International Journal of Swarm Intelligence and Evolutionary Computation

OPEN ACCESS Freely available online

Research Article

Bioinformatics and Artificial Intelligence Approaches for Unraveling Evolutionary Mechanisms in Taxonomic Groups of Neurohypophysial Hormones Family

Alberto F. de Oliveira Jr[1*], Marcelo Querino Lima Afonso[2], Vasco Ariston Azevedo[2], Manuel Lemos[1]

[1]Department of Biochemistry and Genetics, University of Beira Interior, R. Marquês de Ávila e Bolama, 6201-001, Covilhã/Portugal;[2]Department of Biochemistry and Immunology, Federal University of Minas Gerais, Av. Pres. Antônio Carlos, 6627 Pampulha, 31270-901, Belo Horizonte, Minas Gerais/Brasil

ABSTRACT

Together Oxytocin and Vasopressin set the neurohypophysial hormones that form a family of structurally and functionally related peptide hormones. However, the biological function of these proteins may vary depending on their taxonomic classification. In our study, using a broad of bioinformatics and machine learning techniques, we described the role of sets of coevolved amino acids in determining the taxonomic classes of neurohypophysial hormone sequences. Withal, it would be possible to correlate that certain taxonomic classes can still be classified from the presence of specific amino acids from these coevolved sets, bringing more light around how molecular evolution can describe the structure and function.

Keywords: Oxytocin; Vasopressin; Evolution; Coevolution of amino acids; Coevolved sets; Machine learning; Molecular phylogeny; Neurohypophysial hormones

BACKGROUND

Oxytocin and Vasopressin are both the main representatives of neurohypophysial hormones which together form a family of structurally and functionally related peptide hormones. Vasopressins are proteins that are part of a set of the hormonal protein superfamily, which can be observed in both vertebrates and invertebrates [1,2,3].

The biological function of these proteins may vary depending on their taxonomic classification. In humans, this functionality is attributed to their acting as neurohypophyseal hormones, belonging to a family of structurally and functionally related nonapeptides that are synthesized as part of a larger precursor molecule comprising a signal peptide, the nonapeptide hormone, and a neurophysin, required for the targeting of these peptides to the regulated secretory pathway [4]. During processing the signal, peptide is removed in the endoplasmic reticulum and the remaining precursor is then packaged into neurosecretory vesicles in the Golgi apparatus. This packaging occurs during the axonal transport of these vesicles to terminals along with cleavage events by prohormone convertases [5].

Vasopressins also contain Copeptin, which is an additional glycosilated C-terminal peptide whose function is still unknown but has been associated with a chaperone-like activity during the structural assembly of provassopressins by interacting with the calnexin/calreticulin system since the prohormone is reported to inefficiently fold in the absence of this glycopeptide [6]. Due to the observed release of copeptin in the circulation after various stress-related signals, there is also some speculation to a putative signaling function of this peptide. This also is reflected in the use of copeptin as a biomarker for the clinical detection of multiple diseases [7]. In mice, copeptin immunoreactivity was not found to be uniform in different populations of neurons that produced Vasopressin, possibly due to low protease activity and incomplete processing of the prohormone in these neurons [5]. Vasopressin peptides are basic peptides whilst oxytocin peptides are neutral this property is defined by the eighth residue which can be Arg/Lys or Leu/Ile/Gln/Val, respectively [8].

Neurohypophysial hormones can also display selective and non-covalent interactions with their receptors, a characteristic that has been especially described for mammals. Information regarding the evolutionary degree that revolves around these proteins was previously considered to be scarce, but focused efforts in the last two decades yielded a large amount of data to be explored, including especially enzymatic and functional characterizations of various homologous neuropeptides [9-12]. Recently a major phylogenetic analysis of insect Oxytocin/Vasopressin-like neuropeptides focused on discussing the presence of the inotocin signaling system, the

Correspondence to: Alberto F. de Oliveira Jr, Department of Biochemistry and Genetics, University of Beira Interior, R. Marquês de Ávila e Bolama, 6201-001, Covilhã/Portugal, E-mail: afojunior@gmail.com

Received date: March 29, 2021; **Accepted date:** April 12, 2021; **Published date:** April 19, 2021

Citation: Oliveira AFD, Afonso MQL, Azevedo VA, Lemos M (2021) Bioinformatics and Artificial Intelligence Approaches for Unraveling Evolutionary Mechanisms in Taxonomic Groups of Neurohypophysial Hormones Family. Int J Swarm Evol Comput 2021; 10:211

oxytocin/vasopressin orthologue of this clade [13]. A previous gene tree of vertebrate and invertebrate sequences of this family showed that the different genes of this family seen to be clustered in a manner akin to the taxonomic trees of their clades [14]. Other noteworthy mentions also include studies in which taxonomic trees were annotated for the presence or absence of specific Oxytocin/Vasopressin homologs in the literature [15,16]. Various studies are often more focused on the Oxytocin/Vasopressin receptors and solely mention the nonapeptide hormone sequence when considering the evolution of this protein family [17].

From a structural perspective, information is still highly scarce since only the structures of Bovine Oxytocin and Vasopressin alongside their neurophysins (I and II, respectively) have been elucidated by a series of praiseworthy works from the Esther W. Brestow group which provided various insights regarding peptide and inhibitor binding by neurophysins and structural dynamics of neurophysin dimerization [18,19,20]. More frequently found in the literature are studies focused on the structural dynamics of the neuropeptide alone, which have been reported for various related hormones such as Oxytocin, Vasopressin, Conopressin, and Mesotocin, also including mutants in some cases [21-26].

Homologs of Oxytocin and Vasopressin are found in jawed vertebrates (gnasthostomes) but not in jawless fish (Agnatha) which possess a single gene for a related neuropeptide named Vasotocin. As such, a tandem gene duplication event of an Oxytocin/Vasopressin homolog in a common ancestor of the gnasthostome lineage is attributed to be responsible for the presence of both genes in existent clades. Moreover, gene duplication events and Whole-genome duplication events occurring during vertebrate evolution alongside their losses in specific lineages resulted in the variable numbers of these homologs found for different vertebrate lineages, and this also includes the Oxytocin/Vasopressin receptor genes [4].

