Received: 29 April 2022 | Revised: 21 September 2022 | Accepted: 8 October 2022

DOI: 10.1002/pro.4469

FULL-LENGTH PAPER

THE PROTEIN SOCIETY WILEY

Uncovering the promiscuous activity of IL-6 proteins: A multi-dimensional analysis of phylogeny, classification and residue conservation

André da Costa[1,2] | **Ricardo Franco-Duarte**[1,2] | **Raul Machado**[1,2] | **Andreia C. Gomes**[1,2]

[1]CBMA - Centre of Molecular and Environmental Biology, Department of Biology, University of Minho, Braga, Portugal

[2]IB-S Institute of Science and Innovation for Sustainability, University of Minho, Braga, Portugal

Correspondence
André da Costa and Andreia C. Gomes, CBMA - Centre of Molecular and Environmental Biology, Department of Biology, University of Minho, Campus of Gualtar, 4710-057 Braga, Portugal.
Email: andrecosta@bio.uminho.pt (A.D.C.) and agomes@bio.uminho.pt (A.C.G.)

Funding information
Fundação para a Ciência e a Tecnologia, Grant/Award Numbers: CEECIND/00526/2018, POCI-01-0145-FEDER-030568, UIDB/04050/2020

Review Editor: Nir Ben-Tal

Abstract

The IL-6 family of cytokines, known for their pleiotropic behavior, share binding to the gp130 receptor for signal transduction with the necessity to bind other receptors. Leukemia inhibitory factor receptor is triggered by the IL-6 family proteins: leukemia inhibitory factor (LIF), oncostatin-M (OSM), cardiotrophin-1 (CT-1), ciliary neurotrophic factor (CNTF), and cardiotrophin-like cytokine factor 1 (CLCF1). Besides the conserved binding sites to the receptor, not much is known in terms of the diversity and characteristics of these proteins in different organisms. Herein, we describe the sequence analysis of LIF, OSM, and CT-1 from several organisms, and m17, a LIF ortholog found in fishes, regarding its phylogenetics, intrinsic properties, and the impact of conserved residues on structural features. Sequences were identified in seven classes of vertebrates, showing high conservation values in binding site III, but protein-dependent results on binding site II. GRAVY, isoelectric point, and molecular weight parameters were relevant to differentiate classes in each protein and to enable, for the first time and with high fidelity, the prediction of both organism class and protein type just using machine learning approaches. OSM sequences from primates showed an increased BC loop when compared to the remaining mammals, which could influence binding to OSM receptor and tune signaling pathways. Overall, this study highlights the potential of sequence diversity analysis to understand IL-6 cytokine family evolution, showing the conservation of function-related motifs and evolution of class and protein-dependent characteristics. Our results could impact future medical treatment of disorders associated with imbalances in these cytokines.

KEYWORDS

IL-6 cytokine family, leukemia inhibitory factor, machine learning, protein evolution, sequence diversity analysis

1 | INTRODUCTION

The IL-6 family of cytokines is known for its pleiotropic behavior, playing a wide range of context-dependent biological roles and inducing cell type-specific effects. Since the molecular cloning of IL-6 cytokine in 1986,[1] several other members were identified and included in this family. Nowadays, the family comprises 10 members: IL-6, IL-11, IL-27, IL-35, IL-39, oncostatin-M (OSM), leukemia inhibitory factor (LIF), ciliary neurotrophic factor (CTNF), cardiotrophin-1 (CT-1), and cardiotrophin-like cytokine factor 1 (CLCF1).[2] All members of the IL-6 family of cytokines share common features: a signaling receptor subunit glycoprotein with 130 kDa (gp130) for signal transduction[3]; phosphorylation of Janus kinase (JAK) and the signal transducer and activator of transcription (STAT) upon binding to gp130; and their respective receptor as part of a mechanism of different signaling pathways.[2] IL-6 family members are often referred to as pleiotropic due to several different activities that are both systemic and cell-specific, which include effects on egg implantation control during pregnancy, hematopoietic stem cell regulation, glucose homeostasis, wound healing and implications in cancer, arthritis, perinatal development or obesity.[2,4,5] LIF was named after its ability to induce differentiation of myeloid leukemia cells. However, it is involved in several other roles.[6] OSM is a protein similar to LIF and was first introduced as an anticancer agent but with roles in immunity and inflammation.[7] Cardiotrophin-1 was initially described for its effects on the heart, and is now known that CT-1 also acts on other organs such as the liver or the kidneys.[8] The role of some IL-6 cytokine family members in different diseases is controversial, with both beneficial and detrimental effects. For instance, OSM has been shown to have cytostatic effect on lung cancer cells, osteosarcomas and chondrosarcomas but also to induce proliferation of Ewing sarcomas.[9,10] The role of these cytokines is known to be conserved among different organisms with several LIF orthologs studied in other primates, dogs, ungulates, felines, birds, and frogs showing similar functions in stem cell regulation and fetus development as well as tumorigenic effects.[11–17] The conservation and pleiotropic functions of LIF, OSM, and CT-1 made researchers hypothesize that a common ancestor protein arose after more than one gene duplication event.[18] The finding of m17 (termed after the name of the clone m17), a LIF-like cytokine in fishes,[19–21] suggests that the ancestry of IL-6 family members seems to precede the fish-to-tetrapod transition. The first attempts of a phylogenetic study of LIF, OSM, or CT-1 proteins included a small number of protein samples, resulting in odd interactions between orthologs of humans and other animals, and with variations between studies. Adrian-Segarra et al.[18] placed hOSM as being closely related to murine and rat orthologs, whereas Hwang et al.[21] placed it near the cow ortholog. With the most recent available genomes and proteomes, this study can now be conducted in more detail.

Similarly to other cytokines, IL-6 family members are described to have a four helical structure with an up-up-down-down topology that is important for binding to gp130 and a specific or shared receptor required to initiate the signaling process.[3] One such receptor is LIF receptor (LIFR)—a 190 KDa protein with three distinct regions: a cytoplasmatic domain, a small transmembrane domain, and a large extracellular domain.[22] This receptor is common for LIF, OSM, CT-1, CLCF1, and CNTF, whereas the latter two need a binding mediator that can be either CTNFR (CTNF and CLCF1) or IL-6R (CTNF).[2,23] However, while OSM is able to bind LIFR, LIF protein cannot bind OSMR. Molecular simulations and docking-docking studies could aid to scrutinize such behavior.[24] Although common to several proteins, LIFR could be either species-specific or able to bind proteins from different species. Studies using LIF, OSM, and CT-1 focused on the cross-reactivity of both human and mouse variants of these cytokines. In the case of LIF and OSM, hLIF (human LIF) and hOSM (human OSM) are able to bind to murine receptors, but mLIF (murine LIF) and mOSM (murine OSM) are unable to bind hLIFR.[18,25] In CT-1, human CT-1 and murine CT-1 bind either to hLIFR or mLIFR.[26] Cross reactivity is found to be related to conserved motifs such as the FXXK, essential for LIFR binding, found in murine, rat, and human orthologs of CT-1, LIF, and OSM.[18,27–29] A recent study has disclosed other conserved residues important for LIF and OSM binding to LIFR using protein–protein docking, as well as specific residues in each protein.[30] We contemplated whether such motifs exist in other regions of these proteins. Studies on the cross-reactivity of rat and cat orthologs were also performed[31,32] but, to the best of our knowledge, there are no reports on the cross-reactivity from another species to the human receptors.

Focusing on the proteins that activate their signaling pathways through LIFR and gp130 without additional mediators—LIF, OSM, m17, and CT-1—the objective of this work was to unravel the correlation between diversity, conservation, phylogeny, and protein properties solely based on the study of amino acidic sequences. It is, to our knowledge, the first time that a class and a protein type classification is attempted both by phylogenetic studies and by proteins properties extracted from sequence analysis. Focused study on conserved regions, such as the FXXK motif, and not so conserved ones, such as the binding of gp130, were also performed.

2 | MATERIAL AND METHODS

2.1 | Protein sequences

A primary protein sequence dataset was created with sequences retrieved between September 2019 and February 2021 from three different sources: UNIPROT,[33] Ensembl,[34] and NCBI.[35] When available, only the mature protein sequence was considered. LIF sequences were retrieved by initial searches in UNIPROT using the search term "Leukemia inhibitory factor", followed by BLASTp with the LIF human sequence (P15018) against UNIPROTKB reference proteomes plus SWISSPROT as target database. Only sequences with an E-value below $1e^{-15}$ were considered. BLASTp was performed for protein sequences search in Ensembl protein database. Only hits with an E-value below $1e^{-20}$ were considered. For OSM and CT-1, sequences were retrieved after BLASTp against both UNIPROT and Ensembl databases using human OSM sequence (P13725) or human CT-1 sequence (Q16619). Only sequences with an E-value below $1e^{-20}$ were considered. M17 protein sequences were retrieved by employing the same methodology described for OSM and CT-1 while using the sequence from *Danio rerio* (A0PAS6) as reference. Additional analysis was conducted using a reference model for each class (i.e., mammal, reptilian, and bird), randomly selected from the primary protein sequence database, for a BLASTp run (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) while limiting the E-value to $1e^{-10}$.

2.2 | Multiple sequence alignment

Multiple sequence alignment of the retrieved protein sequences was performed using Multiple Sequence Comparison by Log- Expectation (MUSCLE) using the drive5 standalone software.[36] Percentage identity matrices (PIM) were generated using UGene software v33[37] followed by formatting with Microsoft Excel®. Alignment files were trimmed using TrimAl v1.3 with gappyout method.[38] Visualization of the multiple sequence alignments was achieved using Unipro UGENE software v33.[37]

2.3 | Cladogram generation

Cladograms were generated by the Maximum Likelihood method using IQ-tree algorithms[39,40] in protein sequence type and with an automated model finder.[41] The cladograms were bootstrapped 1,000 times[44,45,46] with the Ultrafast Bootstrap algorithm UFBoot[42] and single branch test with 1,000 replicates of SH-aLRT branch test. Generated consensus trees were analyzed and formatted in iTOL,[43] with middle point root and with deletion of nodes with a bootstrap value below 70%. Bayesian inference was performed using Bayesian Evolutionary Analysis Sampling Trees (Beast) v1.10.4.[44] The number of inferences were automatically assigned by the software when creating the Markov chain Monte Carlo chains and ranged from 20,000,000 to 120,000,000 until convergence with the associated burn-in, as analyzed in Tracer v1.7.1.[45] Tree density and substitution models were automatically chosen by the software.

