2007). A review of defensins of diverse origins. Curr. Protein Pept. Sci. *8*, 446–459. https://doi.org/10.2174/138920307782411446.

20. Cuperus, T., Coorens, M., van Dijk, A., and Haagsman, H.P. (2013). Avian host defense peptides. Dev. Comp. Immunol. *41*, 352–369. https://doi.org/10.1016/j.dci.2013.04.019.

21. Hellgren, O., and Ekblom, R. (2010). Evolution of a cluster of innate immune genes (β-defensins) along the ancestral lines of chicken and zebra finch. Immunome Res. *6*, 3. https://doi.org/10.1186/1745-7580-6-3.

22. Cheng, Y., Prickett, M.D., Gutowska, W., Kuo, R., Belov, K., and Burt, D.W. (2015). Evolution of the avian β-defensin and cathelicidin genes. BMC Evol. Biol. *15*, 188. https://doi.org/10.1186/s12862-015-0465-3.

23. Machado, L.R., and Ottolini, B. (2015). An evolutionary history of defensins: A role for copy number variation in maximizing host innate and adaptive immune responses. Front. Immunol. *6*, 115–119. https://doi.org/10.3389/fimmu.2015.00115.

24. Ma, D., Lin, L., Zhang, K., Han, Z., Shao, Y., Liu, X., and Liu, S. (2011). Three novel *Anas platyrhynchos* avian β-defensins, upregulated by duck hepatitis virus, with antibacterial and antiviral activities. Mol. Immunol. *49*, 84–96. https://doi.org/10.1016/j.molimm.2011.07.019.

25. Soman, S.S., Arathy, D.S., and Sreekumar, E. (2009). Discovery of *Anas platyrhynchos* avian β-defensin 2 (Apl_AvBD2) with antibacterial and chemotactic functions. Mol. Immunol. *46*, 2029–2038. https://doi.org/10.1016/j.molimm.2009.03.003.

26. Chapman, J.R., Hellgren, O., Helin, A.S., Kraus, R.H.S., Cromie, R.L., and Waldenström, J. (2016). The evolution of innate immune genes: Purifying and balancing selection on β-defensins in waterfowl. Mol. Biol. Evol. *33*, 3075–3087. https://doi.org/10.1093/molbev/msw167.

27. Guyot, N., Landon, C., and Monget, P. (2022). The two domains of the avian double-β-defensin AvBD11 have different ancestors, common with potential monodomain crocodile and turtle defensins. Biology *11*, 690. https://doi.org/10.3390/biology11050690.

28. Lan, H., Chen, H., Chen, L.-C., Wang, B.-B., Sun, L., Ma, M.-Y., Fang, S.-G., and Wan, Q.-H. (2014). The first report of a Pelecaniformes defensin cluster: Characterization of β-defensin genes in the crested ibis based on BAC libraries. Sci. Rep. *4*, 6923. https://doi.org/10.1038/srep06923.

29. Xiao, Y., Hughes, A.L., Ando, J., Matsuda, Y., Cheng, J.-F., Skinner-Noble, D., and Zhang, G. (2004). A genome-wide screen identifies a single β-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins. BMC Genom. *5*, 56. https://doi.org/10.1186/1471-2164-5-56.

30. Allentoft, M.E., and Rawlence, N.J. (2012). Moa's Ark or volant ghosts of Gondwana? Insights from nineteen years of ancient DNA research on the extinct moa (Aves: Dinornithiformes) of New Zealand. Annals of Anatomy - Anatomischer Anzeiger *194*, 36–51. https://doi.org/10.1016/j.aanat.2011.04.002.

31. Cooper, A. (1997). Studies of avian ancient DNA: From Jurassic Park to modern island extinctions. In Avian Molecular Evolution and Systematics (Elsevier), pp. 345–373. https://doi.org/10.1016/B978-012498315-1/50019-4.

32. Saleh, M.S., Khalil, M.H., Iraqi, M.M., and Camarda, A. (2021). Molecular associations of gallinacin genes with immune response against *Salmonella typhimurium* in chickens. Livest. Sci. *244*, 104315. https://doi.org/10.1016/j.livsci.2020.104315.

