Bioinformatics for prohormone and neuropeptide discovery

Bruce R. Southey,[2] Elena V. Romanova,[1] Sandra L. Rodriguez-Zas,[2] and Jonathan V. Sweedler[1,*]

[1]Department of Chemistry and Beckman Institute for Advanced Science and Technology, and [2]Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, U.S.A.

*Corresponding author: jsweedle@illinois.edu

Abstract

Neuropeptides and peptide hormones are signaling molecules produced via complex post-translational modifications of precursor proteins known as prohormones. Neuropeptides activate specific receptors and are associated with the regulation of physiological systems and behaviors. The identification of prohormones--and the neuropeptides created by these prohormones--from genomic assemblies has become essential to support the annotation and use of the rapidly growing number of sequenced genomes. Here we describe a methodology for identifying the prohormone complement from genomic assemblies that employs widely available public toolsets and databases. The uncovered prohormone sequences can then be screened for putative neuropeptides to enable accurate proteomic discovery and validation.

Key Words: Neuropeptide, Prohormone, Homology, Bioinformatics, Cleavage, Gene prediction

1 Introduction

The increased speed and decreased cost of genomic sequencing has revolutionized neuropeptide identification from primarily experiment-driven discovery to bioinformatics-driven genomic searches. Currently, the number of sequenced genomes exceeds the number of species that have at least one experimentally confirmed prohormone and neuropeptide. The advances in genomics also enable neuropeptide discovery to extend beyond the specific experimental system being investigated, such as a defined tissue or developmental stage. We have developed a bioinformatics toolset that can be used to uncover neuropeptides in a broad range of organisms, and to predict multiple putative novel neuropeptides that have not been previously characterized, even in well-studied species.

The challenge in using genomic data to discover neuropeptides stems from the fact that one neuropeptide gene encodes multiple neuropeptides within a larger precursor protein, or prohormone. Prohormones undergo post-translational enzymatic cleavage and further chemical modifications, resulting in a set of shorter neuropeptides [1], all products of the same gene. The bioinformatics approach to neuropeptide detection follows this biological process, starting with the identification of the prohormone sequence in the genome. The characteristic features of a complete prohormone protein are: (1) a signal peptide at the N-terminus that enables the co-translational translocation of the protein into the secretory pathway [2], and (2) at least one cleavage site that is recognized by endopeptidases that delimits the neuropeptide sequence.

Our efforts to identify prohormones and neuropeptides across taxa and species (e.g., human, mouse, rat, cattle, pig, chicken, song bird, honey bee, flour beetle, California sea slug, fish, camelids [3-10]) have enabled us to identify the bioinformatics steps that are critical for the accurate identification and annotation of prohormones in genomes. These steps provide reliable prohormone gene, protein sequence, and neuropeptide prediction across sequenced genomes, regardless of the degree of experimental validation. Our web application NeuroPred [11,12] expedites the bioinformatics tasks and has been successfully used to predict the putative neuropeptides nested in these prohormone sequences [7-9,11,13]. We have combined the sequence annotations obtained using NeuroPred with other bioinformatics tools into the web portal PepShop [14] to facilitate neuropeptide and prohormone discovery and usage in a wide range of species. Together they provide important resources for the neuropeptide research community.

A major undertaking of the bioinformatics approach for detecting neuropeptides is the identification of candidate prohormones in a target species. The availability of the species genome is not sufficient by itself for accurate neuropeptide identification. Many prohormone genes are predicted using automatized methods to annotate the genome assembly. However, we have demonstrated that some of these prohormone gene predictions encompass inaccurate features due to incorrect exon predictions, and some prohormones have been completely missed. In addition, few neuropeptides from predicted proteins have been manually or empirically confirmed. Accurate prohormone gene prediction also benefits from comparative genomics and recognition of sequence similarities between gene structures of evolutionarily related species because neuropeptides are highly conserved across the metazoans [15].

The prohormone and peptide sequences predicted by our bioinformatics approach (**Fig. 1**) are readily available to support the identification of novel peptide sequences that have been experimentally detected. The symbiotic integration of bioinformatics predictions with robust mass spectrometry (MS) sequencing enables identification of candidate and novel peptides [16]. Sensitive and specific identification of peptides and their PTMs are essential to the advancement of neuropeptide research.

We demonstrated the benefit of integrating our biologically driven informatics approach with MS analytics to accurately predict prohormones and neuropeptides in the particularly challenging genome of *Astatotilapia burtoni*, a teleost fish [4]. The evolutionary ancestor of *A. burtoni* underwent a whole genome duplication event after the divergence of tetrapods and teleosts [17]. Our systematic bioinformatics approach proved to be well-suited for sifting through sequence duplications and rearrangements to harvest useful sequence information from divergent species. Throughout this chapter, we provide examples of our approach by describing the individual steps we used in that work [4] to accurately identify and annotate the *A. burtoni* parathyroid hormone family.

2 Materials

2.1 Genome assembly of desired species.

The bioinformatic identification of prohormone genes integrates genomic, transcriptomic, and proteomic information from the genome repository of the target species. The genomic data include the nucleic assembly, which is typically composed of assembled chromosomes, and scaffolds (genomic sequence interspersed with sections of unknown nucleotides) or contigs (contiguous lengths of known genomic sequence). Transcriptomic data include RNA and expressed sequence tag (EST) sequences, and proteomic data that mainly includes sequences predicted by automated tools that are used to annotate the genome assembly. The National Center for Biotechnology Information (NCBI) [18] and ENSEMBL [19] are examples of public repositories that archive genome assemblies and genome, transcriptome, and protein sequences. These public repositories are regularly updated to include the most updated revisions of the genome assemblies and sequence predictions.

2.2 Prohormone protein sequences of phylogenetically close species.

Evolutionary relationships are shaped in part by molecular similarities at both the DNA and protein levels. Therefore, the prohormone protein sequences of the phylogenetically related reference species are expected to be structurally and functionally similar to the target species. Many of these sequences can also be retrieved from the Universal Protein Resource (UniProt) [20] and NCBI RefSeq [21] databases, albeit in a less streamlined manner since these databases are not dedicated to prohormone sequences. These online resources serve for initial prohormone annotation in the target species using cross-species homology. We have built libraries of annotated prohormones for a wide variety of species, including multiple mammalian species, insects, birds, and mollusks, available at the NeuroPred [22] and PepShop [14] websites (*see* **Note 1**). There are other online databases [23-25] that also include prohormone information.

2.3 Bioinformatic tools.

Essential tools for finding cross-species sequence similarities include routines and software for sequence alignment. Foremost examples of these tools are the Best Linear Alignment Software Tool (BLAST) [26] for pair-wise alignment of a sequence to database sequences, CLUSTAL Omega [27] for multiple sequence alignment, and GeneWise [28] for gene prediction from nucleotide sequences. These web-based tools are highlighted because they are regularly updated, do not require local installation, and outputs can easily be shared, reproduced, and replicated. Additional desirable features of bioinformatic toolsets are enhanced functionality that simplify the steps and allow easy transfer of the output from one program to serve as input for another program. The bioinformatics demonstration described below uses NCBI databases and tools because of their comprehensive scope and ease of integration. These resources will be complemented with CLUSTAL Omega and GeneWise.

2.4 Spreadsheet and text editor applications to record findings.

Gene/protein databases provide important information, notably prohormone gene names, gene locations, and complementary database accession number identifiers that need to be stored for each prohormone gene. The application selected to hold this sequence information must support the addition or update of fundamental information and addendum notes, multifactorial search, and reordering of current information, all within a table format. These features enable the rapid discovery of previously located prohormone genes and entry of new genes and similar genes. Most open-source or commercial spreadsheet applications (e.g., LibreOffice calc and Microsoft Excel) provide these abilities and offer satisfactory functions, including adding new columns and sorting by features such start and end positions of sequence matches or alignments to different genome regions (*see* **Note 2**). A text editor (such as LibreOffice writer or Microsoft Word) primarily enables the accumulation of protein sequences in a table format that can be used for other programs or searched for specific short sequences.

3 Methods

The first step in neuropeptide discovery is the identification of candidate prohormone genes and retrieval of the corresponding protein sequences. Next, thorough confirmation of these sequences is required via additional resources, such as other publically available sequences from desired species and trace archives (raw, short sequences generated by a specific genome project) and comparison of sequences across related species. Finally, the signal peptide and putative neuropeptides contained in the predicted prohormone sequence are calculated. A detailed description of these bioinformatic steps follows.

3.1 Create a list of putative prohormones.

Select the phylogenetically closest species that has extensive characterization of the prohormone complement as the reference species. Use this reference species to generate an initial list of candidate prohormone that will be searched in the target species (*see* **Note 3**).

Arrange the prohormones from the reference species in the rows of a spreadsheet application table. The columns of this table will record the sequence accession numbers of the reference and target species, information on the target genome, including location of the reference sequence matches in the target genomic assembly, and miscellaneous details, e.g., the completeness of the reference sequence in the target genome assembly.