2.4 | Protein parameter analysis and multivariate projection

A total of 1,092 sequences were analyzed with the platform IPC—isoeletric point calculator[46] for calculation of their molecular weight, isoelectric point (Ip), and charge at pH 7.4 (at absolute values). Calculation of the grand average of hydropathy (GRAVY) value was performed in the platform GRAVY CALCULATOR.[47] Matrices were generated for each protein, combining information obtained for all sequences. Data visualization was performed with Orange data mining suite (v. 3.24.1),[48] using scatterplot and FreeViz[49] tools. In particular, FreeViz algorithm was employed to plot the protein molecular weight (KDa, MW), average pI, charge at pH 7.4, and GRAVY, into a two-dimensional visualization, further dividing the sequences by protein class and protein family. The size of the unit vectors projected are then related with the information power of that vector to the classification task, uncovering interactions, and providing information about inter-class similarities.

2.5 | Data analysis using machine-learning algorithms

A set of standard predictive data-mining methods, such as logistic regression, naive Bayesian classifier, CN2 rule

inducer, stochastic gradient descent, adaboost, classification trees, random forest, neural networks, k-nearest-neighbors and support vector machine algorithms, were used for the inference of prediction models in 806 sequences (only complete sequences were considered), using a training set partition of the data. For prediction scoring, the area under the receiver operating characteristics (ROC) curve (AUC),[50] classification accuracy (CA), F1 score, precision, and recall, were used, using 10-fold cross-validation of the test set partition. All classifiers were also tested and scored in Python, using the SciKit-Learn package[51] for consistency. The algorithm with the highest scores, considering all the employed evaluation parameters, was the same using both Orange and Python, and was, in this way, used in further analysis to estimate the probability that the predictive model would correctly differentiate between distinct protein families or between distinct organisms' classes.

2.6 | Overall residues conservation

Untrimmed sequence alignment files were scanned for residue conservation scores calculation using the platform Consurf,[52–54] employing the 3D model structure of human LIF (PDB reference 1EMR), human OSM (1EVS), CT-1 (SWISS-model Q16619) and m17 (HHpred structure from the Consurf platform using MODELLER). Observation and formatting of the obtained 3D conservation structure were performed using UCSF Chimera v1.14.[55]

3 | RESULTS AND DISCUSSION

The pleiotropic activities of the different members of the IL-6 family of cytokines are well known, being some of these functions partially redundant in different members of the same organism. This scenario is the result of sharing gp130 as binding receptor/binding modulator and activating the same initial steps of a complex network of signaling cascades.[2] In fact, several members share not only gp130 as a binding modulator, but also other receptors that were initially thought to be cytokine-specific. That is the case for proteins that bind to LIFR. The proteins LIF, OSM, CT-1, CLCF1, and CNTF have shown the ability to activate signaling cascades through LIFR.[23,56] From these, LIF, OSM and CT-1 have shown to require only gp130 and LIFR, and no other receptors or modulators. As for CLCF1 and CNTF, signaling cascade activation require the binding to gp130, LIFR, and CNTFR too.[23] We retrieved the protein sequences of LIF, OSM, CT-1, and m17 (the LIF ortholog in fishes) of several organisms, to understand the connection between protein diversity, protein properties, and residue conservation in different classes of organisms. Although their sequences are similar to the other family members, we have excluded CLCF1 and CNTF due to the extra receptor requirement that could result in confusing results in terms of conserved regions. While most phylogenetic studies rely on genetic data, our study focused on the extractable data from amino acidic sequences to study not only phylogeny but also unravel motifs conservation connected to the interaction with LIFR or gp130 and aiming at a possible protein classification system. When different isoforms were available, only the sequence most similar to the model was retrieved.

3.1 | Diversity of proteins that bind to LIFR in different organisms

We obtained a total of 1,092 sequences with good quality divided between the four protein families—LIF, OSM, CT-1, and m17, with 285 annotated as partial. LIF had the most representative sample with 534 sequences available (252 partial), followed by OSM with 239 sequences (26 partial), CT-1 with 189 (7 partial), and m17 with 130 sequences (0 partial). The class Mammalia was the most representative with a total of 490 sequences (44.9% of total sequences), followed by the class Aves with 375 (34.3%). This was, since the diversity of genomes and available sequences in mammals are far superior that for any other vertebrate class. An exception lies in LIF protein sequences in which, large projects such as the Bird 10,000 Genomes[57] contributed with several sequences from the class Aves (to a total of 319, mostly partial sequences), surpassing the number of sequences obtained from the class Mammalia (176 sequences).

The cladogram obtained by maximum likelihood (Figure 1) shows the different classes found for each of the studied proteins. These sequences were found in almost all classes of the subphylum Vertebrata, with exception for the Agnatha class. While four-helical cytokines capable of activating the JAK/STAT signaling pathway were already described in Drosophila,[58] we found no matches on Invertebrates in all databases searched. Overall, the cladogram shows a clustering of m17 and CT-1, and OSM and LIF. OSM is known to have its gene sequence linked to that of LIF in humans, on chromosome 22, indicating that it could have resulted from duplication of a common ancestral gene,[59] which is in agreement with our analysis as OSM and LIF are grouped together. In fact, a shared synteny analysis between human, mouse, and two species from the genus *Xenopus* show a noteworthy conservation of the LIF locus in these

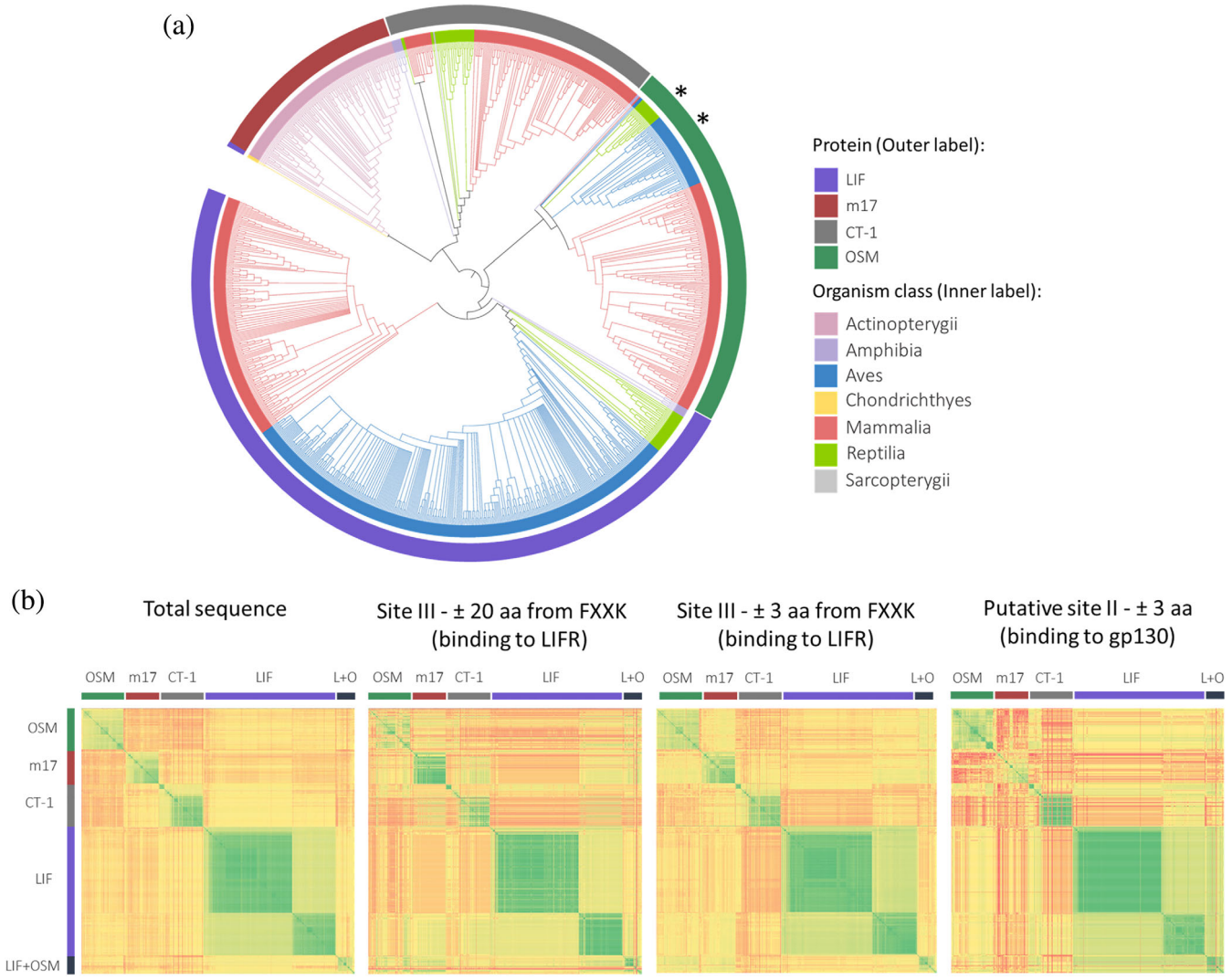

FIGURE 1 Variability and variety of LIF, m17, OSM, and CT-1 proteins. (a) Maximum-likelihood cladogram showing clade distribution of LIF, m17, OSM, and CT-1 proteins in different species. For terminal label information, please see supplementary information for a complete, full resolution cladogram. The cladogram was bootstrapped 1,000 times and nodes with a bootstrap value under 70 were disregarded. (b) Percentage identity matrix (PIM) of the MUSCLE alignment for all sequences and categorized in terms of total sequence; alignment of ±20 amino acids (aa) from FXXK sequence of human LIF; alignment of ±3 aa from FXXK sequence of human LIF; and alignment of ±3 aa from 121 to 127 sequence of human LIF of putative site II. Colour codes from red (lower) to green (higher) denote the similarity between sequences. *–LIF sequences misplaced near OSM sequences

species chromosomes.[15] On the other hand, m17 is described as a LIF ortholog but was grouped with CT-1, probably explained by the fact that these two proteins have more ancient classes in the retrieved sequences. LIF was present in the highest number of classes with representation in Amphibia, Aves, Chondrichthyes, Mammalia, and Reptilia. CT-1 sequences are present in four classes, namely Amphibia, Mammalia, Reptilia, and the classical class Sarcopterygii (now categorized as clade; here referring to the infraclass Coelacanthimorpha); OSM sequences were identified in four classes, Amphibia, Reptilia, Aves, and Mammalia; the LIF ortholog m17 was only found in class Actinopterygii. The results of the likelihood analysis were very similar to those of the Bayesian inference (Figure S1). The main difference lies on the separation of the Mammalia class from the remaining ones in OSM Bayesian inference, which could be explained by the different algorithm used, placing the OSM sequences from birds and reptiles near to those of LIF, by the obtained proximity to these sequences. The similarity obtained for both set of results shows the robustness of the analyses, whereas some differences in estimating the ancestry of proteins by different inferences are documented.[57]