Synteny and conservation analysis of genes from these protein families revealed the presence of the copeptin sequence in the teleost fish isotocin hormone, which does not display an Arginine residue between this sequence and the neurophysin molecule, leading to the predictions of their non-cleavage in these clades. These analyses allowed a detailed description of evolutionary events on the neurohypophyseal gene locus such as an inversion event occurring in placental mammals leading to an inverted orientation of the Oxytocin/Vasopressin genes for species of this clade, and multiple rearrangements in teleost fish not observed in other vertebrates, possibly related to their lineage-specific whole genome duplication event [8]. Further extensions to these analyses described the presence of vasotocin/vasopressin but not oxytocin related genes in jawless vertebrates such as the Japanese lamprey and hagfish. When two copies of the Oxytocin/Vasopressin genes are present in an organism, they have been found to display a high sequence conservation at both the nucleotide and protein level in their second exon. As such, other unexplored regulatory roles have been proposed for this region [27].

Proteins of this family are also found in invertebrates, alongside described roles in reproduction, feeding, and water/salt homeostasis which were recently thoroughly revised [4]. Conopressins compose a particular group of Oxytocin/Vasopressin homologs found in the venoms of piscivorous cone snails, being possibly related to a complex envenomation strategy and possessing a pharmacological interest [21,28]. The origin of Oxytocin/Vasopressin signaling system is thought to be around ~600 million years old [13].

Vasopressin is also an important protein for the most diverse living beings in the animal kingdom, as it is closely related to physiological and neural diseases. As such, studies regarding this protein are of extreme interest for the medical, veterinary and biotechnological areas. An interesting strategy for characterizing these proteins is through computational studies of molecular evolution. However, the evolutionary events that describe the molecular differences and similarities between sequences of this protein family, which would greatly expand the knowledge about these molecules, are still being elucidated. Amino acid co-evolution analyzes are very promising technique, being currently capable of associating various relevant informations in the genetic and three-dimensional protein structure levels [29,30,31]. Furthermore, methods capable of predicting whether a given set of amino acids are specific to proteins of a given clade or class would be useful in molecular evolution research approaches. In this context, machine learning algorithms can allow exploration of such data, elucidating particularities that are still not well understood. These methods have been shown to be effective and interesting when applied to biological sequence analysis, standing out in evolutionary methods for protein engineering [32], in elucidating and predicting structural domains in proteins [33] and also in the prediction of functions in protein families from sequence input [34].

In our study, we described the role of sets of coevolved amino acids in determining the taxonomic classes of neurohypophysial hormone sequences. Withal, it would be possible to correlate that certain taxonomic classes can still be classified from the presence of specific amino acids from these coevolved sets.

MATERIAL AND METHODS

Protein sequences and Multiple Sequence Alignment (MSA)

A multiple sequence alignment of the neurohypophyseal hormone C-terminal domain protein family (Pfam code: PF00184) downloaded from the Pfam database [35] and then subjected to two filtering procedures. First, to remove possible fragments, each sequence in the alignment was compared to the alignment's HMM profile. Sequences with fewer than 80% alignment coverage were removed from the alignment. Next, in order to reduce possible phylogenetic biasing, all sequences were compared to each other, and whenever two sequences shared an identity score higher than 90%, the smaller sequence of the pair was removed. Since the Pfam HMM align PF00184 protein family sequences lacks their N-terminal end that includes the neurohypophyseal hormone peptide, a second multiple sequence alignment was generated. All full UniProt sequences with cross-references to Pfam PF00184 or PF00220 neurohypophyseal hormone protein families were aligned by MAFFT using the L-INS-i protocol on the CIPRES processing facilities [36,37].The Conan web application was used for all procedures involving alignment filtering (pre-processing), conservation and correlation calculations (see below) [38].

Statistical analysis (conservation and correlation)

Two residue+position pairs in the filtered multiple sequence alignment were determined to be part of a co-occurring group by representing this alignment as a bipartite graph composed by set U representing sequence labels, and set V representing all residues present in the alignment (residue+column position). In the monopartite projection of set V groups of co-occurring residues have a tendency to aggregate in communities. All edges in this projection were then subjected to the backbone extraction method

of Tuminello which allows selecting only the most statistically significant edges after the association of a p-value for each edge in the projection according to the number of proteins that residues i and j share [39]. A minimum p-value of 10-10 was used for as a threshold in this network, along with redundancy filters to remove residue+column position pairs which represented either highly conserved or highly variable residues. As such edges which displayed frequencies above 0.8 or below 0.05 were discarded. Community detection was then performed by a hierarchical agglomerative clustering procedure aimed to maximize the average Jaccard similarity coefficient from the columns of the biadjacency matrix of edge frequency co-variation. Similar procedures were applied to the full generated alignment but frequency filters for overly conserved residues were not applied so conserved motifs such as the cleaved neurohypophyseal hormone peptide and disulfide bridge residue correlations could possibly be observed.

Phylogenetic analysis

An evolutionary model was determined using MEGA [40], which examines the alignment, obtains the closest evolutionary model and, thus, assists in building a tree. The popular WAG model, which combines the estimation of transition and scoring matrices by a maximum-likelihood approach, was selected as an evolutionary model. Protein sequences for transcripted Oxytocin/Vasopressin genes were obtained from UniProt and cross-referenced to their respective nucleotide sequences when possible. We used the tool Seaview [41], which is a multi-platform, graphical user interface for multiple sequence alignment and molecular phylogeny, in which the maximum likelihood estimation method from the PhyML software [42] and a multiple sequence alignment generated from ClustalX were used [43]. Branch support consistencies were evaluated using the nonparametric bootstrap test [44] with 1000 replicates and the approximate likelihood ratio test (ALRT) [45]. Finally, the tree was edited highlighting the branches, organisms (sequences) and bootstrap value using Figtree [46,47].

Machine learning approaches

In order to assess whether the correlated amino acid sites could determine a sequence's taxonomic class, the Orange tool [48] was used for testing possible predictive machine learning algorithms. 60% of our data set was used for the algorithm's training and the remaining 40% for validation. The taxonomic classes of proteins in the validation set were not informed to the algorithm in order to assert how precise the trained model was in recognizing the patterns of the different classes. Three different machine learning algorithms were used: K-nearest neighbors (Knn) searches for k closest training examples in feature space using their average for prediction; the process was performed with a number of neighbors 5, euclidian metric for distances and uniform weighting; the CN2 rule induction algorithm is designed for the efficient induction of simple, comprehensible rules of the form "if cond then predict class", even in cases where noise may be present; the process was performed with ordered rules and the exclusive covering algorithm, with the evaluation measure being set as entropic; lastly, the Random Forest method used for classification, regression and other tasks was applied. Tests were performed using [48]. In the prediction validation options of the Orange tool, the screening was designated several times for processing and testing the data. This step divided our dataset into 10 subgroups, using 9 of the subgroups as training and a remaining subgroup for validation. The validation step repeats this procedure 9 more times, each time using a different subgroup for the procedure.