Each candidate in the list of prohormones from the reference species is evaluated in the target species following the bioinformatics steps shown in Fig. 1. Keeping in mind that there is not a single ideal approach to use when evaluating the candidate list, the candidate prohormones are processed individually. Searches can benefit from concurrent evaluation of all the sequences within a prohormone family because of the high sequence similarity among prohormone family members. This sequential processing of the candidate list should be amended to accommodate new putative prohormones that are identified in subsequent bioinformatics steps. Novel prohormones identified in the target species and incorporated in the candidate list should be searched in the reference sequence in an iterative manner.

Putative matches to the candidate sequences identified in the target species must be recursively evaluated before moving to the next candidate prohormone in the list. Information on each putative match, such as gene(s) identifier, genome location(s), significance threshold of detection (e.g., e-value, percentage identity), and comments are added to the table (*see* **Note 4**). The evaluation of the best match (i.e., lower e-value) between the candidate and target sequences must be followed by the evaluation of weaker matches (within 25% of the best match). This extended investigation permits identification of prohormones in the target species that originated from duplication events. Additional matches identified in the target species must be included in the candidate list, and the table entry should be completed with any new information. A candidate prohormone sequence is considered complete once all of the bioinformatics steps in Fig. 1 have been successfully completed.

Some candidate prohormones from the reference species will not have matches in the target species. An example of this is the mammalian relaxin 1 gene that has been lost in ruminants [6]. In other cases the lack of a candidate on the target genome may be associated with reference-specific duplication events after the branching of the reference and target species. An example of this is the insulin-like 4 gene that resulted from a primate-only gene duplication event [29]. Any candidates without a match should be denoted as not being present before moving on to the next candidate in the list.

Our bioinformatics approach was used to identify the *A. burtoni* members of the parathyroid hormone family. Consistent with the *Homo sapiens* (human) parathyroid family [30], the fish parathyroid hormone family fish contains at least three members: parathyroid hormone 1 (PTH1), parathyroid hormone 2 (PTH2), and parathyroid hormone-related protein (PTHLH) [31,32]. Therefore, the initial candidate list for *A. burtoni* had 3 entries: PTH1, PTH2, and PTHLH. This list does not account for the teleost whole-genome duplication because it is uncertain if any of the duplicated prohormone genes resulting from whole-genome duplication have been retained in *A. burtoni*.

3.2 Identification of putative prohormones.

3.2.1 Text search for prohormone and neuropeptide names.

Searches for candidate prohormones include text searches for names and gene symbols of the prohormone genes that have already been annotated in the target species. The existing annotation could be the output of the target genome project or other research efforts (e.g., evolutionary studies of specific prohormone genes). A text search in a gene-centric database, such as NCBI Gene [33], is recommended as the first step rather than a search in a protein database because gene databases are typically more complete and maintain more information at both the nucleotide and amino acid levels than protein databases.

Effective text searches use the common prohormone name or neuropeptide, followed by one or more unique words that are part of the prohormone or neuropeptide name (*see* **Note 5**). Text matches are then recorded in the candidate prohormone table. These details may expedite the entire process since some of this information is identified in the subsequent steps and can assist in the discrimination between duplicated genes and homologous genes from the same prohormone family in the target species.

In our example, the first candidate is parathyroid hormone, so a text search for parathyroid hormone in *A. burtoni* was conducted within the NCBI Gene database using the phrase: "parathyroid hormone"[title] AND ("Astatotilapia burtoni"[orgn]). This phrase uses the '[title]' and '[orgn]' options to limit the search for specific words in the descriptive text, such as gene name and species, respectively. In addition to finding PTH1, this text search also identified the other genes, PTH2 and PTHL, because the phrase "parathyroid hormone" is part of the gene names, as well as part of the name of the prohormone family.

The outcome of this search included 3 prohormone genes and 3 prohormone receptors (Fig. 2), and is summarized in Table 1. An equivalent search within the NCBI protein database [34] only identified 2 proteins. A more general search for the phrase: "parathyroid"[title] AND ("Astatotilapia burtoni"[orgn]) identified 6 proteins corresponding to 3 receptor proteins, PTH1, and the 2 PTHLH protein isoforms. PTH2 was not identified in the previous text search because the database entry referred to a former name, 'tuberoinfundibular peptide of 39 residue' (TIP39).

3.2.2 Homology search against protein databases and genome assembly databases.

The initial search of a protein sequence in a protein database uses BLASTP, a version of BLAST that searches protein sequence databases using a protein sequence as a query. The candidate prohormone protein sequence from the reference species is searched against the protein database of the target species. Consideration of BLASTP matches that have an e-value ≤ 10 using the default BLOSSUM 62 substitution matrix enables the identification of matches that have conserved, albeit small, regions across species. Effective BLASTP searches disable filtering and masking of low-complexity regions to favor the visualization of long sequence alignments. BLASTP searches using a less significant (higher) e-value threshold, as well as substitution matrices that support more distance sequence matches (e.g., BLOSSUM 50), should be investigated when the original search fails to identify matches.

All BLASTP matches that meet the e-value cutoff and have at least one conserved region should be recorded in the table of candidate prohormones, including locations and e-values. Each BLASTP match should be referenced to the candidate list, as all of the entries identified by the text search should also be identified from this sequence similarity search. These existing entries should be updated with the BLASTP match details.

Matches to the candidate prohormone protein sequence could correspond to predicted splice variants resulting from alternative splicing, duplicated genes, other members from the same prohormone family, and other homologs (*see* **Note 6**). Information from the text and prior searches enables the initial differentiation of these redundant matches and elimination of matches already evaluated, or matches to another member the same prohormone family. The best BLASTP match (lowest e-value) should be evaluated for completeness, even if the protein sequence was identified by the previous text search. This enables the identification of possible splice variants or correction of inaccurate protein sequences, such as those generated by automated methods in annotation of the genome assembly.

Each candidate prohormone protein sequence match must be aligned against the prohormone sequence in the reference species, and potentially other species, using tools such as CLUSTAL Omega (*see* **Note 7**). True matches are characterized by alignments that have similar lengths and regions of high amino acid identity or similarity. The information gained from the alignment of the best prohormone sequence match to sequences in multiple species must be added to the table of candidates. These details can include the additional species that contain matches, accession numbers, matching sequences, and the most complete and longest alignments between the candidate sequence and the sequences in other species. All candidate prohormone sequences that have at least one complete protein match that surpasses the e-value cutoff in another species must be marked as completed in the candidate prohormone table. Protein sequence matches supersede text matches and thus, a text match must be replaced with a protein match when available.

The remaining BLASTP matches that surpass the e-value cutoff (other than the best match) include those that contain a high degree of homology (likely duplicated genes), homologous genes from the same gene family, and new prohormone genes. All these matches should be collected into a single table of protein matches and then each entry evaluated one at a time. Entries that match existing complete entries in the candidate prohormone table, or are clearly not prohormone proteins, are removed from the table of protein matches. Entries that match any candidate prohormone or new prohormones should be evaluated in a similar manner to the best match. The resulting complete entries are used to update the table of candidate prohormones, including information on sequence and accession number, before the entry is removed from the table of protein matches.

Some of the remaining entries may be resolved with alternative searches that can result in the identification of a new prohormone that may be unique to the target species. Conducting a protein search in the protein sequences from another species can provide matches that can determine if the current entry is a match to a prohormone, or that can be used instead of the candidate protein sequence. Similarly, searches using the EST database may provide the gene identity or provide a larger region that aids in the resolution of the initial match. Any remaining entries in the table of protein matches must be resolved with genome searches.

In our example, a BLASTP search was initially conducted with the *H. sapiens* PTH1, PTH2, and PTHLH sequences because this species is not encumbered by the teleost whole genome duplication event. Our searches (results not shown) did not uncover additional information to add to our initial text search (Table 1).

Searches with *Danio rerio* (zebrafish) also did not uncover additional protein sequences. However, the *D. rerio* search results indicated that expected gene duplications resulting from the teleost whole genome duplication event were also not present in the *A. burtoni* protein sequences (results not shown).

Genome searches need to be conducted for any candidate prohormones that were not identified in the previous BLASTP searches and any remaining entries in the table of protein matches. A genome search should also be conducted if additional gene duplications are expected, e.g., from whole genome duplications or known tandem duplications in other species. The TBLASTN version of BLAST that searches a protein sequence against translated nucleotide sequences of the genome should be used with the same settings as the previous BLASTP search. The expected matching regions should comprise the exons of the candidate gene as well as exons from genes homologous to the candidate protein sequence. All of these matches should be collected into a single table of genome matches and any duplicate matches should be removed. Any entries that correspond to exons of identified prohormones should be removed, after ensuring that the genomic region only contains the identified prohormone gene.