The maximum likelihood calculation (Figure 1) grouped all protein sequences by protein type and by

organism class. The only four exceptions were: LIF sequences from the class Chondrichthyes, placed near m17 proteins; a group of mammalian sequences from CT-1 displaced from the rest of the mammals; LIF and OSM white-throated tinamou (*Tinamus guttatus*) sequences; and, LIF reptiles *Pelodiscus sinensis* and *Gopherus agassizii* and LIF from the bird species Bar-tailed Godwit (*Limosa lapponica baueri*), placed near the OSM sequences from reptiles (class Reptilia) and birds (class Aves), respectively. The first exception can be explained by the orthologous nature of m17, with the presence of a LIF ortholog in Osteichthyes (bony fishes) and sharks being considered to be evolutionarily more related with bony fishes than to mammals. The mammalian sequences displaced from the remaining mammals in CT-1 sequences are characterized by a conserved LIFR binding motif with the residues FERK, which is similar to the motif found in alligators. On the other hand, the motifs from the remaining mammals have the residues FLAK, FSAK, or FPAK. The white-throated tinamou OSM and LIF sequences are the same and were retrieved from both LIF and OSM searches. As for the LIF sequences from the reptiles and the bird Bar-tailed Godwit placed near OSM sequences, this could be a result of an incorrect gene annotation due to its similarity to OSM. In fact, the MUSCLE alignment depicted as a percentage identity matrix (PIM, Figure 1b, total sequence) reveals that some sequences from OSM and LIF are grouped together. These sequences belong to classes Aves (OSM and LIF) and Reptilia (OSM and LIF). It is therefore unclear if these sequences were misannotated and to which family they belong, or if, in fact, there is a clear similarity between them.

Figure 1b depicts percentage identity matrixes (PIM) for different protein motifs using the human LIF sequence as reference, according to: the total sequence; ±20 and ±3 amino acids from the LIFR binding motif FXXK (site III); and ±3 amino acids from part of the site II in human LIF responsible for gp130 binding (residues 121-127).[60] Total sequence analysis gives a global vision of the sequences' similarity, while site III shows how conserved the LIFR binding site is, either in proximity to the conserved FXXK motif (±3 amino acids) or in an extended form (±20 amino acids). Site II analysis shows the similarity of the gp130 binding site. PIM analysis reveals five different clusters of similar sequences: OSM, m17, CT-1, LIF, and OSM + LIF. Considering full sequences, there is a well-defined separation between each cluster (protein). However, the same separation is not as obvious in relation to the studied binding motifs. Both matrixes from the site III motifs (binding to LIFR) show that a sub-cluster of LIF including bird and reptile sequences shows a degree of similarity with m17 and CT-1, suggesting a possible conservation of the site III motif in these species. On the other hand, in the matrix regarding the binding to gp130, LIF sequences show a higher degree of similarity with OSM, implying a similar motif for gp130 binding.[60,61] CT-1 sequences from reptiles have some degree of similarity to LIF, whereas m17 shows clear dissimilarity to the remaining proteins. Once again, LIF sequences from sharks are placed near those of m17.

The sequences from each protein family were aligned and individually analyzed through maximum likelihood analysis to correlate sequence diversity with different groups of animals (Figures 2 and 3). LIF sequences are dispersed through the highest number of organism classes and, subsequently, show the highest number of artificial animal classifications: 31 (Figure 2a). Most of the used classifications correspond to orders or clades with a similar degree. In addition to maximum likelihood analysis, sequences were also analyzed by Bayesian inference. Except for the division of amphibians into two different clades (frogs and caecilians), the Bayesian inference presented a cladogram similar to the maximum likelihood (Figure S2). LIF sequences from sharks demonstrated to be the most distant, followed by the Bar-tailed godwit, Big-headed turtle (*Platysternon megacephalum*), and the Desert tortoise (*G. agassizii*) that were mixed with OSM sequences in the general cladogram (Figure 1, see Supplementary information for detailed cladogram). Their distance is most likely the consequence of misannotation and being, in fact, OSM sequences. Interestingly, IQ-tree representation shows a clear separation between frogs (order Anura) and caecilians (order Gymnophiona) for class Amphibia. This could be the result of the later divergence of frogs and salamanders,[62] and the result of adaptive evolution regarding the extreme environments inhabited by caecilians.[63] In fact, caecilians, mammals, birds, and reptiles share the same ancestor share the same ancestor. Reptiles are mainly divided in three clusters: lizards and snakes (order Squamata), turtles (order Testudines), and crocodiles (order Crocodilia). The tuatara (*Sphenodon punctatus*) appears isolated as the only member of the order Rhynchocephalia. In mammals, a clear division was found between placentals (order Eutheria) and marsupials (Order Metatheria) grouped with monotremes (order Monotremaria). Due to a higher number of sequences available, placentals had almost all their sequences clearly grouped into orders/families. A small exception was the absence of lemures (Lemuriformes) from the primates clade. They were grouped with other primates but near rodents. Bats were a special case in the cladogram. All the sequences from bats clustered together but suborders Yangochiroptera (microbats) and Yinpterochiroptera (megabats) were

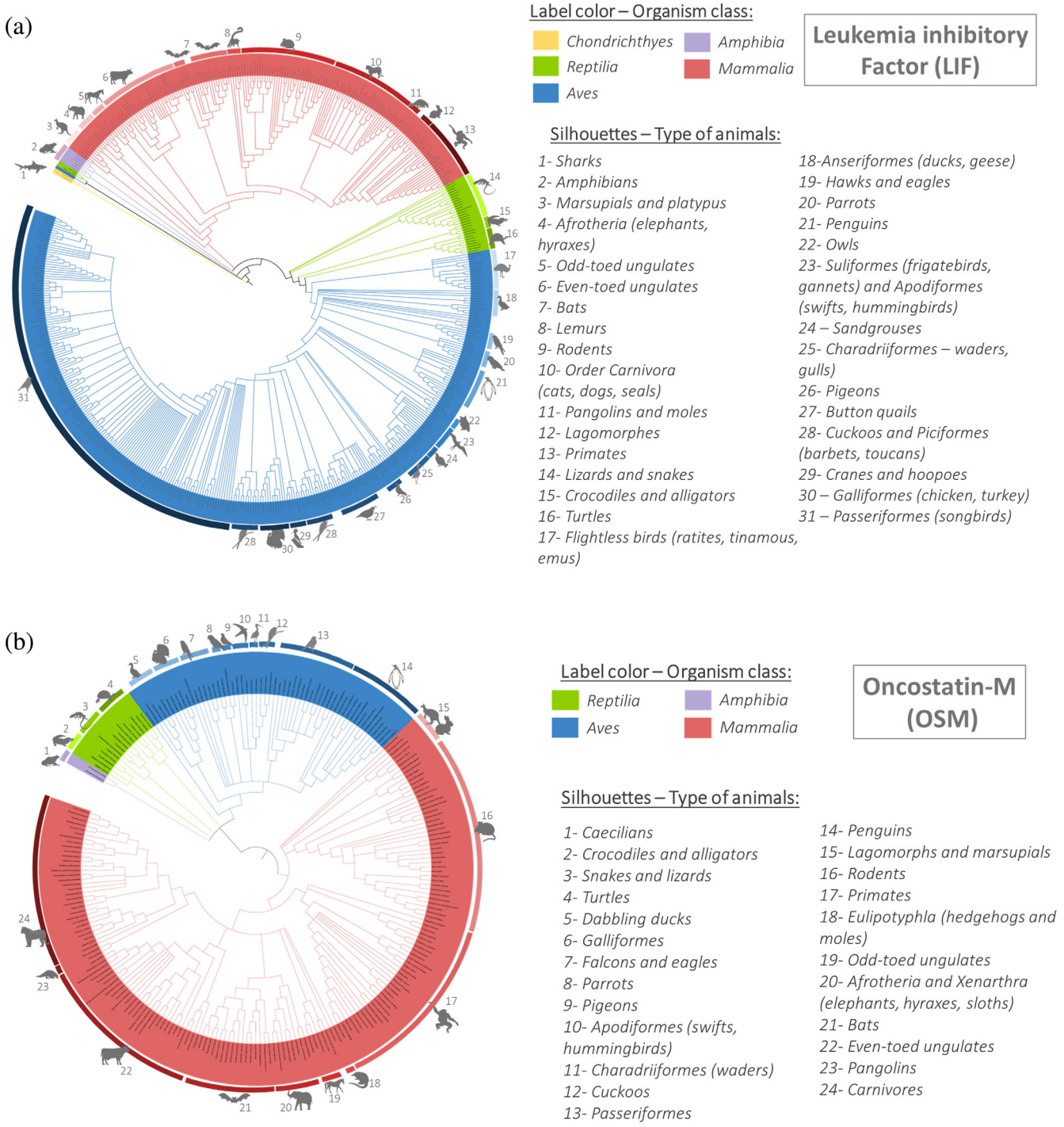

FIGURE 2 Maximum-likelihood cladograms with artificial classification of organisms for Leukemia inhibitory factor (a) and Oncostatin-M (b). Terminal labels show species name (see supplementary information for readable terminal labels). The cladogram was bootstrapped 1,000 times and nodes with a bootstrap value under 70 were deleted. The inner label corresponds to organism class and the outer label to the (order, clade, family) classification of animals

intercalated with the sequences from pig (*Sus scrufa*) and the Chinese tree shrew (*Tupaia chinensis*). It is unknown why it happens, if really similar sequence or a misannotation. Birds, as the class with more sequences, has the highest number of classifiers. While most of the sequences have clustered into different orders, there were some exceptions that are attributed to the higher number of partial sequences in this class. It was the case of crested ibis (*Nipponia nippon*, order Pelecaniformes) that grouped with the secretarybird (*Sagittarius serpentarius*)