RESULTS AND DISCUSSION

Decomposition of related amino acid networks in the PF00184 MSA

Our preliminary results, using the Conan web application [29], indicate the presence of 4 communities of correlated amino acids, which may possibly highlight signs of coevolution or covariation of amino acids in the vasopressin superfamily (Figure 1). Community 1 is composed by a group of amino acids that have a low degree of conservation throughout the filtered multiple sequence alignment. This indicates that such amino acids will mostly appear in a sequence along with the presence of other amino acids in this community. Therefore, it is possible to infer that communities with lower conservation values have correlated amino acids with higher taxonomic specificity than those with higher conservation values when the function exerted by amino acids in this community emerges in a single clade.

In contrast, residues in community 2 display much higher frequencies, possibly exerting more widespread functions in a protein family such as structural roles in either biochemical or physicochemical contexts despite the existence of different taxonomic classes. Communities 3 and 4 curiously stand out from the others by presenting two sets of amino acids capable of differentiating between vertebrate and invertebrate vasopressin sequences. By analyzing the taxonomic lineage of sequences in the multiple sequence alignment, it was possible to associate community 3 to groups of vertebrate animals, including the classes of Mammals, Aves and Actinopterygii. Phe20, from community 2 as an example, is restricted to vertebrates, and is seen as Val20 in community 4, which is found in a more restricted set of sequences belonging to invertebrates, including sequences from Gastropoda, Cephalopoda, Echinoidea, and Polychaeta among others. Consecutively, the amino acids that are correlated with those amino acids described in the respective positions are also restricted to differentiate the sequences according to the large group of vertebrates and invertebrates (Figure 2).

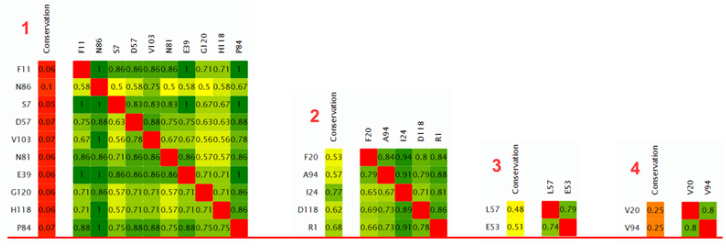

Figure 1: The set of amino acids highlighting 4 different communities with variation of conservation.

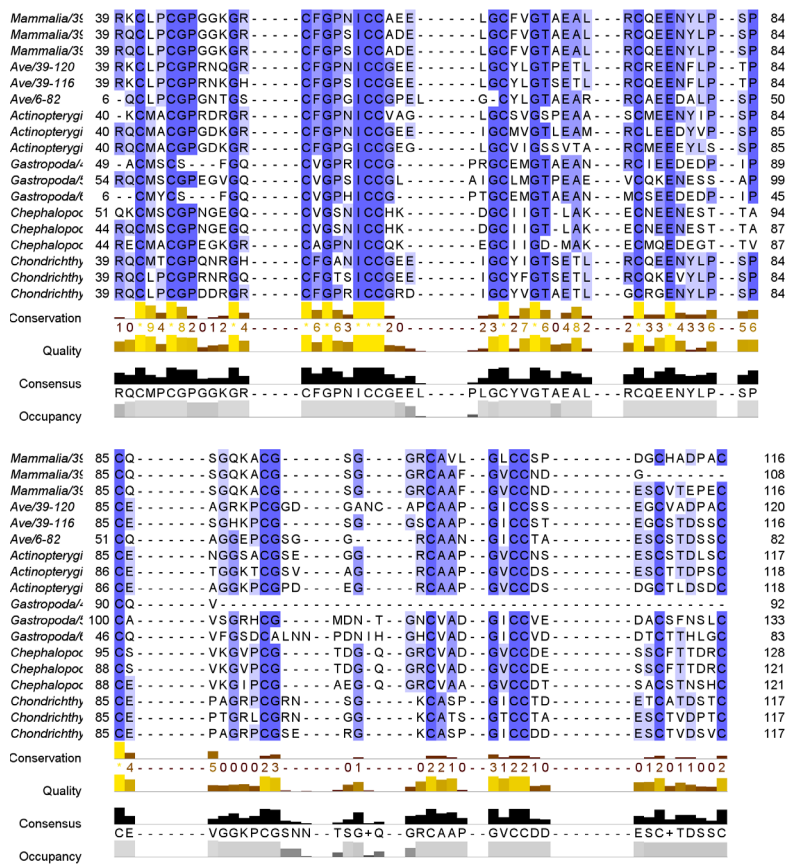

Figure 2: Multiple Sequence Alignment (MSA) to neurohypophyseal hormone super family.

All amino acid position analyzes were based on our multiple sequence alignment. By exploring the initial results further, we noticed that the highlighted positions 11 and 7 present in community 1 display great co-variation patterns in different taxonomic groups. According to the results obtained by the machine learning algorithms, the classes observed according to the composition of correlated amino acids groups, indicate that the Ser7 and Phe11 residues are closely associated the Gastropods class, which is defined by the animals known as snails, slugs, limpets and conches. Throughout all sequences, position 7 is generally occupied by the amino acid Glycine, which is widely distributed in almost all taxonomic classes, showing to be strongly conserved, except for some Gastropods in which this position is occupied by a Serine. Discussions about variations in amino acids referring to the gastropods class were the subject of a study that notified different classes for other neurohypophysial hormones, such as conopressin [49]. Recently, other studies using proteomics and transcriptome techniques have also identified the importance of variations in the hormonal type classification that stood out in invertebrates [50]. According to figure 1, the amino acids Val103 together with Asp57 and Pro84 are correlated. Analyzing the data, once again the taxonomic class of Gastropods shows a close association with this set of amino acids, which reinforces that such amino acids are co-evolving within that taxonomic class. However, position 57 is occupied along with the MSA by a Leucine (Leu), which includes a series of distinct taxonomic classes such as Aves, Mammals and Actinopterygii. In the Insect class, we observed that this position varies from Aspartate to Tyrosine and in the Branchiopod class this position is occupied by a Lysine.