33. Mahnam, K., Foruzandeh, S., Mirakhorli, N., and Saffar, B. (2018). Experimental and theoretical studies of cadmium ions absorption by a new reduced recombinant defensin. J. Biomol. Struct. Dyn. *36*, 2004–2014. https://doi.org/10.1080/07391102.2017.1340851.

34. Torres, A.M., and Kuchel, P.W. (2004). The β-defensin-fold family of polypeptides. Toxicon *44*, 581–588. https://doi.org/10.1016/j.toxicon.2004.07.011.

35. Leonardi, M., Librado, P., Der Sarkissian, C., Schubert, M., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Gamba, C., Willerslev, E., and Orlando, L. (2017). Evolutionary patterns and processes: Lessons from ancient DNA. Syst. Biol. *66*, e1–e29. https://doi.org/10.1093/sysbio/syw059.

36. Hughes, A.L., and Yeager, M. (1997). Coordinated amino acid changes in the evolution of mammalian defensins. J. Mol. Evol. *44*, 675–682. https://doi.org/10.1007/PL00006191.

37. Lay, F.T., Poon, S., McKenna, J.A., Connelly, A.A., Barbeta, B.L., McGinness, B.S., Fox, J.L., Daly, N.L., Craik, D.J., Heath, R.L., and Anderson, M.A. (2014). The C-terminal propeptide of a plant defensin confers cytoprotective and subcellular targeting functions. BMC Plant Biol. *14*, 41. https://doi.org/10.1186/1471-2229-14-41.

38. Zhu, S., Bosmans, F., and Tytgat, J. (2004). Adaptive evolution of scorpion sodium channel toxins. J. Mol. Evol. *58*, 145–153. https://doi.org/10.1007/s00239-003-2534-2.

39. Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Sherry, S.T., Yankie, L., and Karsch-Mizrachi, I. (2024). GenBank 2024 Update. Nucleic Acids Res. *52*, D134–D137. https://doi.org/10.1093/nar/gkad903.

40. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. Database *2020*, baaa062. https://doi.org/10.1093/database/baaa062.

41. Nachtweide, S., and Stanke, M. (2019). Multi-Genome Annotation with AUGUSTUS. In Gene Prediction (New York, NY: Humana), pp. 139–160. https://doi.org/10.1007/978-1-4939-9173-0_8.

42. Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. *32*, W309–W312. https://doi.org/10.1093/nar/gkh379.

43. Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res. *46*, W200–W204. https://doi.org/10.1093/nar/gky448.

44. Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinf. *11*, 431. https://doi.org/10.1186/1471-2105-11-431.

45. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427. https://doi.org/10.1093/nar/gkac993.

46. Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217. https://doi.org/10.1006/jmbi.2000.4042.

47. Chang, J.-M., Di Tommaso, P., and Notredame, C. (2014). TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol. Biol. Evol. *31*, 1625–1637. https://doi.org/10.1093/molbev/msu117.

48. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

49. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., and Minh, B.Q. (2016). W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. *44*, W232–W235. https://doi.org/10.1093/nar/gkw256.

50. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

51. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.

52. Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., et al. (2024). AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. *52*, D368–D375. https://doi.org/10.1093/nar/gkad1011.

53. Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. *26*, 283–291. https://doi.org/10.1107/S0021889892009944.

54. Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR *8*, 477–486. https://doi.org/10.1007/BF00228148.

55. Laskowski, R.A., MacArthur, M.W., and Thornton, J.M. (2012). PROCHECK : Validation of protein-structure coordinates. In International Tables for Crystallography, Online MRW (Wiley), pp. 684–687. https://doi.org/10.1107/97809553602060000882.

56. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein identification and analysis tools on the expasy server. In The Proteomics Protocols Handbook, J.M. Walker, ed. (Humana Press), p. 988. https://doi.org/10.1385/1592598900.

57. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J.C. (2005). GROMACS: Fast, flexible, and free. J. Comput. Chem. *26*, 1701–1718. https://doi.org/10.1002/jcc.20291.

58. Berendsen, H.J.C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. Comput. Phys. Commun. *91*, 43–56. https://doi.org/10.1016/0010-4655(95)00042-E.

59. Páll, S., Zhmurov, A., Bauer, P., Abraham, M., Lundborg, M., Gray, A., Hess, B., and Lindahl, E. (2020). Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. J. Chem. Phys. *153*, 134111. https://doi.org/10.1063/5.0018516.