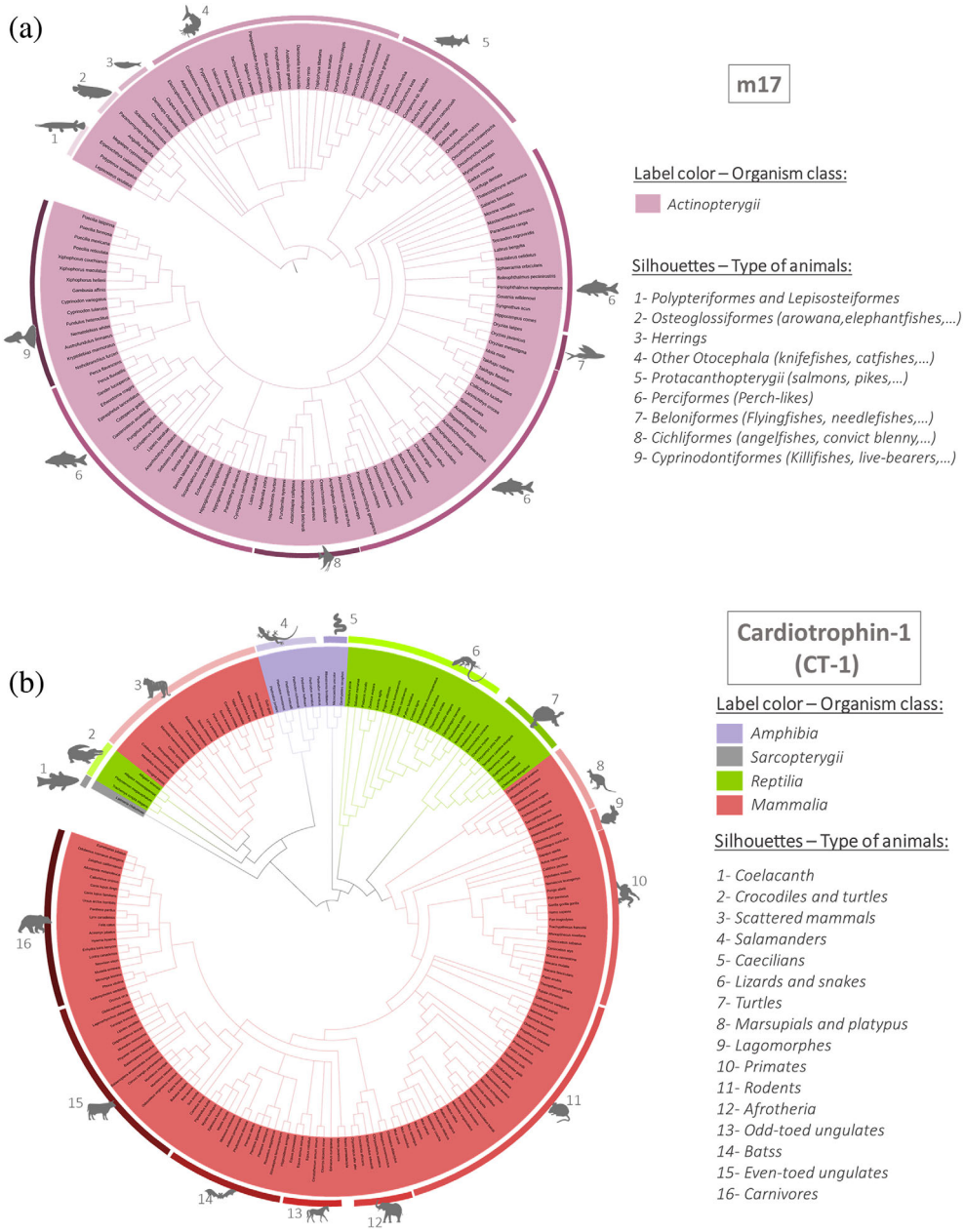

FIGURE 3 Maximum-likelihood cladograms with artificial classification of organisms for m17 (a) and Cardiotrophin-1 (*b*). Terminal labels show species name (see supplementary information for readable terminal labels). The cladogram was bootstrapped 1,000 times and nodes with a bootstrap value under 70 were deleted. The inner label corresponds to organism class and the outer label to the (order, clade, family) classification of animals

and the pied-billed grebe (Podilymbus Podiceps, order Podicipediformes). The presence of sequences from Pelecaniformes, Sulliformes, and Procellariiformes non clustered with others could be related with the polyphyletic nature of the order Pelecaniformes[64] and the unresolved taxonomic relationship between the orders.[65] Ratites and Tinamous birds clustered together, which is in line with literature as they are jointly designated as Palaeognathes[66]; and the appearance of a large cluster of Passeriformes correlates with it being the largest order in the Aves class.[65] In general, LIF protein sequences seem to have evolved with different speciation events, resulting in the clades found. Protein functions should have remained similar, as *M. musculus* and *R. norvegicus* LIFs have similar functions to hLIF,[25] despite being evolutionary distant from each other. Experimentation with orthologs using human LIFR and gp130 could resolve some of these questions.

Analysis of OSM sequences led to similar results as those of LIF, in terms of clade organization (Figure 2b). Again, the Bayesian inference analysis retrieved very similar results (Figure S2). OSM sequences are available for four classes: Amphibia, Aves, Mammalia, and Reptilia; with Mammals clearly separated from the remaining classes. In the Bayesian inference, the ancestry of amphibians and reptilians is placed differently to the maximum-likelihood analysis. OSM sequences in birds have a less prominent Passeriformes clade when

compared to LIF and fewer classifiers due to the lower number of Aves sequences present. In mammals, Lagomorphs and Marsupials are grouped together and sequences from Rodents seem to have a different ancestry compared to the other placentals. Primates, Ungulates, Carnivores, Bats, and Afrotheria appear to share a common ancestor sequence. Pangolins grouped near the order Carnivora, forming the clade Ferae,[67,68] and bats clustered together, although with a clear separation between Yangochiroptera and Yinpterochiroptera. The OSM sequences of lemures also grouped together with those of the remaining primates, in contrast to what happened with LIF.

Sequences for m17 were only retrieved from the class Actinopterygii, since it was described as a LIF ortholog in bony fishes.[69] Four big clades were found for the sequences of this protein (Figure 3a): on containing the orders Polypteriformes, Lepisosteiformes, and Osteoglossiformes, as the group with higher ancestry; the clade with herrings and other Otocephala; the superorder Protacanthopterygii, and the last group with orders Perciformes, Beloniformes, Cichliformes and Cyprinodontiformes, all with the same ancestor organism. As previously noted, the Bayesian inference cladogram showed similar results (Figure S3).

CT-1 sequences (Figure 3b) grouped differently when compared to LIF and OSM. First, CT-1 does not suggest orthologs in Aves but does so in Reptilia, Amphibia and Sarcopterygii, and Mammals. The CT-1 gene is localized in chromosome 16, which could indicate a different evolutionary ancestral when compared to LIF and OSM, explaining the differences in its presence for the different organism classes. Two clear separations, with the same ancestor, are visible: one containing the ancient class Sarcopterygii, Alligators and ancient turtles from Reptilia, salamanders from Amphibia, and a group of mammals from different orders; and the other with caecilians, and the remaining reptiles and mammals. This was confirmed in the Bayesian inference analysis (Figure S3). The scattered mammals found in the first group were already discussed and grouped near crocodiles. Of notice is the fact that it was the only protein family where sequences from salamanders were found. In the remaining placentals, a big clade group ungulates, bats, Afrotheria, and carnivores show similarities between these groups.

3.2 | Protein parameters influence on protein classification

The majority of the sequences were clustered as commonly accepted for the phylogenetic distribution of the studied species.[70,71] However, as expected, there is high variability between sequences of the same protein family. To verify if these differences result in different protein properties and, therefore, different functions, we analyzed all the complete protein sequences (total of 1,092) through a multivariate projection using the FreeViz algorithm.[49] Sequences were projected in respect to the following parameters: Molecular weight (MW, KDa), average Isoelectric point (pI), charge at pH 7.4, and the grand average of hydropathy (GRAVY). In FreeViz, each sample is plotted as a single point on a two-dimensional surface, with the distance between two points being determined by their overall similarity. The background color is displayed as an intensity gradient related to cluster positioning and the size of the axis. Figure 4 shows the FreeViz clustering visualization, underlying protein stratification and allowing to understand several correlations. To note that in some cases we were unable to find the sequence mature form, which could result in discrepancies, especially at the level of MW. LIF protein sequences (Figure 4a) show low separation between organism classes, revealing high conservation between them. Nevertheless, mammal sequences were populated at higher GRAVY values, whereas sequences from class Aves are characterized by a higher pI and positive charge at pH 7.4, and lower GRAVY values. In general, Amphibia showed higher MW, and Reptilia are spread-out with no defined grouping at any characteristic. In OSM (Figure 4b), sequences from reptiles and amphibians show a higher GRAVY and MW, whereas mammals and birds are close together with a tendency for a higher pI in birds sequences. In CT-1 chart (Figure 4c), sequences are scattered throughout the chart with no clear distinction between groups. Contributions were high from all factors tested with an exception for the charge at pH 7.4. As m17 has only sequences in one organism class, this analysis was not performed.

To find patterns between groups, all sequences from all organism classes were considered (Figure 4d), and results show that mammalian sequences populate two different groups: one with higher GRAVY and another with a higher charge at pH 7.4. Sequences from reptiles are in general characterized by a higher GRAVY and MW, although with some exceptions showing low values. As for the sequences from the class Aves, the conclusion is similar to the analysis of the individual protein families: lower GRAVY and higher pI, and protein charge at pH 7.4 for almost all sequences, divided in two groups. Overall, the multivariate projections demonstrate that bird sequences are characterized by a lower MW and higher average pI for both LIF and CT-1 which could be defining characteristics. For the remaining organism classes, defining characteristics are more protein dependent. When analyzing sequences grouped by protein family