Our results show that the found groups of correlated amino acids in the PF00184 multiple sequence alignment can be easily attributed to specific taxons. Even so, it is possible to discuss with a higher level of detail how homologous biomolecules can present characteristics indicative of evolutionary divergences. Considering the large groups of vertebrates and invertebrates at the molecular level by analyzing protein structures, since Gastropods, Insects and Branchiopod present a degree of residue variations for the characterized positions not found in vertebrate groups such as Aves, Mammals and Actinopterygii, in which these positions display lesser variant flexibility and possess instead a higher degree of conservation.

Positions 20 and 94 are closely related to the two large groups of vertebrates and invertebrates as discussed earlier in some specific sites. Along the alignment the Phe20 position is correlated to Ala94 for the vertebrate group, the same is not seen in the invertebrates, in which the same positions are occupied by different amino acids, with Val20 being correlated with Val94. Interestingly, such positions are found in the region of the protein that characterizes the neurophysin portion that acts as a carrier [51]. This could possibly be associated with hormones transport differentiation in the different organisms, considering the types of taxonomic classes in the large groups of vertebrates and invertebrates. However, no work has been found in the literature that could highlight experiments associated with different functions related to neurophysin considering different taxonomic classes in these groups. As previously described, community 3 highlights the amino acids Glu53 and Leu57, which are correlated exclusively in taxonomic classes belonging to the vertebrate group. However,

other taxonomic groups, inserted in the vertebrate group, have a high residue type variability for position 53, and the same is observed for position 57. Both positions are part of the neurophysin protein domain, as well as the positions of community 4 (see also correlations in the full alignment for possible functions associated with Comm. 4). These findings reinforce the idea that possible variations in the mechanism of transport of neurohypophysial hormones can occur in relation to different taxonomic classes, especially when comparing vertebrates and invertebrates.

Correlation of amino acids and the three-dimensional structure of the preprohormone and function differentiation

The protein structure of the one neurohypophysial hormone can be seen in Figure 3 above. Data provided by the literature, indicate that the cleavage region between the nonapeptide motif and the neurophysin motif is three amino acids, more precisely Gly-Lys-Arg [52]. Analyzing the three-dimensional structure, we observed that the nonapeptide does not interact directly to the cleavage site, despite being close to it, but with amino acids (considering mammalian sequences in our MSA) Glu54, Leu57, Ser61 and Cys63 (In green on Figure 3). The position Leu57 was detected is under coevolution in our previous results. Variations in this same position indicate that Leu57 is closely correlated to sequences of vertebrates, while invertebrates can present in their vast majority Ser57 Asp. Leu57 is interconnected to less than 3 Angstroms of Cys1 that is part of the 9 amino acids of the functional polypeptide. These results indicate the hypothesis of different functional interpretations of neurohypophysial hormones. It is possible that the cleavage events of the 9 nanopeptide portion suffer variations or may even explain the functional differences in the different taxonomic groups, since in view of the results of the structure analysis, Leu57 is directly interacting with the Cys1 of the nonapeptide set, which is an extremely conserved amino acid.

Recent studies associated with molecular evolution in vertebrates highlight supposedly that the Oxytocin/Vasopressin genes arose from a common ancestor of Vasotocin via gene duplication, which resulted in the appearance of such genes in a wide variety of different species such as mammals, fish and birds [53].

Correlations in the full preprohormone uniProt alignment

A total of ten different Communities containing seventy two different residues were found for the full preprohormone uniprot alignment generated by MAFFT. Fifty-three of these residues were found in Community 1, which represented highly conserved residues found in most sequences. This Community's average conservation is 83%, with thirty-one residues possessing frequencies equal or higher than 85% and forty-five residues with frequencies higher than 70% in the alignment. Even the remaining eight residues have at least medium frequency values ranging from 55% to 69%. Highly conserved positions often reflect residues essential for functions widespread in most sequences in a multiple sequence alignment, such as fundamental residues for structural maintenance, catalytic activity or proper signaling in their biological contexts (Supplementary materials Table S1).

Mapping residues from this community to the Bovine Neurophysin II structure complexes with oxytocin confirms such roles. Seven of the nine following hormone nonapeptide amino acids are found in this community (in uppercase CYIqNCPlG in human oxytocin). All of the fourteen disulfides bridged cysteine residues described in this structure can also be found in this community [20]. This community also contains thirteen conserved Glycine and Proline residues which are known to be important for structural stability and conserved protein functional dynamics along disulfide bridges. Also present in this Community is the tripeptide processing signal conserved dibasic amidation motif (GKR) which signals the enzymatic cleavage responsible for prohormone maturation and release of the hormone peptide [54]. Five of the twenty positions of the signal peptide also belong to this community, including two positions located at a cleavage site located right before the nonapeptide sequence. Bovine Neurophysin Leu81 (Leu50 in 1NPO structure numbering) and Glu78 (Glu47 in 1NPO structure numbering) which make important hydrogen bonds and electrostatic interactions with the hormone nonapeptide are also present in this Community [20].