The remaining entries in the genome table that are likely to be from the same gene should be grouped together. These groups should contain the genomic regions of at least one of the prohormone gene exons, and then evaluated one at a time. Each group must evaluated for assembly artifacts that must be resolved before continuing (*see* **Note 8**). Subsequently, the genome region of the group entry is expanded by ~ 2000 bp beyond the 5' and 3' ends of the group entry and the sequence from this extended region corresponding to the correct reading fame of the TBLASTN match is then extracted. This strategy maximizes the likelihood that the complete prohormone gene will be identified. A gene prediction program(e.g., GeneWise) is used to predict a protein sequence from this extracted genome region (*see* **Note 9**). If the predicted protein sequence is a new prohormone, then the table of candidate prohormones is updated accordingly. All entries in the genome table corresponding to genome group are then removed.

We performed genome searches for PTH1, PTH2, and PTHLH in the *A. burtoni* genome using TBLASTN (results not shown). All matches from the protein sequences of the 3 genes were entered as a single list of matches since the same match can occur with different members of the same prohormone family. Any duplicate entries were removed so that each match was only processed once. The first entry processed was a single match to PTH2, corresponding to the previously identified PTH2 gene (Table 1). This result indicated that there is only one copy in the genome, so this entry was removed from the list of matches.

The next entries processed were 2 matches to PTH1 (Fig. 3). The first of these entries matched the region of the NW_005179605.1 contig, which contains the previously identified PTH1 gene (XP_005936534.1; Table 1). This entry did not provide additional information and was removed from the list of matches. The second PTH1 entry corresponded to another contig, NW_005179496.1, and exhibited high similarity to the start of the second exon of PTH1. This discovery likely corresponds to the duplicated PTH1 gene resulting from the teleost whole genome duplication event.

To confirm that this match was a duplication of PTH1, we used the 'GenBank' link in NCBI output to retrieve the genome record and the matched sequence. The TBLASTN search had located a match in the complement strand, and a customized view was developed by selecting the reverse complement option and 'update view' in NCBI (Fig. 4). This region was expanded ~2,000 bp from the 5' and 3' ends and extracted. The GeneWise program was then used to predict a protein sequence from this region. The input (Fig. 5) included the *A. burtoni* PTH1 protein sequence (denoted as PTH1A) and the extracted *A. burtoni* genomic region. The output confirms our hypothesis of a duplicated PTH1 gene, providing the complete predicted gene sequence (Fig. 6), including an intron that is consistent with the known PTH1 gene. The previous PTH1 gene was renamed as PTH1A, and this new prediction denoted as PTH1B; the letters after the initial gene symbol (PTH1) are used to differentiate the different copies of the parathyroid hormone 1 gene. The protein sequence and information pertaining to PTH1B were added to candidate list and the entry was removed from the list of matches.

Our TBLASTN search of PTHLH in the target *A. burtoni* genome uncovered 5 matches with an e-value <10 (Fig. 7). The best match (lower e-value) was to the region of the NW_005179673.1 contig containing the previously identified PTHLH gene (XP_005939766.1; Table 1). This entry was removed from the list of matches because no additional information was obtained. The two matches with the highest e-values matched non-prohormone genes and these entries were discarded from the list of matches. The second-best match exhibited high similarity to the signal peptide region of PTHLH and the third-best match mapped to the start of the PTHLH sequence (Fig. 7). The locations of the second and third matches excluded the other known members of the parathyroid hormone family. This implied that these could be 3 *A. burtoni* PTHLH genes. Following the same approach previously used to obtain the PTH1B gene, augmented genome regions corresponding to these matches were extracted and the 2 different protein sequences were predicted. The initial PTHLH was renamed as PTHLH1 and the 2 predictions were denoted as PTHLH2 and PTHLH3 to differentiate the different PTHLH genes. After updating the candidate list and entering the proteins for these 3 PTHLH genes, these final entries were removed from the list of matches, marking the successful completion of the homology search step.

3.2.3 Novel detection based on neuropeptide motifs.

We recommend homology searches centered on conserved regions, such as neuropeptides, and using relatively lax criteria as helpful strategies to identifying novel prohormones. A more general approach is to search for neuropeptide motifs.

This approach has successfully led to the discovery of new prohormones [35,36]. However, searches for more generic motifs [37] or conserved regions inferred in silico using machine learning algorithms, such as hidden Markov models [38-40], have a high false-positive rate [41].

3.2.4 Validation of predicted prohormone protein sequences.

The predicted protein sequences must be validated for accuracy. This validation includes additional support from data not included in genome repository of the target species and information from other species. Homology searches for the prohormone protein sequences should be conducted in EST, RNA, and protein sequence databases that have not been incorporated into the genomic repository of the target species. These databases can also provide valuable complementary information, such as single nucleotide polymorphisms (SNPs), insertions, and deletions. Moreover, this complementary information can be helpful when the coverage of the target genome assembly varies across regions or is incomplete.

Multiple matches between the protein sequence from the target species and the EST sequences, including the full sequence or overlapping regions, offer the strongest empirical evidence for the gene. Gaps in the alignment could be attributed to alternative splicing and thus, all EST forms should be extracted and tested for sequence accuracy and completeness. When protein isoforms from the same prohormone are available, then all isoforms should be predicted from the same genomic region.

Confirmation using sequence information from non-genomic databases of other phylogenetically close species can be used to further validate the prohormone predictions. Matches to EST or other transcriptomic databases provide evidence of the presence of the predicted prohormone. However, the actual protein sequence cannot be completely verified since differences between sequences can be actual sequence differences between species, as well as a result of sequencing errors.

Our predicted *A. burtoni* PTH1B gene has no supporting confirmatory experimental evidence in any of the current *A. burtoni* resources. A suitable match in *Neolamprologus brichardi*, a species very closely related to *A. burtoni*, was found to the NCBI-predicted protein, "XP_006804843.1 PREDICTED: parathyroid hormone-like". This cross-species analysis offers further confirmation of our *A. burtoni* PTH1B prediction.

The new *A. burtoni* PTHLH2 and PTHLH3 prohormone protein sequences were validated using the NCBI *A. burtoni* transcriptomic-based databases. The PTHLH2 protein sequence had multiple matches in the NCBI *A. burtoni* 'Non-RefSeq RNA' database (this database contains RNA sequences that are present in GenBank but have not been included in the RefSeq database). One of these additional matches was an *A. burtoni* RNA sequence in the Transcriptome Shotgun Assembly (TSA) database [42] (titled "GBDH01011309.1 TSA: Haplochromis burtoni comp20762_c0_seq1 transcribed RNA sequence"). The

PTHLH3 protein sequence matched the *A. burtoni* EST, DY630955.1. Extracting and translating these sequences from each source confirmed the previously predicted PTHLH2 and PTHLH3 protein sequences. Recently PTHLH3 was identified in zebrafish (*D. rerio*) as parathyroid hormone 4 (PTH4; XP_005168342.1) and was shown to interact with the known parathyroid zebrafish hormone receptors [43].

3.3 Sequence verification of predicted prohormone proteins.

The comprehensive validation of the predicted prohormone in the target species enhances confidence in the detection. Further biological support can be gained using motif- or block-centered multiple sequence alignments. The alignments of the predicted prohormone sequence to databases of protein families using tools such as Pfam [44] further ensures that the target sequence is a member of the expected protein family. This additional validation step is particularly helpful when the candidate sequence or the validation information corresponds to a phylogenetically distant related species.

We recommend an additional bioinformatics step for accurate annotation of prohormone paralogs and orthologs in the target species, particularly in protein families that have a high level of homology. This step encompasses multiple sequence alignments of the candidate sequences and known prohormone sequences from multiple species, including phylogenetically distant species, using tools such as CLUSTAL Omega. In addition to discrimination between prohormone family members and identification of new duplicates in the target species, multiple sequence alignments support the assessment of expected structural features, such as conserved regions that are often characteristic for bioactive neuropeptides. For example, in our *A. burtoni* annotation [4], we confirmed the loss of melanocyte-stimulating hormone (MSH) peptide, γ-MSH [45].

To validate the accuracy of our parathyroid prohormone predictions, the 7 *A. burtoni* sequences were searched against the NCBI-predicted genes from 4 related cichlid species with sequenced genomes: *Oreochromis niloticus* (Nile tilapia), *N. brichardi, Pundamilia nyererei*, and *Maylandia zebra*. Matches to all of our parathyroid prohormone predictions were identified in multiple species (Table 2). Two PTH2 protein isoforms with corresponding mRNA evidence were detected in *O. niloticus*. Based on this evidence, both PTH2 protein isoforms were subsequently predicted from the same *A. burtoni* genomic region using GeneWise. Our bioinformatics approach uncovered 2 *A. burtoni* PTH2 isoforms that are the result from alternative splicing events.

Phylogenetic trees depicting the relationship between the sequences from the same prohormone or prohormone family across species can also be constructed using tools such as ETE3 [46]. These trees facilitate the identification of potential sources of prohormone sequence alignment errors, such as unusual or unexpected prediction, gaps, large mismatched regions, or misnamed predictions. Also, these phylogenetic trees provide insights into the evolutionary relationship between the sequences and can be used to discover novel taxa-specific prohormone genes.