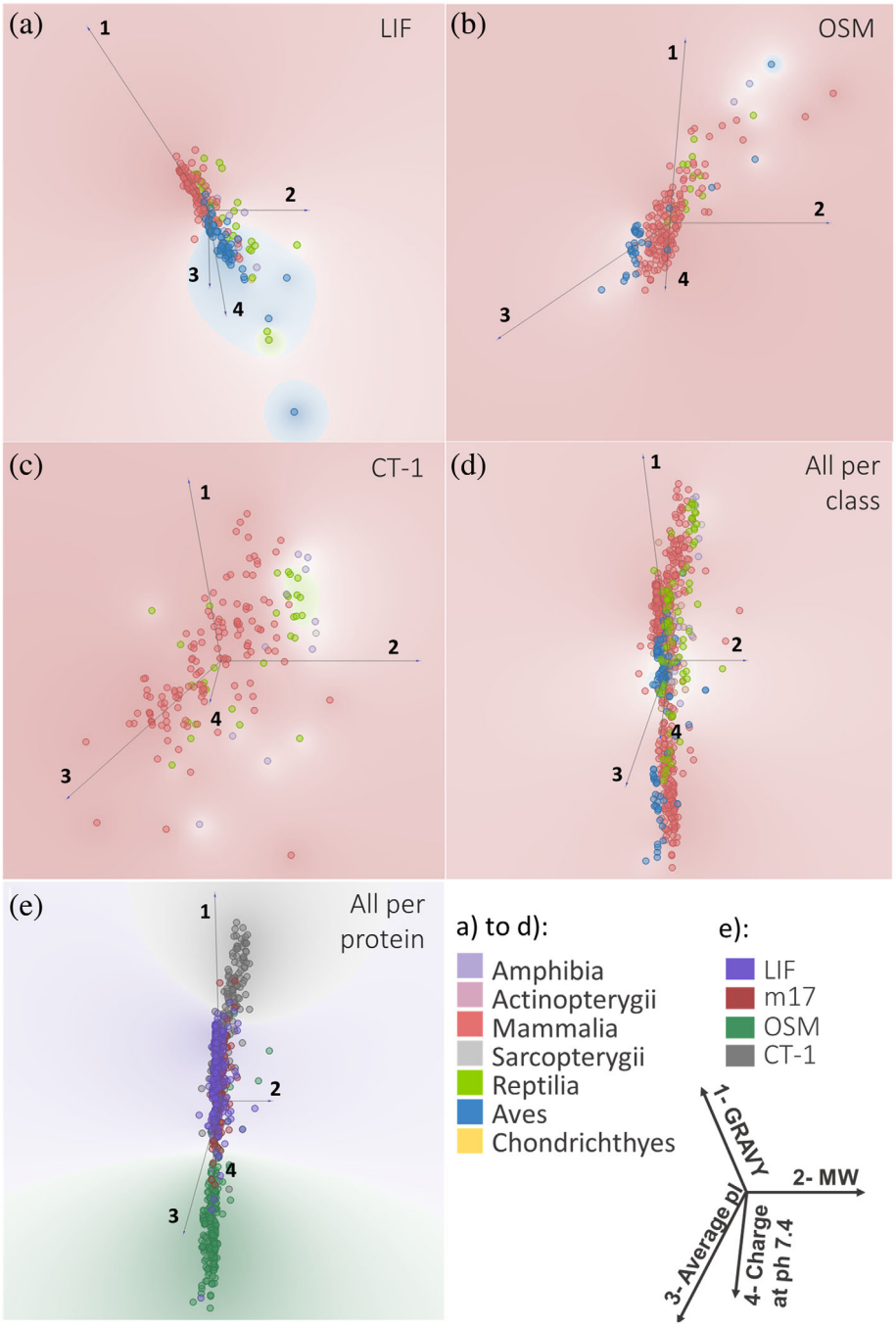

FIGURE 4 Protein parameters and influence on classification. FreeViz plots of multivariate projections using the protein characteristics molecular weight (KDa, MW), average pI, charge at pH 7.4 and grand average of hydropathy (GRAVY) on sequences from LIF (a), OSM (b), CT-1 (c), and all obtained sequences divided by organism class (d) and by protein family (e). The direction and size of each vector indicates the relative prevalence of that feature to explain group stratification. Background color is an intensity gradient related to cluster positioning and size

(Figure 4e), an interesting scenario was found. The OSM sequences are in general separated from the other groups, characterized by higher average pI and charge, with no influence of GRAVY and MW values. On the other hand, the CT-1 sequences demonstrate to be greatly influenced by GRAVY values, spanning over a wide range. As for LIF and m17, these were clustered together as described orthologs with some degree of conservation.

To understand if these characteristics are essential in protein classification, machine-learning data mining algorithms were used for both organism class and protein prediction. For this assessment, only complete sequences (total of 806) were considered. The predictive performance of the different models (plus a Constant one for validation) was evaluated in terms of area under the Receiver-Operating-Characteristics (ROC) curve, using 10-fold cross-validation. The scores obtained for each model in both protein family and organism class classification can be found in Tables S1 and S2. In the case of protein family classification, all the tested models (except Constant and kNN) revealed scores of Area Under the Curve (AUC) above 0.9 (Table S1). As for the organism class classification test, 4 out of 11 algorithms showed AUC scores over 0.9 (Table S2). Random Forest classifier,

a meta estimator that fits a number of decision tree classifiers to establish predictions, retrieved the best results for both classifications approaches, and considering all evaluation parameters (AUC, classification accuracy, F1 score, precision, and recall), was selected for further analyses. In protein classification by protein family, Random Forest classifier obtained an AUC score of 0.990, after 10-fold cross-validation, whereas in classification by organism class, the same model obtained an AUC score of 0.962. Since AUC scores above 0.75 are currently considered as good,[50] we conclude that both values obtained in our classifications are considered as extremely good,[50] reflecting the good accuracy of the predictive models.

In the face of these results, Random Forest classifier was used to predict protein family and organism class classification for the entire dataset and compared with real classifications. Results were presented in the form of confusion matrices (Figures 5a,b) and scatterplots, representing the algorithm correct assignments for the most significant parameters (Figures 5c,d). Regarding protein classification by protein family (Figure 5a), the selected model was able to correctly classify 732 (90.7%) of a total of 806 sequences (misclassified sequences are presented in Table S3). Of notice, none of the m17 proteins were misclassified, showing strong conservation of not only sequences, as previously discussed, but also of their physical–chemical properties. The scatter plot in Figure 5c shows that MW and average pI were the most relevant properties for classification. It is interesting that, even with limited information on the mature form and size for some proteins, the molecular weight demonstrated to exert a strong influence on the classification. Regarding the molecular weight, while all these proteins are characterized by a characteristic four helical structure, differences in the structural organization or extra structural features could aid in creating distinguishable properties that are noticeable by a change in the MW.[27,28,72] Family classification for members of LIF, CT-1, and OSM demonstrated a number of mismatched sequences lower than 10% with, for instance, eight LIF sequences (out of 282) classified as OSM, and 14 OSM sequences (out of 213) classified as LIF.

In the classification by organism class (Figure 5b), 713 (88%) of all 806 sequences were correctly classified (misclassified sequences are presented in Table S4). The algorithm was able to be fairly precise in the classification of a protein's class even when belonging to different protein types, showing high conservation of protein

(a)

Actual \ Predicted	CT-1	LIF	OSM	m17	Σ
CT-1	168	11	3	0	182
LIF	8	266	8	0	282
OSM	1	14	198	0	213
m17	0	0	0	130	130
Σ	177	291	209	130	807

(b)

Actual \ Predicted	Actinopterygii	Amphibia	Aves	Chondrichthyes	Mammalia	Reptilia	Sarcopterygii	Σ
Actinopterygii	130	0	0	0	0	0	0	130
Amphibia	0	4	1	0	6	7	0	18
Aves	0	1	86	0	12	4	0	103
Chondrichthyes	0	0	0	0	2	0	0	2
Mammalia	0	3	12	0	453	12	0	480
Reptilia	0	3	3	0	27	40	0	73
Sarcopterygii	0	0	0	0	1	0	0	1
Σ	130	11	102	0	501	63	0	807

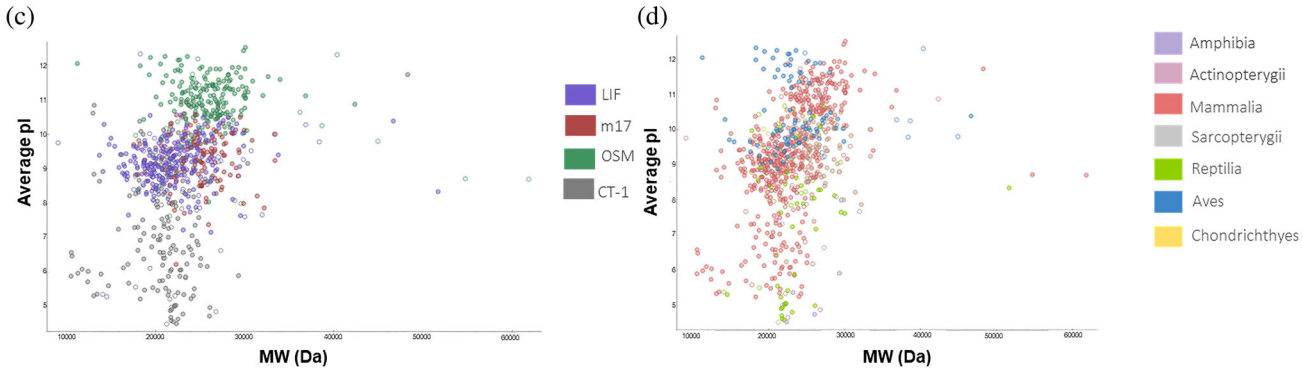

FIGURE 5 Modelling analysis for protein classification. (a) Confusion Matrix indicating the protein family classification of the 806 sequences predicted via Random Forest model (AUC = 0.990); (b) Confusion Matrix indicating the organism class of the 806 protein sequences predicted via Random Forest model (AUC = 0.961); Scatter plots of protein family (c) and organism class (d) classifications relatively to most important characteristics. Filled circles: correct classifications; non-filled circles: incorrect classifications

properties in some organism classes. While for m17, the analysis based solely on organism class could seem obvious, the results obtained for the Mammalia class were fairly precise. We also found that the model's precision increased with more sequences available for analysis, ranging from 94% accuracy in mammals (501 sequences available) to 0% in Sarcopterygii (one sequence available) and Chondrichthyes (two sequences available). The small size of sequences available for some classes impairs the prediction of a strong classification based on the protein properties.[73] Once again, most of the sequences misplaced in the cladogram of Figure 1 were also part of the misclassified list in this classification (Table S4).