Mapping this Community to the Vasopressin-Neurophysin I sequence in UniProt also revealed that twenty of the positions in this Community are the cause of Neurohypophyseal Diabetes

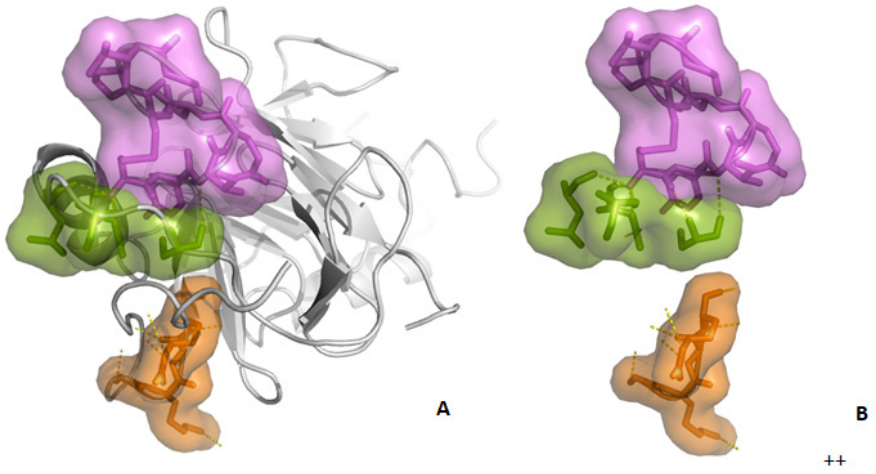

Figure 3: Schematic of the three-dimensional structure of the 1NPO protein, obtained by accessing the PDB database (Protein Data Bank). (A) Representation of the complete structure highlighting in orange the region described in the literature on cleavage with the nonapeptide; in green color the amino acids observed in our results, interacting with the amino acids in the nonapeptide region; in pink color the nine amino acids that make up the nanopeptide. (B) Zoom representation of the highlighted amino acids showing the intramolecular interactions between them.

insipidus (NDI) when mutated, including five of the remaining undescribed nine residues.

Two of these positions, E78 (E47 in 1NPO) and L81 (L50 in 1NPO) have been described for making important electrostatic interactions and hydrogen bonds, respectively, to the nonapeptide hormone [18,20]. Y80 (Y49 in 1NPO) in this Community was described to display perturbations upon hormone binding to the Neurophysin protein and to cause a diminution in binding when mutated due to important contacts with the hormone nonapeptide [18]. E71 (E40) was shown to make important dimerization hydrogen bond contacts along with another three other amino acids. One of these three remaining previously described residues corresponds to a position found in Community 5 (T69-T38 in 1NPO), and another is A99 (A38 in 1NPO) found in Community 7. The side chain of this Glutamate was also observed to show one of the largest displacements related to structural changes made upon the nonapeptide hormone-binding leading to small alterations across the dimerization interface [55]. A99 (A38 in 1NPO) displacements in a mutant that displays weak neurophysin dimerization have also been described in an NMR study [19].

This Community can be generally described as unspecific for any taxonomic class, albeit lower frequencies are seen for Gastropoda and Insecta for positions conserved in Aves, Mammalia and Actinopteri (Supplementary materials Figure S2 – Community 1).

The other nine communities are composed of residues with mostly low frequencies when seen alone in the multiple sequence alignment but high co-occurrence values (Supplementary materials Table S1).

Community 4 includes a pair of residues, A490 located in the signal peptide region and Q664 located in the hormone nonapeptide ("cyiQncpig" in human oxytocin). Q664 is mutually exclusive to S664 in Community 10 and A490 to L490 in Community 9 (Supplementary materials Figure S2).

Community 9 also contains Q501 which is mutually exclusive to S501 in Community 1 in the signal peptide region. The remaining residue in Community 10 is C500 with no mutual exclusions but also located in the signal peptide (Supplementary materials Figure S2).

Examining the taxonomic profile of residue frequencies for these communities yields that Community 4 is seen highly frequent in Mammalia and Aves and infrequent in Gastropoda and Insecta. Actinopteri show frequency values for this Community close to its overall frequency in the alignment. In contrast, Community 9 is highly specific for Gastropoda and infrequent in Mammalia, Aves, Insecta and Actinopteri, reflecting unique signal peptide characteristics often found in this clade for this protein family.

Community 10 is highly specific to Actinopteri and has low amino acid frequencies for all other clades. (Supplementary materials Figure S3). Among different clades, sequences of the nonapeptides of the oxytocin/vasopressin protein family present marked differences with the Serine residue in position 664 of the multiple sequence alignment reflecting the sequence of Isotocin, which replaces along with vasotocin the mammalian oxytocin and vasopressin in bony fishes. The Glutamine residue of Community 4 in this position is found in all other major vertebrate homologs in this protein family such as Vasopressin/Lysipressin/Phenypressin, Vasotocin, Oxytocin and Mesotocin [56,53]. Substitution of this Glutamine by an Arginine in the Conopressin of Conus miliaris leads to the low activity of this peptide with human and zebrafish Oxytocin and Vasopressin receptors [21].

Mutually exclusive correlations between Communities 9 and 10 to Communities 1 and 4 probably reflect evolutionary divergences for the signal peptide and nonapeptide hormone region in this protein family found between derived and basal clades in vertebrates. The previous mutation of position 664 suggests that these divergences might reflect preferential receptor binding from different classes of Oxytocin/Vasopressin neuropeptides specific to each of these clades.

Communities 2, 3, 6 and 8 are all located in the C-terminal Copeptin peptide cleaved from Vasopressin genes and absent in Oxytocin genes. The function of Copeptin is still unknown but has been proposed to be related to the appropriate folding of Vasopressin Neurophysin-II during precursor processing.

Taxonomically, residues from these communities have their frequencies enriched only in mammals, with the exception of Community 3, which is also found more frequently in Aves (Supplementary materials Figure S3).

Community 6 corresponds to the N-Glycosylated Asparagine residue of this peptide and an Arginine residue, which is the cleavage site between the Neurophysin-II portion of the synthesized precursor and Copeptin [7]. As such, this correlation might reflect the two main characteristics of Copeptin, possibly directly related to its function. Exploring the effect of mutating such positions might shed light on the biological role of Copeptin, which is still elusive up to this point. Since no structures and few mutagenesis studies are available for Copeptins, functions possibly exerted by the amino acids of communities 2, 3 and 8 are unknown.

Community 5 contains S873 that was observed to make important dimerization contacts in Neurophysins T69 (T38 in 1NPO structure) [55]. S873 is correlated to G850 whose function is also still elusive but whose position in the Neurophysin-I structure was noted to be part of a β-turn and might be structurally relevant in these proteins [19]. This Community is particularly prevalent in Actinopteri (Supplementary materials Figure S3).

Community 7 is composed of a pair of Valine residues (V956 and V833) corresponding to Community 4 of the multiple sequence alignment containing exclusively the PF00184 domain. This Community has a high prevalence in the Gastropoda class and is mutually exclusive to two highly frequent residues in Community 1 (A956 and F833-Supplementary materials Figure S2). A956 corresponds to the A99 residue in Bovine Neurophysin I (A68 in 1NPO) for which a mutation related to Neurohypophyseal Diabetes insipidus has been described. This position has also been described for making important hydrogen bonding interactions at the dimerization interface of Neurophysins [55].