All of the known sequences from the *A. burtoni*, *D. rerio*, and *H. sapiens* parathyroid prohormone family were aligned together using CLUSTAL Omega. A phylogenetic tree (Fig. 8) depicting the relationship between these parathyroid prohormone family sequences was constructed using the GenomeNet ETE 3 (v3.0.0b32) implementation [47]. This tree confirms that the sequences appear to be correctly named and illustrates the expected relationships between members of the parathyroid prohormone family. The branching in the PTHLH tree suggests 2 gene duplication events. The first is the teleost-specific whole genome duplication that resulted in PTHLH1 and PTHLH2; the second event occurred prior to the teleost-specific whole genome duplication and resulted in PTHLH3 (PTH4) and PTHLH. There is no *H. sapiens* PTHLH3 because PTHLH3 (PTH4) was lost in eutherian mammals after the eutherian-metatherian split [43]. These discoveries highlight the potential of our bioinformatic approach to annotate prohormones in target species, discover novel taxa-specific prohormone genes, and identify losses of prohormone genes.

3.4 Peptide prediction from prohormone protein sequences.

3.4.1 Signal peptide prediction.

Tools such as SignalP [48] can be used to predict the signal peptide and associated cleavage site in the N-terminal region of the predicted prohormone sequence in the target species (*see* **Note 10**). Rigorous application of the previous prohormone prediction steps guarantee the identification of a complete or nearly complete prohormone sequence. Incomplete protocol implementation or limited sequence information could produce partial, incorrect, or chimeric sequences. The predicted prohormone protein sequence must be revised when the signal peptide cleavage site cannot be identified. The sequence must also be revised when the location of the signal peptide cleavage site is predicted >40 amino acids from the prohormone initiation methionine site because this location is generally atypical.

3.4.2 Prediction of putative peptides.

The predicted prohormone sequence must be subsequently analyzed using NeuroPred. This public web service supports the prediction of putative peptide cleavage sites in a protein sequence, assessment of the likelihood of cleavage, and evaluation of potential PTMs. This resource accepts one or more protein sequences, and signal peptides are identified using default or user-defined specifications (*see* **Note 11**). NeuroPred encompasses multiple peptide cleavage prediction methods that have been optimized for multiple species (*see* **Note 12**). After removal of the signal peptide, NeuroPred predicts the cleavage sites and, depending on the selected options, identifies the resulting peptides. The peptide sequences, PTMs, and masses predicted by NeuroPred have been successfully used in support of neuropeptide identification in MS analyses.

We used NeuroPred to predict the peptides in the *A. burtoni* PTH1B protein sequence (Figs. 9 and 10). The NeuroPred input options depicted in Fig. 9 include the model used to predict the cleavage sites (the ‘Known Motif’ and ‘Mammalian’ cleavage models were selected for the *A. burtoni* example). The NeuroPred output depicted in Fig. 10 includes a cleavage diagram that localizes the signal peptide used and all predicted cleavage sites. The output also includes information on the peptides resulting from cleavage at the predicted sites.

4 Notes

1. The accuracy of the prohormone gene prediction increases exponentially with increasing sequence similarity between species studied. Thus, the species phylogenetically closest to the target species should be used in the first bioinformatics steps. Sequences from more phylogenetically distant species should be used to resolve contradictory findings among closest species and whenever substantial uncertainty arises from poor matches or unreliable predictions during the final bioinformatic steps.
2. The ability to identify the start and end locations of exons from multiple genome matches is critical to the discrimination between duplicated genes

located on the same chromosome or contigs in the target species. Furthermore, accurate exon delimitation enables the reduction of multiple matches to the same region by different genes, typically from the same gene family.

3. The nomenclature used to name a prohormone on a target species should follow community annotation guidelines. When these guidelines are not in place, the recommendation is to follow the guidelines of a major genome annotation project within the taxa. Accepted gene symbols should be used for prohormones to facilitate search and validation, especially across species, and to avoid confusion between proteins and peptides and between duplicated genes.
4. Characterization of prohormones resulting from tandem duplication requires recursive iteration of the bioinformatic steps described. An example of this scenario is mammalian calcitonin, which has varying copy numbers among mammalian species. Knowledge of potential tandem duplication aids in the accurate identification of the individual prohormone genes.
5. Known genes may not be found using text searches because the names of genes, proteins, and neuropeptides can vary between species and across time. In the absence of identifications using text searches, subsequent searches should explore gene symbols and synonyms. In some cases, text searches in other species may uncover alternative annotation, nomenclature, and search terms. The candidate prohormone table should include all synonyms of gene names and symbols.

6. Duplicated genes or homologs may match to the same location as well as different locations in the genome of the target species. Expected locations based on comparative mapping information should be evaluated first to identify the primary ortholog gene, and nearby matches are likely to be the result of tandem duplication. Also, incomplete coverage or assembly may cause one gene sequence to be mapped to substantially different regions. In these cases, the matches tend to be short and lack conserved regions. Other types of sequences such as ESTs can be very helpful to address gaps or low quality sequences.
7. Multiple sequence alignment is the favored method for identifying inaccuracies in prohormone protein prediction. An incorrect initiation codon will result in predicted protein isoforms that are contained within a subset of a larger protein isoform. Incorrect termination codons will result in the last exon being incorrectly predicted or completely missed. Gaps and mismatches indicate incomplete coverage or species differences that can only be resolved by other sources. Multiple sequence alignment should be used to resolve inaccuracies in the prediction, starting from the conserved regions and extending to both ends.
8. There are various assembly artifacts that have varying impacts on prohormone gene identification. Often many of these artifacts, such as matches in different strands or where exons of the gene are not sequentially located on the same contig, can be resolved by manually placing the translated sequence from exons in the correct order based on the candidate prohormone protein

sequence. A different resolution is required for erroneous insertions that result in part of all of an exon located multiple times in nearby regions. Typically only one of these regions will provide a complete gene prediction. If remaining incorrect regions does not provide a complete prohormone gene prediction, then the matches to those regions can be safely discarded.

9. Alternative specifications of the gene prediction tool should be investigated when the extracted genome sequence does not support the prediction of a complete or nearly complete protein, or when the predicted protein is substantially shorter or longer than expected. Effective strategies include varying the sequence region being analyzed, and removing excessive gaps or strings of ambiguous nucleotides. The global model option in GeneWise should be used when a high degree of homology is expected between the protein sequence and the genome of the target species. Using the gene prediction tool to predict the protein from the genome sequence of other species can offer insights into the expected prediction. Searching for the prohormone in other nucleotide databases of the target species could improve the prohormone prediction. Protein sequences should not be predicted using TBLASTN because this tool may provide a low prediction accuracy at the intron-exon splice boundaries, and may fail to identify all prohormone exons.
10. Signal peptide cleavage sites are typically located between 15 and 40 amino acids from the N-terminal start of the protein sequence. Signal peptides are highly conserved and thus, the predicted signal peptide cleavage is identical or very similar across related species, and often across members within a

prohormone family. Multiple sequence alignment of prohormone sequences from multiple species that have known or predicted signal peptide cleavage sites will enhance the accuracy of the prediction when the predicted prohormone sequence in the target species is partial or encompasses highly uncertain positions. If a site is not predicted, alternative specifications of the signal peptide prediction tool or different tools should be investigated.

11. The default specifications to predict peptides cleaved from protein sequences in NeuroPred were designed based on information for a large number of prohormones across multiple species and accommodate the sequences of many prohormones and neuropeptides. Users can overwrite these default specifications, e.g., by specifying the position of the signal peptide cleavage site when the input sequence is incomplete. The NeuroPred specifications for false-positive and false-negative cleavage predictions can also be adjusted according to the goals of the study.
12. Multiple species-specific models to predict cleavage sites and resulting peptides are available in NeuroPred. Cleavage and resulting peptide prediction models from the species phylogenetically closest to the target species should be used first. Models from more phylogenetically distant species should be examined for confirmation. NeuroPred supports the estimation of cleavage probability for each predicted cleavage site, which further empowers users to prioritize among the predicted peptides.