3.3 | Conservation of protein structure

The importance of protein structure for function is well established and residue conservation has a role in protein evolution, structure, and function.[74,75] IL-6 family cytokines are described to have a conserved four-helical structure with an up-up-down-down topology that is important for their function and is characteristic of many cytokine families.[3] We assessed residue conservation for all the retrieved sequences using the ConSurf server[52,53] with 3D structures of human LIF and human OSM, and predicted protein structures for human CT-1 and m17 as base models. Figure 6 shows the conservation scores for the tested proteins with a color-base scheme. All protein structures, even the predicted ones, show the characteristic four helices, with OSM (Figure 6b) displaying an extra loop which is important for its specificity toward the OSM receptor.[28]

Binding of LIF, OSM, and CT-1 to the Immunoglobulin-like domain of LIF receptor is mediated via the FXXK motif located at the N terminus of helix D (termed site III, Figure 6).[27] Indeed, both F and K residues demonstrated to be highly conserved (max score) for all of the retrieved sequences. This analysis is in accordance with available literature,[27,29,60] supporting the importance of this motif for binding to the LIF receptor. Table 1 shows the percentage of conservation for each of the Phe or Lys residues. Both residues have values above 90% in the four proteins tested. The only exception was found in the K residue for OSM (84.9% of conservation), showing a substitution by the polar amino acid arginine (Arg, R) with 14.6% of conservation in several mammal species, mainly in order Carnivora. However, this substitution was not found in pangolin sequences (genus Manis), pointing out to a divergence in the order Carnivora, appearing 64 million years ago after the clade Ferae ancestor.[67] Huyton et al. have disclosed the crystal structure of hLIF bound to the murine LIFR, revealing a binding interface comprised of 22 amino acids of hLIF[61] having the FXXK domain lying in the center of the interface. While affinity could change between a human ligand and a murine receptor, several other residues from this interface are also highly conserved, such as Pro51 and Lys153,[60] already described as important, with a conservation percentage of 97.0 and 92.1%,

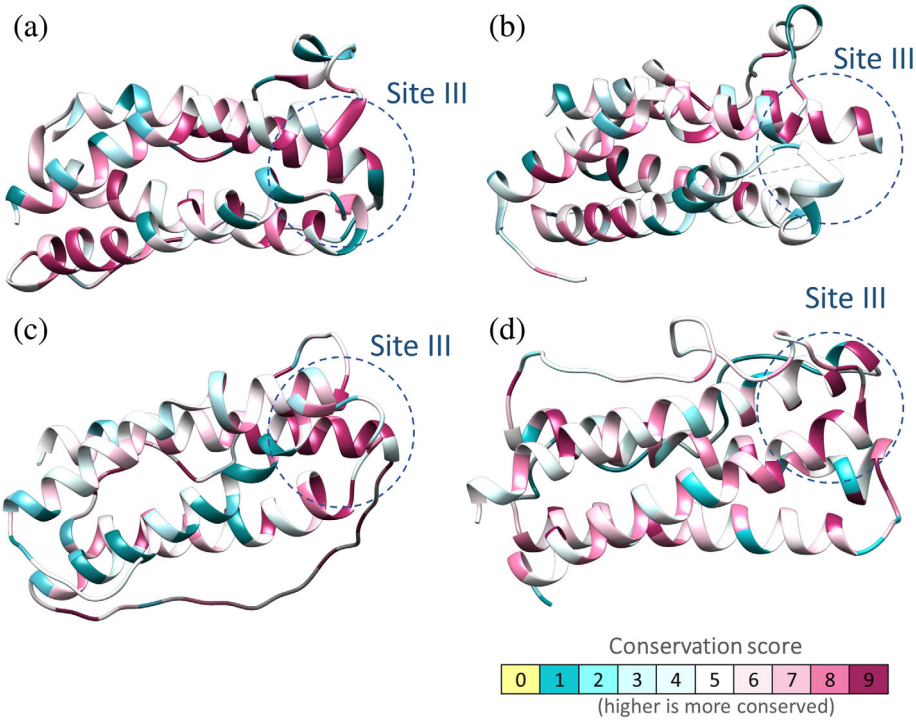

FIGURE 6 Residue conservation in LIF (a), OSM (b), m17 (c) and CT-1 (d) after analysis by the ConSurf server. 3D model structures of human LIF (PDB reference 1EMR), human OSM (PDB reference 1EVS), CT-1 (SWISS-model Q16619), and m17 (HHpred structure from the Consurf platform using MODELLER) were used as base models. The conservation score has a color code and numbered score (0–9). Site III of binding to LIFR is highlighted in blue

TABLE 1 Conservation percentage (%) of the residues in the FXXK Motif of LIF, OSM, m17, and CT-1 proteins

Protein	LIF		OSM		m17		CT-1		Total (as from new alignment)	
FXXK motif	F	K	F	K	F	K	F	K	F	K
F	99.1	-	97.5	-	98.5	-	97.3	-	98.4	-
K	-	98.9	-	84.9	-	100	-	96.2	-	95.5
R	-	0.4	-	14.6	-	-	-	-	-	3.3
Others	0.4	0.2	1.2	-	1.5	-	1.1	1.1	1.0	0.5
Gaps	0.6	0.6	1.3	0.5	-	-	1.6	2.7	0.6	0.7

Note: The percentage of conservation was obtained by the consensus values obtained after alignment in UGene software. Total (as from new alignment) refers to the percentage of conservation obtained from an alignment with all 607 sequences and not to the average of the alignments from each protein.

respectively. Other highly conserved residues from this interface are Phe52 with 98.1% conservation, Pro106 with 98.1%, and Asn105 with 97.4%. On the other hand, some residues from this interface have very low conservation and probably low impact on the ligand-receptor binding. These receptors are Leu103 with 8.1%, Ser107 with 9.8%, Asp154 with 32.3% and Val155 with 29.7% conservation. In CT-1, Pro51 have 98.4% conservation and Phe52 has 95.1%, whereas we found no structural similarity to the Lys153 from LIF protein. OSM was described to have the Pro51 substituted by Leu40 structurally[61] and this residue presented a 96.7% conservation. Once again, no comparable Lys153 residue was found in OSM. In m17 we found no relatable residues. A recent study using docking-docking simulations with LIF, OSM, and LIFR revealed some residues in LIF and OSM for binding to LIFR.[30] Although in LIF most of the residues identified in the study correlate with our analysis in terms of conservation, for OSM, we found that only the residues near site III and related to LIFR binding were conserved in our alignment.

The conservation of most of site III residues explains the cross reactivity between proteins from different species.[18,25,26] Nevertheless, a question arises as the human orthologs of LIF and OSM bind to both human and murine receptors but not the opposite.[18,25,26] Structural and sequence differences in both cytokines and receptors could influence specific binding interactions as, for instance, it is described that specific residues of the AB loop of mouse OSM prevents the binding to the human receptor.[18] Our alignment results demonstrate that the AB loop region in OSM is a highly variable region, and could influence the cross reactivity of this protein to receptors of different species. Besides the highly conserved Gln residue found at the end of helix A in all four proteins, a low conservation percentage associated in the AB loop residues seems to be consistent for the proteins tested. Regarding the two central residues in the FXXK motif, we found high protein to protein and organism class to class variation, highlighting their limited relevance for cytokine activity.[27–29] Still, these residues are highly conserved in m17, indicating low variation over time in Actinopterygii.

IL-6 family is known to interact with gp130 using site II at different locations in this receptor.[76,77] Point mutations in the OSM binding site of gp130 inhibited the binding of OSM but not of LIF.[77] This is attributed to the particular promiscuous behavior of gp130, in which it could behave more as a converter protein than as a binding receptor.[78] In a review with unpublished data, Bravo and Heat[79] disclosure three residues essential for hLIF binding to gp130: Gln29, Asn128, and Gly124. We observed that these residues were highly conserved in both LIF and OSM (Figure 7a,b). In fact, site II placement is very similar in LIF and OSM, facing outwards protein helixes A and C. These results are in agreement with what was found in the PIM of the site II in Figure 1b. We did not find the same pattern in either m17 (Figure 7c) or CT-1 (Figure 7d). Instead, in m17, the B-helix and CD loop have some conserved residues namely, Ile58, Gly60, Ser69, Pro125, and Asn126 (Figure 7c). It could be hypothesized that these conserved residues can be related with gp130 binding as this receptor was already described in fishes.[80] The lack of identity found in the PIM of Figure 1b also corroborates these results. CT-1 has conserved residues near both the N- and C- terminal regions facing to the core of the protein, in opposition to what is found for LIF and OSM. Residues Ile27, Leu81, Arg86, Leu140, and Asp191 have conservation values over 90% in the UGene software and over nine in the Consurf algorithm, and could indicate a different binding behavior to gp130 from CT-1. Some of these residues were also organism class dependent. For instance, Ser28 in hLIF is present only in primates, being substituted by an L-asparagine in the remaining mammals, and by an L-arginine or L-lysine in birds. In CT-1, Leu140 from the human variant is present in mammals but substituted by a methionine in reptiles and amphibians. To understand

if the structure prediction of CT-1 influences the results, we compared the Consurf analysis with CT-1 structure available in the Swiss Modeler website with the one predicted by the algorithm I-TASSER[81] (Figure S4). Both structures were similar and no relevant differences were found in terms of residue conservation.

Other highly conserved residues include cysteines, indicating that the tertiary conformation is important for activity, as confirmed for human recombinant versions of LIF.[82] The six cysteines in LIF protein sequences are maintained at the same positions (for the formation of three disulfide bonds) in most of the sequences. Exceptions are for the bar-tailed godwit (*Limosa lapponica baueri*), with only two cysteines, and for the brownbanded bambooshark (*Chiloscyllium punctatum*), showing four cysteine residues and much greater similarity to m17 sequences. As for OSM and m17, these have generally four highly conserved cysteines corresponding to two disulfide bonds, and CT-1 only presents two conserved cysteines, corresponding to one disulfide bond. In this case, while one Cys residue in CT-1 is highly conserved, the placement of the remaining residue is highly variable.

OSM has shown activities that overlap with those described for LIF.[2] While similar in activity, human OSM displays an extra loop (BC) in its structure, compared with human LIF (Figure 8a). Studies demonstrate that removal of this BC loop potentiates OSM binding,

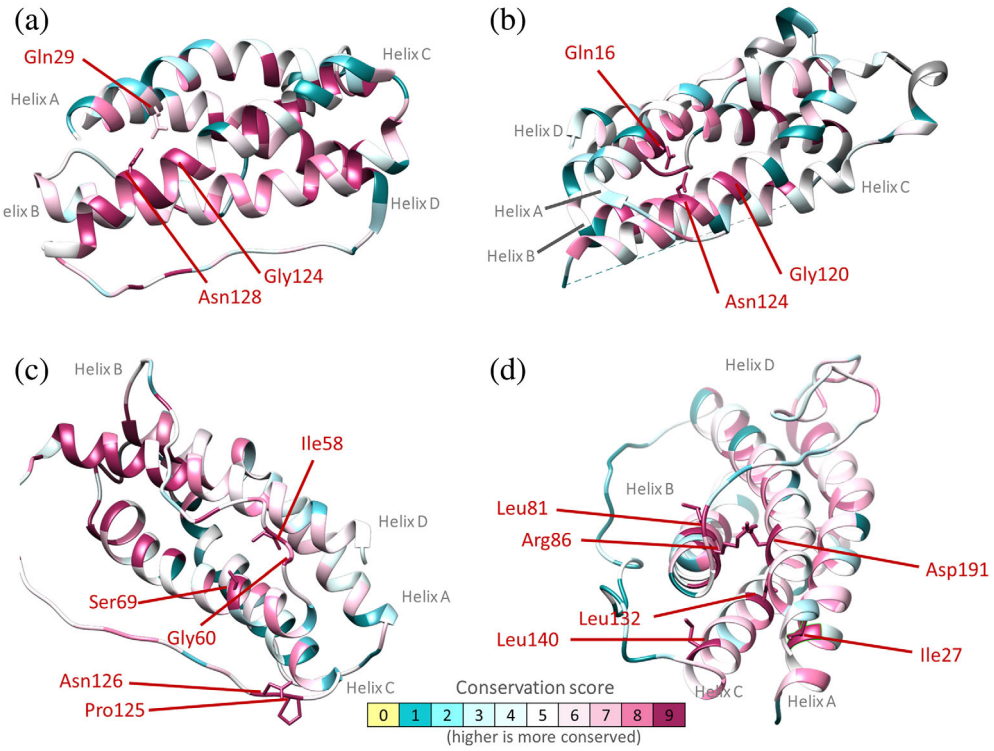

FIGURE 7 Residue conservation in LIF (a), OSM (b), m17 (c) and CT-1 (d) associated with gp130 binding. 3D model structures of human LIF (PDB reference 1EMR), human OSM (PDB reference 1EVS), CT-1 (SWISS-model Q16619), and m17 (HHpred structure from the Consurf platform using MODELLER) were used as base models. The conservation score has a color code and numbered score (0–9)