Despite the elusive function of residues in Community 5 and 7, previous structural and disease-related variant studies, the clade specificity of these communities and the role of some positions in the dimerization interface of Neurophysins could reflect the evolution of dimer contacts in different clades for this protein family. This appears to be the case especially for Community 7, mutually exclusive to positions in Community 1 and found in both multiple sequence alignments. Future studies involving mutagenesis of these positions are required in order to clarify their possible importance.

Phylogenetic model analysis

Due to the high conservation of these sequences at both the protein and nucleotide level, poor bootstraps support values were obtained despite various approaches including varying substitution models and column filtering procedures. Despite this, separation of these proteins into different phylogenetic classes in a similar manner to their taxonomic classification could be coarsely observed such as a grouping of Conopressins, Mesotocins, Oxytocins and Vasopressins in different branches of the obtained tree (Supplementary materials Table S4).

Amino acid coevolution and prediction of taxonomic classes

In order to identify particularities about the amino acid sites, which could explain if coevolution signals from the residues correlation results could determine the formation of different groups within the taxonomic classes, machine-learning methods were applied to correlate amino sites from the four communities in the PF00184 multiple sequence alignment. Between the three different algorithms used, the K-Nearest neighbors (Knn) showed the best hits, both for taxonomic classes prediction using only the correlated amino acids and for validation values. Our preliminary results about the predictions made by the algorithms, more specifically the K-nn, showed that the amino acids highlighted as correlated by the tests made later could predict the main classes. Table 1 shows the percentage of values obtained by the hits and the valid values for the main classes.

Table 1: Percentage of hits prediction of ML algorithms.

Class	Percentage	F1	CA	AUC
Actinopterygii	98%			
Aves	86%			
Mammals	97%			
Reptilia	16%			
Amphibia	100%	0.83	0.86	0.9
Cephalopoda	100%			
Chondrichthyes	100%			
Gastropods	85%			
Insecta	33%			

Using Multidimensional Scale (MDS) methods that can project items on a plane adjusted distances between points, it was possible to analyze with greater clarity how residue coevolution can influence taxonomic differentiation within a family of proteins (Figure 4). The results obtained by the analysis of MDS clearly show that they can use correlated amino acids to classify the different groups. In figure 4, it is possible to see that a large part of Actinopterygii sequences, in red, are interconnected with the Chondrichthyes sequences. It makes sense in an evolutionary context to note that both sequences are shown to be close in the results of prediction by machine learning algorithms since both belong to the fish superclass. The set of amino acid groups capable of determining the class of Aves shows that some reptilian sequences are also close to each other. Aves are an evolutionarily related group to the Crocodilia order since they represent the only surviving members of a "reptilian" clade, the Archosauria (Figure 5).

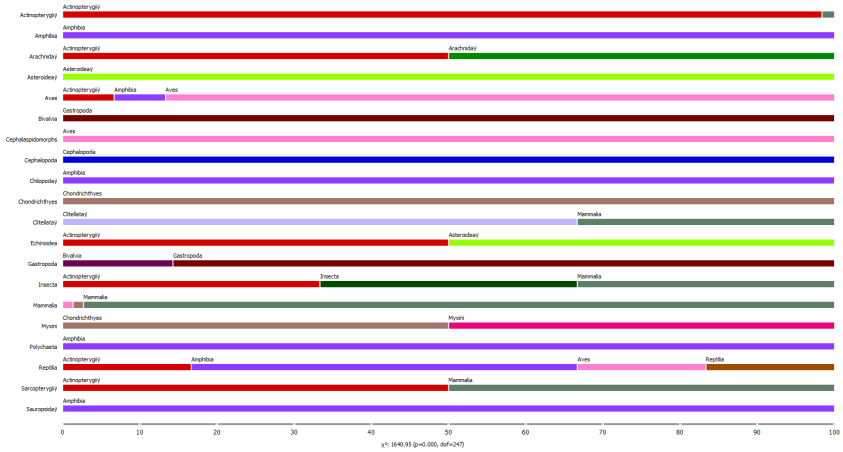

Figure 4: A simplified and summarized representation of boxplot graphs of the total samples predicted by the algorithms and their margin of accuracy.

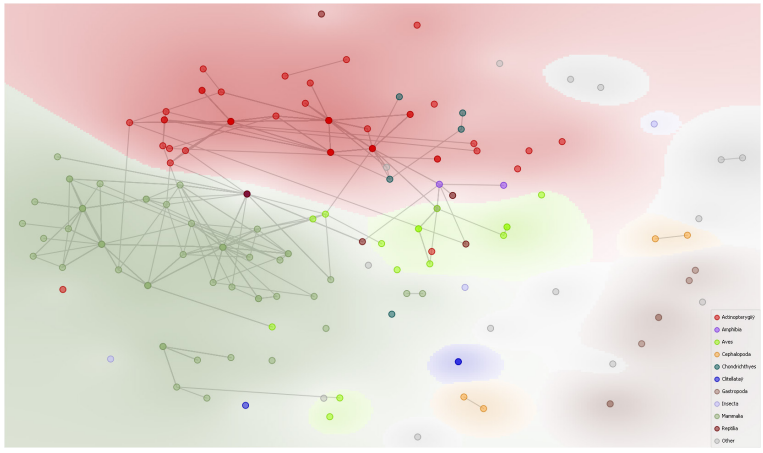

Figure 5: Algorithm's capabilities in predicting distinguishing sites, between the groups of vertebrates and invertebrates.

Near these two, the mammalian class can be found. These groupings reflect the algorithm's capabilities in predicting distinguishing sites, between the groups of vertebrates and invertebrates from the single use of the coevolving amino acids. When analyzing this figure, it is possible to see that such groupings do not coarsely mix. As seen for example, in invertebrates such as Gastropods, and Cephalopods. This demonstrates that analyzing co-evolving amino acid sites can be a relevant way to understand evolutionary convergences and divergences between biopolymers and their relation to taxonomic diversity, contributing as an aid for a better understanding of the evolutionary information contained within clades.