Acknowledgements

This work was supported by the National Institutes of Health, Award No. P30 DA018310 from the National Institute on Drug Abuse (NIDA), the US Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) project No. ILLU-538-909, and the National Science Foundation, Award No. CHE-16-06791. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

5 References

1. Burger E (1988) Peptide hormones and neuropeptides. Proteolytic processing of the precursor regulatory peptides. Arzneimittelforschung 38 (5):754-761

2. von Heijne G (1990) The signal peptide. The Journal of Membrane Biology 115 (3):195-201. doi:10.1007/bf01868635

3. Amare A, Hummon AB, Southey BR, Zimmerman TA, Rodriguez-Zas SL, Sweedler JV (2006) Bridging neuropeptidomics and genomics with bioinformatics: Prediction of mammalian neuropeptide prohormone processing. J. Proteome Res. 5 (5):1162-1167. doi:10.1021/pr0504541

4. Hu CK, Southey BR, Romanova EV, Maruska KP, Sweedler JV, Fernald RD (2016) Identification of prohormones and pituitary neuropeptides in the African

cichlid, Astatotilapia burtoni. BMC Genomics 17 (1):660. doi:10.1186/s12864-016-2914-9

5. Porter KI, Southey BR, Sweedler JV, Rodriguez-Zas SL (2012) First survey and functional annotation of prohormone and convertase genes in the pig. BMC Genomics 13:582. doi:10.1186/1471-2164-13-582

6. Southey BR, Rodriguez-Zas SL, Sweedler JV (2009) Characterization of the prohormone complement in cattle using genomic libraries and cleavage prediction approaches. BMC Genomics 10:228. doi:10.1186/1471-2164-10-228

7. Southey BR, Sweedler JV, Rodriguez-Zas SL (2008) A python analytical pipeline to identify prohormone precursors and predict prohormone cleavage sites. Front. Neuroinform. 2:7. doi:10.3389/neuro.11.007.2008

8. Southey BR, Sweedler JV, Rodriguez-Zas SL (2008) Prediction of neuropeptide cleavage sites in insects. Bioinformatics 24 (6):815-825. doi:10.1093/bioinformatics/btn044

9. Tegge AN, Southey BR, Sweedler JV, Rodriguez-Zas SL (2008) Comparative analysis of neuropeptide cleavage sites in human, mouse, rat, and cattle. Mamm Genome 19 (2):106-120. doi:10.1007/s00335-007-9090-9

10. Murphy D, Alim FZD, Hindmarch C, Greenwood M, Rogers M, Gan CK, Yealing T, Romanova EV, Southey BR, Sweedler JV (2016) Seasonal

Adaptations of the Hypothalamo-Neurohypophyseal System of the Arabian One-Humped Camel. Paper presented at the Plant and Animal Genome, San Diego, CA, USA,

11. Southey BR, Amare A, Zimmerman TA, Rodriguez-Zas SL, Sweedler JV (2006) NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. Nucleic Acids Res 34 (Web Server issue):W267-272. doi:10.1093/nar/gkl161

12. Southey BR AA, Zimmerman TA, Rodriguez-Zas SL, Sweedler JV (2017) NeuroPred application. http://neuroproteomics.scs.illinois.edu/neuropred.htm. Accessed 2/21/2017 2017

13. Southey BR, Rodriguez-Zas SL, Sweedler JV (2006) Prediction of neuropeptide prohormone cleavages with application to RFamides. Peptides 27 (5):1087-1098. doi:10.1016/j.peptides.2005.07.026

14. Southey BR, Rodriguez Zas SL (2017) PepShop application. http://stagbeetle.animal.uiuc.edu/pepshop. Accessed 2/21/2017 2017

15. Grimmelikhuijzen CJ, Hauser F (2012) Mini-review: the evolution of neuropeptide signaling. Regul Pept 177 Suppl:S6-9. doi:10.1016/j.regpep.2012.05.001

16. Romanova EV, Sweedler JV (2015) Peptidomics for the discovery and characterization of neuropeptides and hormones. Trends Pharmacol. Sci. 36 (9):579-586. doi:10.1016/j.tips.2015.05.009

17. Glasauer SM, Neuhauss SC (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol. Genet. Genomics 289 (6):1045-1060. doi:10.1007/s00438-014-0889-2

18. Coordinators NR (2017) Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res 45 (D1):D12-D17. doi:10.1093/nar/gkw1071

19. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P (2016) Ensembl 2016. Nucleic Acids Res 44 (D1):D710-716. doi:10.1093/nar/gkv1157

20. UniProt C (2015) UniProt: a hub for protein information. Nucleic Acids Res 43 (Database issue):D204-212. doi:10.1093/nar/gku989

21. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42 (Database issue):D756-763. doi:10.1093/nar/gkt1114

22. Southey BR AA, Zimmerman TA, Rodriguez-Zas SL, Sweedler JV (2017) NeuroPred sequence data. http://stagbeetle.animal.uiuc.edu/neuropred/sequences/Sequencedata.html. Accessed 2/21/2017 2017

23. Liu F, Baggerman G, Schoofs L, Wets G (2008) The construction of a bioactive peptide database in Metazoa. J. Proteome Res. 7 (9):4119-4131. doi:10.1021/pr800037n

24. Burbach JP (2010) Neuropeptides from concept to online database www.neuropeptides.nl. Eur J Pharmacol 626 (1):27-48. doi:10.1016/j.ejphar.2009.10.015

25. Falth M, Skold K, Norrman M, Svensson M, Fenyo D, Andren PE (2006) SwePep, a database designed for endogenous peptides and mass spectrometry. Mol. Cell. Proteomics 5 (6):998-1005. doi:10.1074/mcp.M500401-MCP200

26. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (17):3389-3402

27. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539. doi:10.1038/msb.2011.75

28. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res. 14 (5):988-995. doi:10.1101/gr.1865504

29. Wilkinson TN, Speed TP, Tregear GW, Bathgate RA (2005) Evolution of the relaxin-like peptide family. BMC Evol Biol 5:14. doi:10.1186/1471-2148-5-14

30. Wysolmerski JJ (2012) Parathyroid hormone-related protein: an update. J. Clin. Endocrinol. Metab. 97 (9):2947-2956. doi:10.1210/jc.2012-2142

31. Bhattacharya P, Yan YL, Postlethwait J, Rubin DA (2011) Evolution of the vertebrate pth2 (tip39) gene family and the regulation of PTH type 2 receptor (pth2r) and its endogenous ligand pth2 by hedgehog signaling in zebrafish development. J Endocrinol 211 (2):187-200. doi:10.1530/JOE-10-0439

32. Guerreiro PM, Renfro JL, Power DM, Canario AV (2007) The parathyroid hormone family of peptides: structure, tissue distribution, regulation, and potential

functional roles in calcium and phosphate balance in fish. Am J Physiol Regul Integr Comp Physiol 292 (2):R679-696. doi:10.1152/ajpregu.00480.2006

33. NCBI (2017) Gene database. https://www.ncbi.nlm.nih.gov/gene/. Accessed 2/21/2017 2017

34. NCBI (2017) Protein database. https://www.ncbi.nlm.nih.gov/protein/. Accessed 2/21/2017 2017

35. Nathoo AN, Moeller RA, Westlund BA, Hart AC (2001) Identification of neuropeptide-like protein gene families in Caenorhabditiselegans and other species. Proc Natl Acad Sci U S A 98 (24):14000-14005. doi:10.1073/pnas.241231298

36. Hummon AB, Richmond TA, Verleyen P, Baggerman G, Huybrechts J, Ewing MA, Vierstraete E, Rodriguez-Zas SL, Schoofs L, Robinson GE, Sweedler JV (2006) From the genome to the proteome: uncovering peptides in the Apis brain. Science 314 (5799):647-649. doi:10.1126/science.1124128

37. Gustincich S, Batalov S, Beisel KW, Bono H, Carninci P, Fletcher CF, Grimmond S, Hirokawa N, Jarvis ED, Jegla T, Kawasawa Y, LeMieux J, Miki H, Raviola E, Teasdale RD, Tominaga N, Yagi K, Zimmer A, Hayashizaki Y, Okazaki Y, Group RG, Members GSL (2003) Analysis of the mouse

transcriptome for genes involved in the function of the nervous system. Genome Res. 13 (6B):1395-1401. doi:10.1101/gr.1135303

38. Shi L, Ko ML, Abbott LC, Ko GY (2012) Identification of Peptide lv, a novel putative neuropeptide that regulates the expression of L-type voltage-gated calcium channels in photoreceptors. PLoS ONE 7 (8):e43091. doi:10.1371/journal.pone.0043091

39. Mirabeau O, Perlas E, Severini C, Audero E, Gascuel O, Possenti R, Birney E, Rosenthal N, Gross C (2007) Identification of novel peptide hormones in the human proteome by hidden Markov model screening. Genome Res. 17 (3):320-327. doi:10.1101/gr.5755407

40. Sonmez K, Zaveri NT, Kerman IA, Burke S, Neal CR, Xie X, Watson SJ, Toll L (2009) Evolutionary sequence modeling for discovery of peptide hormones. PLoS Comput. Biol. 5 (1):e1000258. doi:10.1371/journal.pcbi.1000258

41. Ozawa A, Lindberg I, Roth B, Kroeze WK (2010) Deorphanization of novel peptides and their receptors. AAPS J 12 (3):378-384. doi:10.1208/s12248-010-9198-9

42. NCBI (2017) Transcriptome Shotgun Assembly database https://www.ncbi.nlm.nih.gov/genbank/tsa/. Accessed 2/21/2017 2017

43. Suarez-Bregua P, Torres-Nunez E, Saxena A, Guerreiro P, Braasch I, Prober DA, Moran P, Cerda-Reverter JM, Du SJ, Adrio F, Power DM, Canario AV, Postlethwait JH, Bronner ME, Canestro C, Rotllant J (2017) Pth4, an ancient parathyroid hormone lost in eutherian mammals, reveals a new brain-to-bone signaling pathway. FASEB J. 31 (2):569-583. doi:10.1096/fj.201600815R

44. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44 (D1):D279-285. doi:10.1093/nar/gkv1344

45. Dores RM, Baron AJ (2011) Evolution of POMC: origin, phylogeny, posttranslational processing, and the melanocortins. Ann N Y Acad Sci 1220:34-48. doi:10.1111/j.1749-6632.2010.05928.x

46. Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33 (6):1635-1638. doi:10.1093/molbev/msw046

47. ETE G (2017) GenomeNet ETE3 application. http://www.genome.jp/tools/ete/. Accessed 2/21/2017 2017

48. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8 (10):785-786. doi:10.1038/nmeth.1701

Figure Captions

Fig 1. Outline of the bioinformatic steps leading to the accurate identification and annotation of a prohormone.