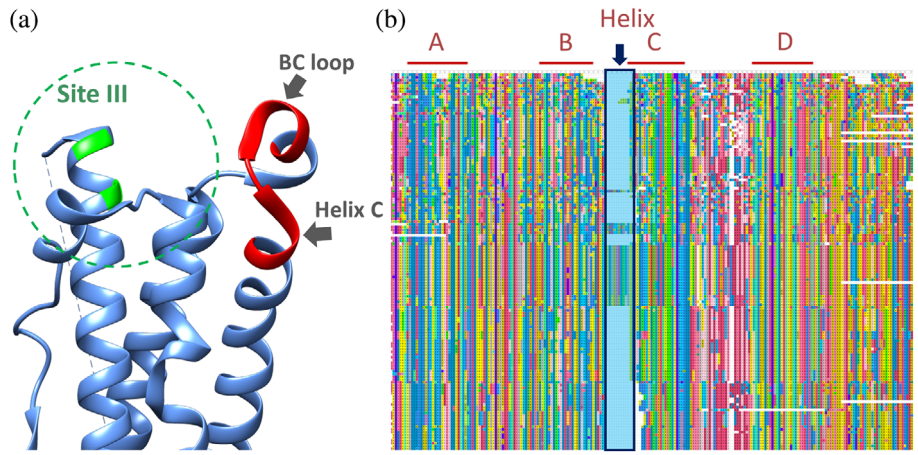

FIGURE 8 Sequence differences in OSM from primates when compared to the remaining mammals. (a) 3D structure of hOSM (PDB reference 1EVS) highlighting the FXXK motif (green) of site III and the region near the BC only present in primates (red); (b) sequence alignment of all OSM mammal sequences denoting the extra region near the BC loop (dark blue arrow) found for primates and their placement in terms of helix distribution

resulting in increased activity.[28] We performed an alignment of all OSM mammal sequences and identified a section with an increased number of residues that only appears in primates (Figure 8b). This extra section includes part of the BC loop and the C-helix, and indicates a speciation event that occurred in all primates. Considering the inhibitory effect of the BC loop size in humans, it may be hypothesized that this is a control step to reduce OSM activity in specific contexts, as undesirable effects of high concentrations of cytokine are described for several pathways.[83,84]

With the ascending number of sequenced genomes and proteomic data, the study of protein phylogeny became less complicated and gives a stronger and more realistic perspective of species-to-species variability. Herein, we described the sequence analysis of four members of the IL-6 cytokine family – LIF, OSM, CT-1, and m17 – which share the ability to direct signal transduction through LIFR. We obtained 1,092 different sequences spanning seven classes of vertebrates, although each protein has its own set of sequences from different classes. We hypothesize that these proteins evolved from common ancestors through gene duplication events and had specialized functions, while maintaining features common to all IL-6 family members. Furthermore, physical–chemical properties such as the molecular weight, pI and the grand average of hydropathy (GRAVY) demonstrated to, on one hand, be central to establish protein type and class of organisms they belong to, and on the other hand, being conserved not only within the same protein type but also in the same class for the different proteins. It is, to our knowledge, the first time an organism class classification system was created based solely on the properties obtainable by protein sequences. We also observed that LIF, OSM, m17, and CT-1 have high conservation values of the characteristic LIFR-binding motif FXXK across all proteins and in all classes from which sequences were retrieved. On the other hand, gp130 binding motif has shown similarity between LIF and OSM, but a different placement in CT-1, while being absent in m17. Furthermore, residues related to the maintenance of protein typology, such as cysteine residues, are also highly conserved. OSM protein displayed the highest variability in the FXXK motif, with an event in time that changed the lysine to arginine in the order Carnivora, although probably maintaining similar functions. In fact, OSM sequence analysis has pinpointed another event in primates due to an extended BC loop, when compared to the remaining mammals, that influences binding to OSM receptor.

This study shows that sequence diversity analysis could provide much more information than just phylogeny, contributing not only to the understanding of the IL-6 cytokine family evolution, their organism class, and protein-dependent characteristics, but also to the confirmation on the importance of conserved motifs. Furthermore, the results here obtained demonstrate the potential use of machine learning approaches for the creation of new antagonists that will impact future medical treatment of disorders associated with imbalances in these cytokines.

AUTHOR CONTRIBUTIONS

Raul Machado: Conceptualization (equal); funding acquisition (equal); writing – original draft (equal); writing – review and editing (equal). **Andreia C. Gomes:** Conceptualization (equal); funding acquisition (equal); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

This work was supported by the "Contrato-Programa" UIDB/04050/2020 funded by national funds through the FCT I.P. and project FUN2CYT: Harnessing the potential for biomedical applications of pleiotropic cytokines LIF and oncostatin M (POCI-01-0145-FEDER-030568) supported by Programa Operacional Competitividade e Internacionalização (FEDER) and FCT, I.P. Raul Machado acknowledges FCT I.P. for funding within the Scientific Employment Stimulus instrument (CEECIND/00526/2018).

DATA AVAILABILITY STATEMENT

All the alignment files and dataset are available in the supplementary materials. Other data requests underlying this article will be shared on reasonable request to the corresponding authors.

ORCID

André da Costa https://orcid.org/0000-0002-4098-2083
Ricardo Franco-Duarte https://orcid.org/0000-0002-2333-6127
Raul Machado https://orcid.org/0000-0002-9477-9945
Andreia C. Gomes https://orcid.org/0000-0002-0567-064X

REFERENCES

1. Hirano T, Yasukawa K, Harada H, et al. Complementary DNA for a novel human interleukin (BSF-2) that induces B lymphocytes to produce immunoglobulin. Nature. 1986;324:73–76.
2. Murakami M, Kamimura D, Hirano T. Pleiotropy and specificity: Insights from the interleukin 6 family of cytokines. Immunity. 2019;50:812–831.
3. Rose-John S. Interleukin-6 family cytokines. Cold Spring Harb Perspect Biol. 2018;10:1–17.
4. Pinho V, Fernandes M, da Costa A, Machado R, Gomes AC. Leukemia inhibitory factor: Recent advances and implications in biotechnology. Cytokine Growth Factor Rev. 2020;52:25–33.