Balance selection evidenced by neutrality tests

In order to assess whether the correlated amino acid sites could influence different taxonomic categories, we initiated some statistical approaches to determine variables such as nucleotide segregation, frequency and the presence of polymorphisms. For this, neutrality tests have been proposed, such as Tajima D [57] and Mixed Effects Model of Evolution (MEME) [58]. Our results point to values that fit into interpretations of sudden population contraction, which corresponds to proper balancing between the sequences. Such data inform that the correlated amino acids identified are strongly involved in the selection of the different types of protein taxonomic classes, confirming that such amino acids are key in the protein prediction capacity.

CONCLUSION

Our results point out that coevolved amino acid sites reflect conserved features in the neurohypophyseal hormone protein family, including the evolutionary characteristics of different classes and nonapeptide sequences. This expands the previous knowledge of molecular evolution between sequences and structural features for this protein family, since these sites display-varying frequencies between sequences of different taxonomic groups.

Machine learning methods have highlighted that correlated amino acids may be key in predicting different proteins from their taxonomic classes. Such information brings more detail into the relationships between taxon classification and molecular evolution. Our results also reflect protein cleavage variations between vertebrate and invertebrate groups.

Since studies involving evolutionary concepts for the neurohypophysial hormones protein family are scarce, we hope this study is a useful addition capable of highlighting the importance and necessity of more future studies regarding this topic.

The results obtained in this work can serve as a basis for a progressive understanding of evolutionary convergences and divergences in this group of biomolecules, which possess potential in the medical, veterinary and biotechnological areas.

ACKNOWLEDGMENTS

The author would like to thank Prof. Noel Mendonça for the scholarship provided by the project "machine learning algorithms to detect disease risk factors" and the space provided by UBImedical, affiliated to the Universidade da Beira Interior.

REFERENCES

1. Gruber CW. Physiology of invertebrate oxytocin and vasopressin neuropeptides. Exp Physiol. 2014; 99(1): 55-61.
2. Muratspahić E, Monjon E, Duerrauer L, Rogers SM, Cullen DA, Broeck JV, et al. Oxytocin/vasopressin-like neuropeptide signaling in insects. Vitam Horm. 2020;113: 29-53.
3. Wacker D, Ludwig M. The role of vasopressin in olfactory and visual processing. Cell Tissue Res. 2019; 375(1): 201-15.
4. Odekunle EA, Elphick MR. Comparative and evolutionary physiology of vasopressin/oxytocin-type neuropeptide signaling in invertebrates. Front Endocrinol. 2020; 11: 225.
5. Kawakami N, Otubo A, Maejima S, Talukder AH, Satoh K, Oti T, et al. Variation of pro-vasopressin processing in parvocellular and magnocellular neurons in the paraventricular nucleus of the hypothalamus: Evidence from the vasopressin-related glycopeptide copeptin. J Comp Neurol. 2020.
6. Barat C, Simpson L, Breslow E. Properties of human vasopressin precursor constructs: Inefficient monomer folding in the absence of copeptin as a potential contributor to diabetes insipidus. Biochem. 2004; 43(25): 8191-8203.
7. Christ-Crain M. Vasopressin and Copeptin in health and disease. Rev Endocr Metab Disord. 2019; 20(3): 283-294.
8. Gwee PC, Amemiya CT, Brenner S, Venkatesh B. Sequence and organization of coelacanth neurohypophysial hormone genes: Evolutionary history of the vertebrate neurohypophysial hormone gene locus. BMC Evol Biol. 2008; 8(1): 1-2.
9. Rougon-Rapuzzi G, Cau P, Boudier JA, Cupo A. Evolution of vasopressin levels in the hypothalamo-posthypophysial system of the rat during rehydration following water deprivation. Neuroendocrinol. 1978; 27(1-2): 46-62.
10. Van Kesteren RE, Smit AB, De Lange RP, Kits KS, Van Golen FA, Van Der Schors RC, et al. Structural and functional evolution of the vasopressin/oxytocin superfamily: vasopressin-related conopressin is the only member present in Lymnaea, and is involved in the control of sexual behavior. J Neurosci. 1995; 15(9): 5989-5998.
11. Van Kesteren RE, Tensen CP, Smit AB, van Minnen J, Kolakowski Jr LF, Meyerhof W, et al. Co-evolution of ligand-receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. J Biol Chem. 1996; 271(7): 3619-3626.
12. Wallis M. Molecular evolution of the neurohypophysial hormone precursors in mammals: Comparative genomics reveals novel mammalian oxytocin and vasopressin analogues. Gen Comp Endocrinol. 2012; 179(2): 313-318.
13. Liutkeviciute Z, Koehbach J, Eder T, Gil-Mansilla E, Gruber CW. Global map of oxytocin/vasopressin-like neuropeptide signalling in insects. Sci Rep. 2016; 6(1): 1-9.
14. Gorson J, Ramrattan G, Verdes A, Wright EM, Kantor Y, Rajaram Srinivasan R, et al. Molecular diversity and gene evolution of the venom arsenal of terebridae predatory marine snails. Genome Biol Evol. 2015; 7(6): 1761-1778.
15. Feldman R, Monakhov M, Pratt M, Ebstein RP. Oxytocin pathway genes: Evolutionary ancient system impacting on human affiliation, sociality, and psychopathology. Biol Psychiatry. 2016; 79(3): 174-184.
16. Lockard MA, Ebert MS, Bargmann CI. Oxytocin mediated behavior in invertebrates: An evolutionary perspective. Dev Neurobiol. 2017; 77(2): 128-142.

17. Paré P, Paixão-Côrtes VR, Tovo-Rodrigues L, Vargas-Pinilla P, Viscardi LH, Salzano FM, et al. Oxytocin and arginine vasopressin receptor evolution: Implications for adaptive novelties in placental mammals. Genet Mol Biol. 2016; 39(4): 646-657.

18. Chen LQ, Rose JP, Breslow E, Yang D, Chang WR, Furey WF, et al. Crystal structure of a bovine neurophysin II dipeptide complex at 2.8 A determined from the single-wavelength anomalous scattering signal of an incorporated iodine atom. Proc Natl Acad Sci. 1991; 88(10): 4240-4244.

19. Nguyen TL, Breslow E. NMR analysis of the monomeric form of a mutant unliganded bovine neurophysin: Comparison with the crystal structure of a neurophysin dimer. Biochem. 2002; 741(18): 5920-5930.