Fig. 2. Result of a text search for parathyroid hormone family genes in *Astatotilapia burtoni* within the NCBI Gene database where *Haplochromis burtoni* is the former name of *A. burtoni*.

Fig. 3. Output of the NCBI TBLASTN search of the parathyroid hormone 1 (PTH1) gene in the *Astatotilapia burtoni* genome.

Fig. 4. Output of the sequence of the second TBLASTN match of PTH1 in the *Astatotilapia burtoni* genome.

Fig. 5. Input to the GeneWise program using the *Astatotilapia burtoni* parathyroid hormone 1 protein sequence and genome region containing the suspected duplicated parathyroid hormone 1 gene.

Fig. 6. Output from GeneWise depicting the alignment between the amino acid sequence from the region extracted and the *Astatotilapia burtoni* PTH1 prohormone sequence and the predicted protein sequence of duplicated parathyroid hormone gene (denoted as PTH1B).

Fig. 7. Output of the *Astatotilapia burtoni* parathyroid hormone like prohormone TBLASTN search.

Fig. 8. Phylogenetic tree of parathyroid prohormone genes obtained using the GenomeNet ETE3. Leaves represent the gene symbol followed by species suffix: *Astatotilapia burtoni* (Ab); Danio rerio (Dr); *Maylandia zebra* (Mz); *Oreochromis niloticus* (On); *Neolamprologus brichardi* (Nb); *Pundamilia nyererei* (Pn); and *Homo sapiens* (human).

Fig. 9. Input to the NeuroPred tool to predict cleavage sites for *Astatotilapia burtoni* parathyroid hormone 1B (PTH1B) prohormone protein sequence with the 'Known Motif' and 'Mammalian' cleavage models selected.

Fig. 10. Output of the NeuroPred tool showing the cleavage diagram where the predicted cleavage sites are denoted by the letter 'C' for 'Known Motif' and 'Mammalian' cleavage models, and sequences and masses of first putative peptides.

Table 1. Summary of a text-based search for parathyroid hormone family genes in *Astatotilapia burtoni* within the NCBI Gene database.

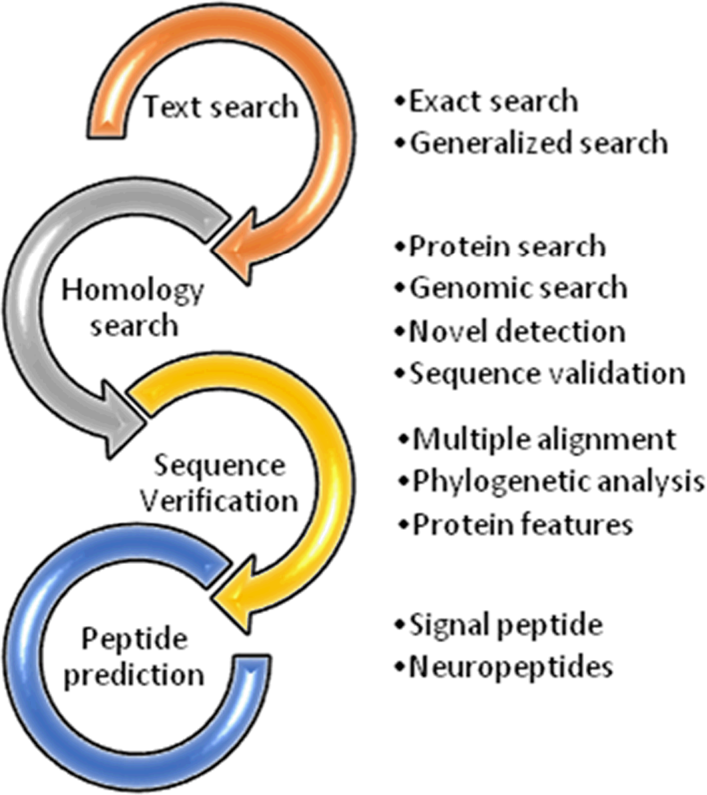
		Genomic Information		
Prohormone	**Gene ID**	**NCBI Accession Number of Contig**	**Position with Contig**	**NCBI Protein Accession Number**
PTH1	102295615	NW_005179605.1	11468..12171, complement	XP_005936534.1
PTH2	106633640	NW_005179731.1	310219..312857	XP_014195500.1
PTHLH	102305631	NW_005179673.1	98846..102040	XP_005939766.1; XP_014194690.1[a]

[a] PTHLH has 2 predicted protein isoforms due to different initiation codons. Subsequent evaluation indicated that the isoform (XP_005939766.1) has the same sequence length as the mammalian protein homologs and thus, will be entered in the prohormone table of candidate prohormones.

Table 2. Matches of the 7 predicted parathyroid prohormone family protein sequences to the protein databases for *Astatotilapia burtoni*, *Oreochromis niloticus*, *Neolamprologus brichardi, Pundamilia nyererei*, and *Maylandia zebra*.

	NCBI Protein Accession Number				
Gene Symbol	***A. burtoni***	***O. niloticus***	***N. brichardi***	***P. nyererei***	***M. zebra***
PTH1A	XP_005936534.1		XP_006780577.1	XP_005724427.1	XP_004543577.1
PTH1B			XP_006804843.1		
PTH2_v1		NP_001266421.1			XP_012779771.1
PTH2_v2	XP_014195500.1	XP_013120746.1		XP_013771025.1	
PTHLH1	XP_005939766.1	XP_003443941.1	XP_006782617.1	XP_005728183.1	XP_004563786.1
PTHLH2		XP_003448833.1	XP_006781115.1		XP_004564921.1
PTHLH3			XP_006783019.1	XP_005729696.1	

Type of file: figure
Label: Fig1
Filename: Figure1_SoutheyChapter.tiff


Text search
•Exact search
•Generalized search
Homology search
•Protein search
•Genomic search
•Novel detection
•Sequence validation
Sequence Verification
•Multiple alignment
•Phylogenetic analysis
•Protein features
Peptide prediction
•Signal peptide
•Neuropeptides

Type of file: figure
Label: Fig10
Filename: Figure10_SoutheyChapter.tiff

Number of sequences detected = 1

Individual Precursors

PTH1B

- Cleavage Prediction Diagram
- Mass of Predicted Peptides

PTH1B

TOP

Cleavage Prediction Diagram

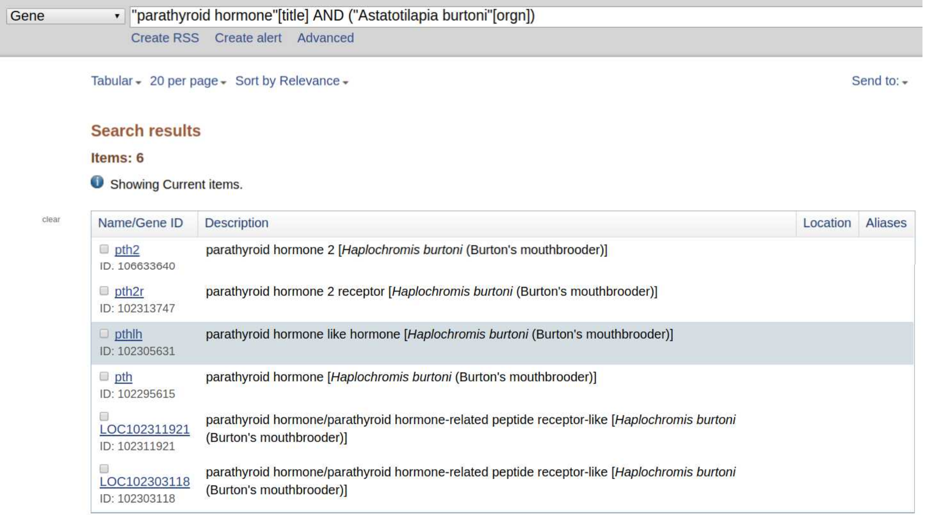
```
Sequence      MGKTDYKILL ISLCLLHFSV HCQGRPLSKR TVSEVQFMHN LGEHKQVQER
Known Motif   ssssssssss sssss..... ........rC .......... .........r
Mammal        ssssssssss sssss..... ........rC .......... .........r
Consensus     ssssssssss sssss..... ........rC .......... .........r
Sequence      REWLQMRLRG IHTAGARNSS RETTGRRRRR WPLRLEEMDE LSDLTSDEIQ
Known Motif   C......... .......... .....rCCCC .......... ..........
Mammal        C.......C. ......C... .....rCCCC .......... ..........
Consensus     C.......C. ......C... .....rCCCC .......... ..........
Sequence      HALNVLDELL KSE
Known Motif   .......... ...
Mammal        .......... ...
Consensus     .......... ...
```