5. Kazakov AS, Sokolov AS, Permyakova ME, et al. Specific cytokines of interleukin-6 family interact with S100 proteins. Cell Calcium. 2022;101:102520.
6. Jorgensen MM, de la Puente P. Leukemia inhibitory factor: An important cytokine in pathologies and cancer. Biomolecules. 2022;12:217.
7. Masjedi A, Hajizadeh F, Beigi Dargani F, et al. Oncostatin M: A mysterious cytokine in cancers. Int Immunopharmacol. 2021;90:107158.
8. López-Yoldi M, Moreno-Aliaga MJ, Bustos M. Cardiotrophin-1: A multifaceted cytokine. Cytokine Growth Factor Rev. 2015;26: 523–532.
9. Pan CM, Wang ML, Chiou SH, Chen HY, Wu CW. Oncostatin M suppresses metastasis of lung adenocarcinoma by inhibiting SLUG expression through coordination of STATs and PIASs signalings. Oncotarget. 2016;7:60395–60406.
10. David E, Tirode F, Baud'Huin M, et al. Oncostatin M is a growth factor for Ewing sarcoma. Am. J. Clin. Pathol. 2012; 181:1782–1795.
11. Ali SA, Malakar D, Kaushik JK, Mohanty AK, Kumar S. Recombinant purified buffalo leukemia inhibitory factor plays an inhibitory role in cell growth. PLoS One. 2018;13:1–18.
12. An L, Liu J, Du Y, et al. Synergistic effect of cysteamine, leukemia inhibitory factor, and Y27632 on goat oocyte maturation and embryo development in vitro. Theriogenology. 2018;108: 56–62.
13. Hanson JM, Mol JA, Meij BP. Expression of leukemia inhibitory factor and leukemia inhibitory factor receptor in the canine pituitary gland and corticotrope adenomas. Domest Anim Endocrinol. 2010;38:260–271.
14. Dutton LC, Dudhia J, Guest DJ, Connolly DJ. Inducing pluripotency in the domestic cat (Felis catus). Stem Cells Dev. 2019; 28:1299–1309.
15. Jalvy S, Veschambre P, Fédou S, Rezvani HR, Thézé N, Thiébaud P. Leukemia inhibitory factor signaling in *Xenopus embryo*: Insights from gain of function analysis and dominant negative mutant of the receptor. Dev Biol. 2019;447:200–213.
16. Horiuchi H, Tategaki A, Yamashita Y, et al. Chicken leukemia inhibitory factor maintains chicken embryonic stem cells in the undifferentiated state. J Biol Chem. 2004;279:24514–24520.
17. Murphy MJ, Halow NG, Royer PA, Hennebold JD. Leukemia inhibitory factor is necessary for ovulation in female rhesus macaques. Endocrinology. 2016;157:4378–4387.
18. Adrian-Segarra JM, Sreenivasan K, Gajawada P, Lörchner H, Braun T, Pöling J. The AB loop of oncostatin M (OSM) determines species-specific signaling in humans and mice. J Biol Chem. 2018;293:20181–20199.
19. Hanington PC, Belosevic M. Interleukin-6 family cytokine M17 induces differentiation and nitric oxide response of goldfish (*Carassius auratus* L.) macrophages. Dev Comp Immunol. 2007;31:817–829.
20. Fujiki K, Nakao M, Dixon B. Molecular cloning and characterisation of a carp (*Cyprinus carpio*) cytokine-like cDNA that shares sequence similarity with IL-6 subfamily cytokines CNTF, OSM and LIF. Dev Comp Immunol. 2003;27:127–136.
21. Hwang JY, Santos MD, Kondo H, Hirono I, Aoki T. Identification, characterization and expression of a novel cytokine M17 homologue (MSH) in fish. Fish Shellfish Immunol. 2007;23: 1256–1265.
22. Davis SM, Pennypacker KR. The role of the leukemia inhibitory factor receptor in neuroprotective signaling. Pharmacol Ther. 2018;183:50–57.
23. Jones SA, Jenkins BJ. Recent insights into targeting the IL-6 cytokine family in inflammatory diseases and cancer. Nat Rev Immunol. 2018;18:773–789. https://doi.org/10.1038/s41577-018-0066-7.
24. Du Q, Qian Y, Xue W. Molecular simulation of Oncostatin M and receptor (OSM–OSMR) interaction as a potential therapeutic target for inflammatory bowel disease. Front Mol Biosci. 2020;7.
25. Owczarek CM, Zhang Y, Layton MJ, Metcalf D, Roberts B, Nicola NA. The unusual species cross-reactivity of the leukemia inhibitory factor receptor α-chain is determined primarily by the immunoglobulin-like domain. J Biol Chem. 1997;272: 23976–23985.
26. Pennica D, Swanson TA, Shaw KJ, et al. Human cardiotrophin-1: Protein and gene structure, biological and binding activities, and chromosomal localization. Cytokine. 1996;8:183–189.
27. Plun-Favreau H, Perret D, Diveu C, et al. Leukemia inhibitory factor (LIF), cardiotrophin-1, and oncostatin M share structural binding determinants in the immunoglobulin-like domain of LIF receptor. J Biol Chem. 2003;278:27169–27179.
28. Chollangi S, Mather T, Rodgers KK, Ash JD. A unique loop structure in oncostatin M determines binding affinity toward oncostatin M receptor and leukemia inhibitory factor receptor. J Biol Chem. 2012;287:32848–32859.
29. Perret D, Guillet C, Elson G, et al. Two different contact sites are recruited by cardiotrophin-like cytokine (CLC) to generate the CLC/CLF and CLC/sCNTFRα composite cytokines. J Biol Chem. 2004;279:43961–43970.
30. Du Q, Qian Y, Xue W. Cross-reactivity of two human IL-6 family cytokines OSM and LIF explored by protein-protein docking and molecular dynamics simulation. Biochim Biophys Acta (BBA) - General Subjects. 2021;1865:129907.
31. Drechsler J, Grötzinger J, Hermanns HM. Characterization of the rat oncostatin m receptor complex which resembles the human, but differs from the murine cytokine receptor. PLoS One. 2012;7:1–12.
32. Kanegi R, Hatoya S, Tsujimoto Y, et al. Production of feline leukemia inhibitory factor with biological activity in *Escherichia coli*. Theriogenology. 2015;86:604–611.
33. Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–D515.
34. Cunningham F, Achuthan P, Akanni W, et al. Ensembl 2019. Nucleic Acids Res. 2019;47:D745–D751.
35. Agarwala R, Barrett T, Beck J, et al. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2016;44:D7–D19.
36. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.
37. Okonechnikov K, Golosova O, Fursov M, et al. Unipro UGENE: A unified bioinformatics toolkit. Bioinformatics. 2012; 28:1166–1167.
38. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–1973.

39. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–274.
40. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44:W232–W235.
41. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–589.
42. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the ultrafast bootstrap approximation. Molecular biology and evolution. Mol Biol Evol. 2018;35:518–522.
43. Letunic I, Bork P. Interactive tree of life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. 2019;47:256–259.
44. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4:1–5.
45. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst Biol. 2018;67:901–904.
46. Kozlowski LP. IPC - Isoelectric Point Calculator. Biol Direct. 2016;11:1–16.
47. Fuchs S Gravy Calculator.
48. Demšar J, Curk T, Erjavec A, et al. Orange: Data mining toolbox in python. J. Mach. Learn. Res. 2013;14:2349–2353.
49. Demšar J, Leban G, Zupan B. FreeViz-An intelligent multivariate visualization approach to explorative analysis of biomedical data. J Biomed Inform. 2007;40:661–671.
50. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
51. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825–2830.
52. Ashkenazy H, Abadi S, Martz E, et al. ConSurf 2016: An improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016;44:W344–W350.
53. Landau M, Mayrose I, Rosenberg Y, et al. ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. 2005;33:299–302.
54. Glaser F, Pupko T, Paz I, et al. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics. 2003;19:163–164.
55. Pettersen EF, Goddard TD, Huang CC, et al. UCSF chimera - a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–1612.
56. Heinrich PC, Behrmann I, Haan S, Hermanns HM, Müller-Newen G, Schaper F. Principles of interleukin (IL)-6-type cytokine signalling and its regulation. Biochem J. 2003;374:1–20.
57. Feng S, Stiller J, Deng Y, et al. Dense sampling of bird diversity increases power of comparative genomics. Nature. 2020;587:252–257.
58. Oldefest M, Nowinski J, Hung CW, et al. Upd3 - An ancestor of the four-helix bundle cytokines. Biochem Biophys Res Commun. 2013;436:66–72.
59. Rose TM, Lagrou MJ, Fransson I, et al. The genes for oncostatin m (OSM) and leukemia inhibitory factor (LIF) are tightly linked on human chromosome 22. Genomics. 1993;17:136–140.
60. Hudson KR, Vernallis AB, Heath JK. Characterization of the receptor binding sites of human leukemia inhibitory factor and creation of antagonists. J Biol Chem. 1996;271:11971–11978.
61. Huyton T, Zhang JG, Luo CS, et al. An unusual cytokine:Ig-domain interaction revealed in the crystal structure of leukemia inhibitory factor (LIF) in complex with the LIF receptor. Proc Natl Acad Sci U S A. 2007;104:12737–12742.
62. San Mauro D. A multilocus timescale for the origin of extant amphibians. Mol Phylogenet Evol. 2010;56:554–561.
63. Torres-Sánchez M, Gower DJ, Alvarez-Ponce D, Creevey CJ, Wilkinson M, Di SM. What lies beneath? Molecular evolution during the radiation of caecilian amphibians. BMC Genomics. 2019;20:1–13.
64. Mayr G. The phylogenetic affinities of the shoebill (*Balaeniceps rex*). J Ornithol. 2003;144:157–175.
65. Huang T, Peng J, Zhao Y, Xu Z. The complete mitochondrial genome of *Pelecanus occidentalis* (Pelecaniformes: Pelecanidae) and its phylogenetic analysis. Mitochondrial DNA B Resour. 2018;3:782–783.
66. Peng QL, Nie LW, Pu YG. Complete mitochondrial genome of Chinese big-headed turtle, *Platysternon megacephalum*, with a novel gene organization in vertebrate mtDNA. Gene. 2006;380:14–20.
67. Gaubert P, Antunes A, Meng H, et al. The complete phylogeny of pangolins: Scaling up resources for the molecular tracing of the Most trafficked mammals on earth. J Hered. 2018;109:347–359.
68. Amrine-Madsen H, Koepfli KP, Wayne RK, Springer MS. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Mol Phylogenet Evol. 2003;28:225–240.
69. Abe T, Mikekado T, Haga S, et al. Identification, cDNA cloning, and mRNA localization of a zebrafish ortholog of leukemia inhibitory factor. Comp Biochem Physiol B Biochem Mol. 2007;147:38–44.
70. Upham NS, Esselstyn JA, Jetz W. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. Plos Biol. 2019;17(12):e3000494.
71. Hinchliff CE, Smith SA, Allman JF, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A. 2015;112:12764–12769.
72. Layton MJ, Owczarek CM, Metcalf D, et al. Conversion of the biological specificity of murine to human leukemia inhibitory factor by replacing 6 amino acid residues. J Biol Chem. 1994;269:29891–29896.
73. Qi Y. Random Forest for bioinformatics. Ensemble machine learning. Boston, MA: Springer US, 2012; p. 307–323.
74. Fantini M, Lisi S, De Los RP, Cattaneo A, Pastore A. Protein structural information and evolutionary landscape by in vitro evolution. Mol Biol Evol. 2020;37:1179–1192.
75. Donald JE, Hubner IA, Rotemberg VM, Shakhnovich EI, Mirny LA. CoC: A database of universally conserved residues in protein folds. Bioinformatics. 2005;21:2539–2540.
76. Boulanger MJ, Bankovich AJ, Kortemme T, Baker D, Garcia KC. Convergent mechanisms for recognition of divergent cytokines by the shared signaling receptor gp130. Mol Cell. 2003;12:577–589.

77. Olivier C, Auguste P, Chabbert M, Lelièvre E, Chevalier S, Gascan H. Identification of a gp130 cytokine receptor critical site involved in oncostatin M response. J Biol Chem. 2000;275:5648–5656.
78. Müller-newen G. The cytokine receptor gp130: faithfully promiscuous. Sci STKE. 2012;2003(201):PE40.
79. Bravo J, Heat JK. Receptor recognition by gp130 cytokines. EMBO J. 2000;19:2399–2411.
80. Kaneda M, Odaka T, Suetake H, Tahara D, Miyadai T. Teleost IL-6 promotes antibody production through STAT3 signaling via IL-6R and gp130. Dev Comp Immunol. 2012;38:224–231.
81. Yang J, Zhang Y. I-TASSER server: New development for protein structure and function predictions. Nucleic Acids Res. 2015;43:W174–W181.
82. Schmelzer CH, Harris RJ, Butler D, Yedinak CM, Wagner KL, Burton LE. Glycosylation pattern and disulfide assignments of recombinant human differentiation-stimulating factor. Arch Biochem Biophys. 1993;302:484–489.
83. Stawski L, Trojanowska M. Oncostatin M and its role in fibrosis. Connect Tissue Res. 2019;60:40–49.
84. Stephens JM, Elks CM. Oncostatin M: Potential implications for malignancy and metabolism. Curr Pharm Des. 2017;23:3645–3657.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Costa Ad, Franco-Duarte R, Machado R, Gomes AC. Uncovering the promiscuous activity of IL-6 proteins: A multi-dimensional analysis of phylogeny, classification and residue conservation. Protein Science. 2022;31(11):e4469. https://doi.org/10.1002/pro.4469