20. Rose JP, Wu CK, Hsiao CD, Breslow E, Wang BC. Crystal structure of the neurophysin—oxytocin complex. Nat Struct Biol. 1996; 3(2): 163-169.

21. Giribaldi J, Ragnarsson L, Pujante T, Enjalbal C, Wilson D, Daly NL, et al. Synthesis, pharmacological and structural characterization of novel conopressins from Conus miliaris. Mar Drugs. 2020; 18(3): 150.

22. Haensele E, Saleh N, Read CM, Banting L, Whitley DC, Clark T. Can simulations and modeling decipher NMR data for conformational equilibria? Arginine–Vasopressin. J Chem Inf Model. 2016; 56(9): 1798-1807.

23. Kulkarni AK, Ojha RP. Conformations of a model cyclic hexapeptide, CYIQNC: 1H-NMR and molecular dynamics studies. J Biomol Struct Dyn. 2015; 33(9): 1850-1865.

24. Lubecka EA, Sikorska E, Marcinkowska A, Ciarkowski J. Conformational studies of neurohypophyseal hormones analogues with glycoconjugates by NMR spectroscopy. J Pept Sci. 2014; 20(6): 406-414.

25. Sikorska E, Rodziewicz-MotowidŁo S. Conformational studies of vasopressin and mesotocin using NMR spectroscopy and molecular modelling methods. Part I: Studies in water. J Pept Sci. 2008; 14(1): 76-84.

26. Yedvabny E, Nerenberg PS, So C, Head-Gordon T. Disordered structural ensembles of vasopressin and oxytocin and their mutants. J Phys Chem B. 2015; 119(3): 896-905.

27. Gwee PC, Tay BH, Brenner S, Venkatesh B. Characterization of the neurohypophysial hormone gene loci in elephant shark and the Japanese lamprey: Origin of the vertebrate neurohypophysial hormone genes. BMC Evol Biol. 2009; 9(1): 1-5.

28. Robinson SD, Li Q, Bandyopadhyay PK, Gajewiak J, Yandell M, Papenfuss AT, et al. Hormone-like peptides in the venoms of marine cone snails. Gen Comp Endocrinol. 2017; 244: 11-18.

29. da Fonseca Jr NJ, Afonso MQ, de Oliveira LC, Bleicher L. A new method bridging graph theory and residue co-evolutionary networks for specificity determinant positions detection. Bioinformatics. 2019; 35(9): 1478-1485.

30. Oliveira A, Teixeira P, Barh D, Ghosh P, Azevedo V. Key Amino Acids in Understanding Evolutionary Characterization of Mn/Fe-Superoxide Dismutase: A Phylogenetic and Structural Analysis of Proteins from Corynebacterium and Host. Trends Artif Intell. 2017; 1(1): 1–11.

31. Oliveira A, Bleicher L, Schrago CG, Junior FP. Conservation analysis and decomposition of residue correlation networks in the phospholipase A2 superfamily (PLA2s): Insights into the structure-function relationships of snake venom toxins. Toxicon. 2018; 146: 50-60.

32. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nat Methods. 2019; 16(8): 687-694.

33. Iakovlev D, Kobchenko A, Semina E. Protein structural domain prediction via machine learning approach. In Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS\SB-2018) 2018; 104-104.

34. Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, et al. SVM-Prot 2016: A web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. PloS one. 2016; 11(8): 1-14.

35. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: The protein families database. Nucleic Acids Res. 2014; 42.

36. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013; 30(4): 772-780.

37. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES science gateway for inference of large phylogenetic trees. 2010 Gateway Computing Environments Workshop (GCE). 2010; 1–8.

38. Fonseca NJ, Afonso MQ, Carrijo L, Bleicher L. CONAN: A web application to detect specificity determinants and functional sites by amino acids co-variation network analysis. Bioinformatics. 2020.

39. Tumminello M, Micciche S, Lillo F, Piilo J, Mantegna RN. Statistically validated networks in bipartite complex systems. PloS one. 2011; 6(3).

40. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013; 30(12): 2725-2729.

41. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Bio Evol. 2010; 27(2): 221-224.

42. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online: A web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 2005; 33: 557-559.

43. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23(21): 2947-2948.

44. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 1985; 39(4):783-791.

45. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006; 55(4): 539-552.

46. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000; 155(3): 1429-1437.

47. Suchard MA, Rambaut A. Many-core algorithms for statistical phylogenetics. Bioinformatics. 2009; 25(11): 1370-1376.

48. Janez D, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, et al. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research. 2013;14: 2349-2353.

49. Dutertre S, Croker D, Daly NL, Andersson A, Muttenthaler M, Lumsden NG, et al. Conopressin-T from Conus tulipa reveals an antagonist switch in vasopressin-like peptides. J Biol Chem. 2008; 283(11): 7100-7108.

50. Kumar S, Vijayasarathy M, Venkatesha MA, Sunita P, Balaram P. Cone snail analogs of the pituitary hormones oxytocin/vasopressin and their carrier protein neurophysin. Proteomic and transcriptomic identification of conopressins and conophysins. Biochim Biophys Acta Proteins Proteom. 2020; 1868(5):140391.

51. De Bree FM, Burbach JP. Structure–function relationships of the vasopressin prohormone domains. Cell Mol Neurobiol. 1998; 18(2):173-191.

52. Miller WL. Molecular genetics of familial central diabetes insipidus. J Clin Endocrinol Metab 1993; 77(3): 592-595.

53. Donaldson ZR, Young LJ. Oxytocin, vasopressin, and the neurogenetics of sociality. Science. 2008; 322(5903):900-904.

54. Grinevich V, Desarménien MG, Chini B, Tauber M, Muscatelli F. Ontogenesis of oxytocin pathways in the mammalian brain: Late maturation and psychosocial disorders. Front Neuroanat. 2015; 8: 164.

55. Wu CK, Hu B, Rose JP, Liu ZJ, Nguyen TL, Zheng C et al. Structures of an unliganded neurophysin and its vasopressin complex: Implications for binding and allosteric mechanisms. Protein Sci. 2001; 10(9): 1869-1880.

56. Acher, R. Neurohypophysial Hormone Regulatory Systems. In Encyclopedia of Endocrine Diseases. 2004; 314–329.

57. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123(3):585-595.

58. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SL, et al. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012; 8(7).