TOP

Mass of Predicted Peptides

Abb. Peptide	NCut	CCut	PTM applied	Predicted Aver. Mass	Predicted Mono. Mass	Peptide sequence
L16_R30	Signal Peptidase	Known Motif, Mammal	Cleaved	1765.069800	1763.941690	LHFSVHCQGRPLSKR
L16_S28	Signal Peptidase	Known Motif, Mammal	TrimKR	1480.708200	1479.745620	LHFSVHCQGRPLS
T31_R51	Known Motif, Mammal	Known Motif, Mammal	Cleaved	2552.851300	2551.276500	TVSEVQFMHNLGEHKQVQERR

Type of file:	figure
Label:	Fig2
Filename:	Figure2_SoutheyChapter.tiff



"parathyroid hormone"[title] AND ("Astatotilapia burtoni"[orgn])

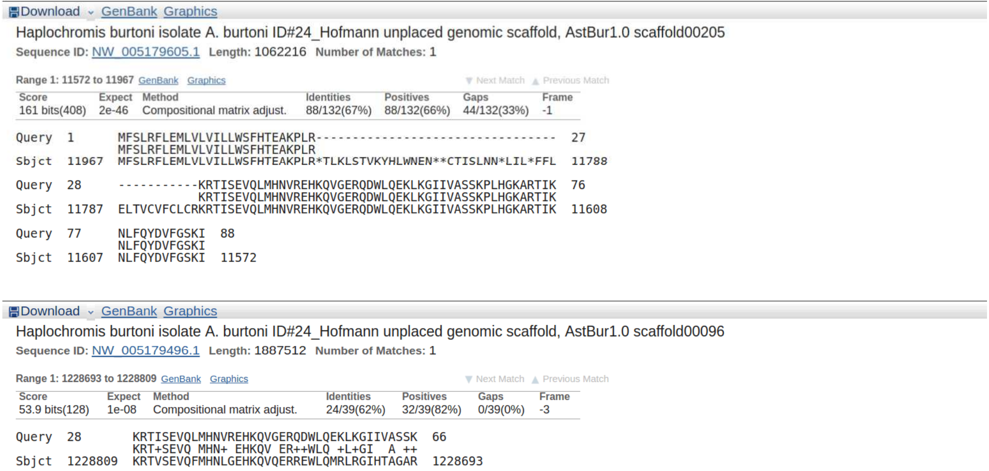
Search results

Items: 6

Showing Current items.

Name/Gene ID	Description	Location	Aliases
☐ pth2 ID: 106633640	parathyroid hormone 2 [*Haplochromis burtoni* (Burton's mouthbrooder)]		
☐ pth2r ID: 102313747	parathyroid hormone 2 receptor [*Haplochromis burtoni* (Burton's mouthbrooder)]		
☐ pthlh ID: 102305631	parathyroid hormone like hormone [*Haplochromis burtoni* (Burton's mouthbrooder)]		
☐ pth ID: 102295615	parathyroid hormone [*Haplochromis burtoni* (Burton's mouthbrooder)]		
☐ LOC102311921 ID: 102311921	parathyroid hormone/parathyroid hormone-related peptide receptor-like [*Haplochromis burtoni* (Burton's mouthbrooder)]		
☐ LOC102303118 ID: 102303118	parathyroid hormone/parathyroid hormone-related peptide receptor-like [*Haplochromis burtoni* (Burton's mouthbrooder)]		

Type of file: figure
Label: Fig3
Filename: Figure3_SoutheyChapter.tiff



Download ⌄ GenBank Graphics

Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00205

Sequence ID: NW_005179605.1 **Length:** 1062216 **Number of Matches:** 1

Range 1: 11572 to 11967 GenBank Graphics

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
161 bits(408)	2e-46	Compositional matrix adjust.	88/132(67%)	88/132(66%)	44/132(33%)	-1

```
Query  1      MFSLRFLEMLVLVILLWSFHTEAKPLR---------------------------------  27
              MFSLRFLEMLVLVILLWSFHTEAKPLR
Sbjct  11967  MFSLRFLEMLVLVILLWSFHTEAKPLR*TLKLSTVKYHLWNEN**CTISLNN*LIL*FFL  11788

Query  28     -----------KRTISEVQLMHNVREHKQVGERQDWLQEKLKGIIVASSKPLHGKARTIK  76
                         KRTISEVQLMHNVREHKQVGERQDWLQEKLKGIIVASSKPLHGKARTIK
Sbjct  11787  ELTVCVFCLCRKRTISEVQLMHNVREHKQVGERQDWLQEKLKGIIVASSKPLHGKARTIK  11608

Query  77     NLFQYDVFGSKI  88
              NLFQYDVFGSKI
Sbjct  11607  NLFQYDVFGSKI  11572
```

Download ⌄ GenBank Graphics

Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00096

Sequence ID: NW_005179496.1 **Length:** 1887512 **Number of Matches:** 1

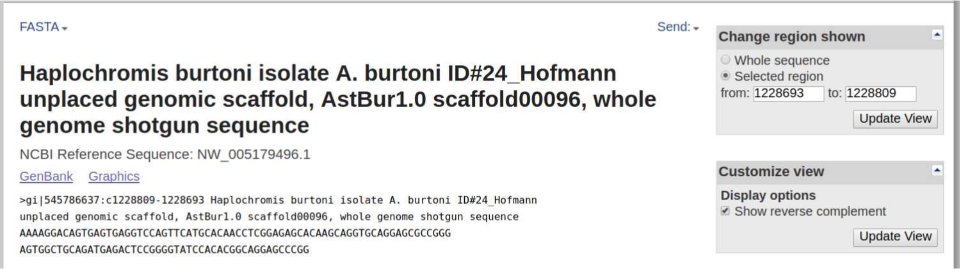
Range 1: 1228693 to 1228809 GenBank Graphics

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
53.9 bits(128)	1e-08	Compositional matrix adjust.	24/39(62%)	32/39(82%)	0/39(0%)	-3

```
Query  28       KRTISEVQLMHNVREHKQVGERQDWLQEKLKGIIVASSK  66
                KRT+SEVQ MHN+ EHKQV ER++WLQ +L+GI  A ++
Sbjct  1228809  KRTVSEVQFMHNLGEHKQVQERREWLQMRLRGIHTAGAR  1228693
```

Type of file: figure
Label: Fig4
Filename: Figure4_SoutheyChapter.tiff



Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00096, whole genome shotgun sequence

NCBI Reference Sequence: NW_005179496.1

```
>gi|545786637:c1228809-1228693 Haplochromis burtoni isolate A. burtoni ID#24_Hofmann
unplaced genomic scaffold, AstBur1.0 scaffold00096, whole genome shotgun sequence
AAAAGGACAGTGAGTGAGGTCCAGTTCATGCACAACCTCGGAGAGCACAAGCAGGTGCAGGAGCGCCGGG
AGTGGCTGCAGATGAGACTCCGGGGTATCCACACGGCAGGAGCCCGG
```

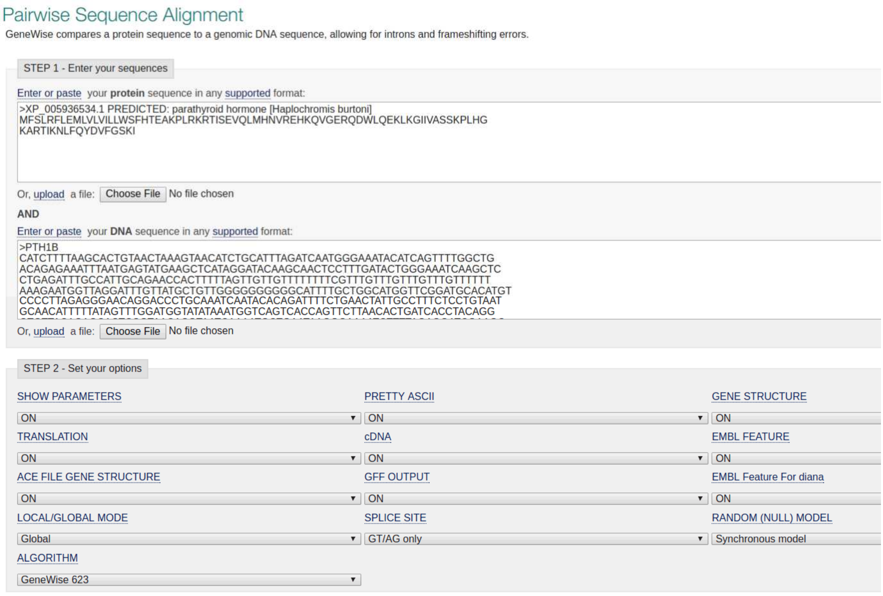
Change region shown

○ Whole sequence
● Selected region
from: 1228693 to: 1228809

Customize view

Display options

☑ Show reverse complement



Pairwise Sequence Alignment

GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

STEP 1 - Enter your sequences

Enter or paste your **protein** sequence in any supported format:

```
>XP_005936534.1 PREDICTED: parathyroid hormone [Haplochromis burtoni]
MFSLRFLEMLVLVILLWSFHTEAKPLRKRTISEVQLMHNVREHKQVGERQDWLQEKLKGIIVASSKPLHG
KARTIKNLFQYDVFGSKI
```

Or, upload a file: Choose File No file chosen

AND

Enter or paste your **DNA** sequence in any supported format:

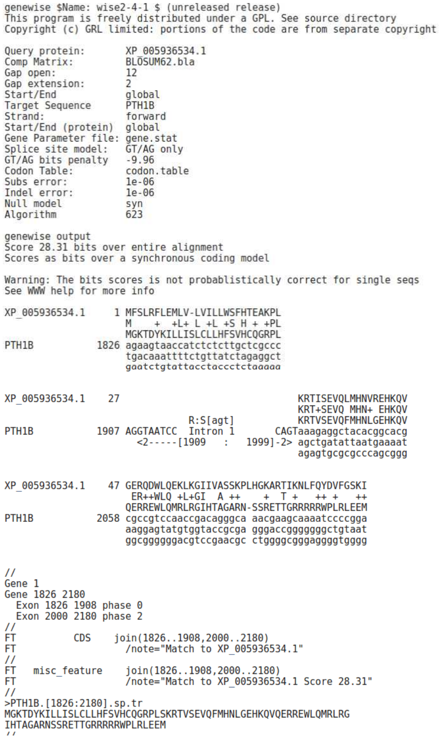
```
>PTH1B
CATCTTTTAAGCACTGTAACTAAAGTAACATCTGCATTTAGATCAATGGGAAATACATCAGTTTTGGCTG
ACAGAGAAATTTAATGAGTATGAAGCTCATAGGATACAAGCAACTCCTTTGATACTGGGAAATCAAGCTC
CTGAGATTTGCCATTGCAGAACCACTTTTTAGTTGTTGTTTTTTTTTCGTTTGTTTGTTTGTTTGTTTTTT
AAAGAATGGTTAGGATTTGTTATGCTGTTGGGGGGGGGGCATTTTGCTGGCATGGTTCGGATGCACATGT
CCCCTTAGAGGGAACAGGACCCTGCAAATCAATACACAGATTTTCTGAACTATTGCCTTTCTCCTGTAAT
GCAACATTTTTATAGTTTGGATGGTATATAAATGGTCAGTCACCAGTTCTTAACACTGATCACCTACAGG
```

Or, upload a file: Choose File No file chosen

STEP 2 - Set your options

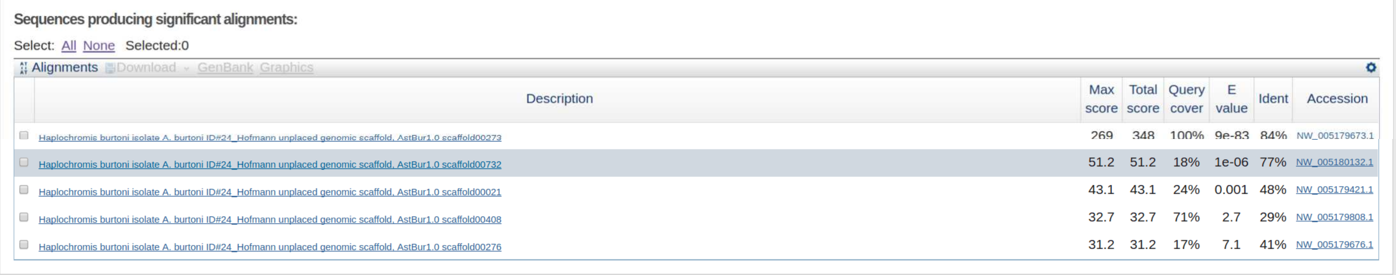
SHOW PARAMETERS	PRETTY ASCII	GENE STRUCTURE
ON	ON	ON
TRANSLATION	cDNA	EMBL FEATURE
ON	ON	ON
ACE FILE GENE STRUCTURE	GFF OUTPUT	EMBL Feature For diana
ON	ON	ON
LOCAL/GLOBAL MODE	SPLICE SITE	RANDOM (NULL) MODEL
Global	GT/AG only	Synchronous model
ALGORITHM		
GeneWise 623		

Type of file: figure
Label: Fig6
Filename: Figure6_SoutheyChapter.tiff



```
genewise $Name: wise2-4-1 $ (unreleased release)
This program is freely distributed under a GPL. See source directory
Copyright (c) GRL limited: portions of the code are from separate copyright

Query protein:       XP_005936534.1
Comp Matrix:         BLOSUM62.bla
Gap open:            12
Gap extension:       2
Start/End            global
Target Sequence      PTH1B
Strand:              forward
Start/End (protein)  global
Gene Parameter file: gene.stat
Splice site model:   GT/AG only
GT/AG bits penalty   -9.96
Codon Table:         codon.table
Subs error:          1e-06
Indel error:         1e-06
Null model           syn
Algorithm            623

genewise output
Score 28.31 bits over entire alignment
Scores as bits over a synchronous coding model

Warning: The bits scores is not probablistically correct for single seqs
See WWW help for more info

XP_005936534.1      1 MFSLRFLEMLV-LVILLWSFHTEAKPL
                      M    +  +L+ L +L +S H + +PL
                      MGKTDYKILLISLCLLHFSVHCQGRPL
PTH1B            1826 agaagtaaccatctctcttgctcgccc
                      tgacaaattttctgttatctagaggct
                      gaatctgtattacctaccctctaaaaa


XP_005936534.1     27                               KRTISEVQLMHNVREHKQV
                                                    KRT+SEVQ MHN+ EHKQV
                                 R:S[agt]           KRTVSEVQFMHNLGEHKQV
PTH1B            1907 AGGTAATCC  Intron 1       CAGTaaagaggctacacggcacg
                        <2-----[1909   :   1999]-2> agctgatattaatgaaaat
                                                    agagtgcgcgcccagcggg


XP_005936534.1     47 GERQDWLQEKLKGIIVASSKPLHGKARTIKNLFQYDVFGSKI
                       ER++WLQ +L+GI  A ++    +  T +   ++ +   ++
                      QERREWLQMRLRGIHTAGARN-SSRETTGRRRRRWPLRLEEM
PTH1B            2058 cgccgtccaaccgacagggca aacgaagcaaaatccccgga
                      aaggagtatgtggtaccgcga gggaccgggggggctgtaat
                      ggcgggggacgtccgaacgc ctggggcgggaggggtgggg


//
Gene 1
Gene 1826 2180
  Exon 1826 1908 phase 0
  Exon 2000 2180 phase 2
//
FT          CDS      join(1826..1908,2000..2180)
FT                   /note="Match to XP_005936534.1"
//
FT   misc_feature    join(1826..1908,2000..2180)
FT                   /note="Match to XP_005936534.1 Score 28.31"
//
>PTH1B.[1826:2180].sp.tr
MGKTDYKILLISLCLLHFSVHCQGRPLSKRTVSEVQFMHNLGEHKQVQERREWLQMRLRG
IHTAGARNSSRETTGRRRRRWPLRLEEM
//
```

Type of file:	figure
Label:	Fig7
Filename:	Figure7_SoutheyChapter.tiff

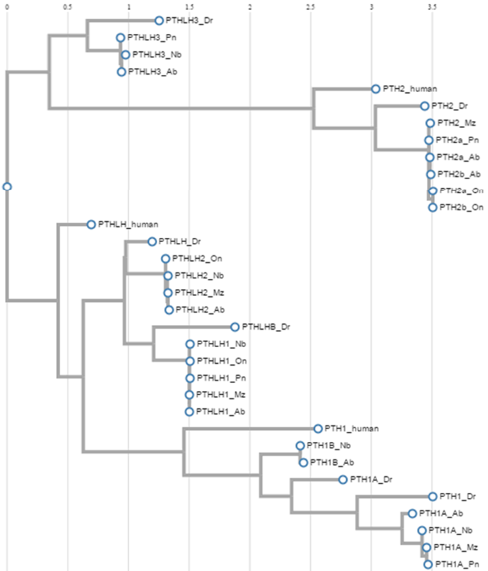


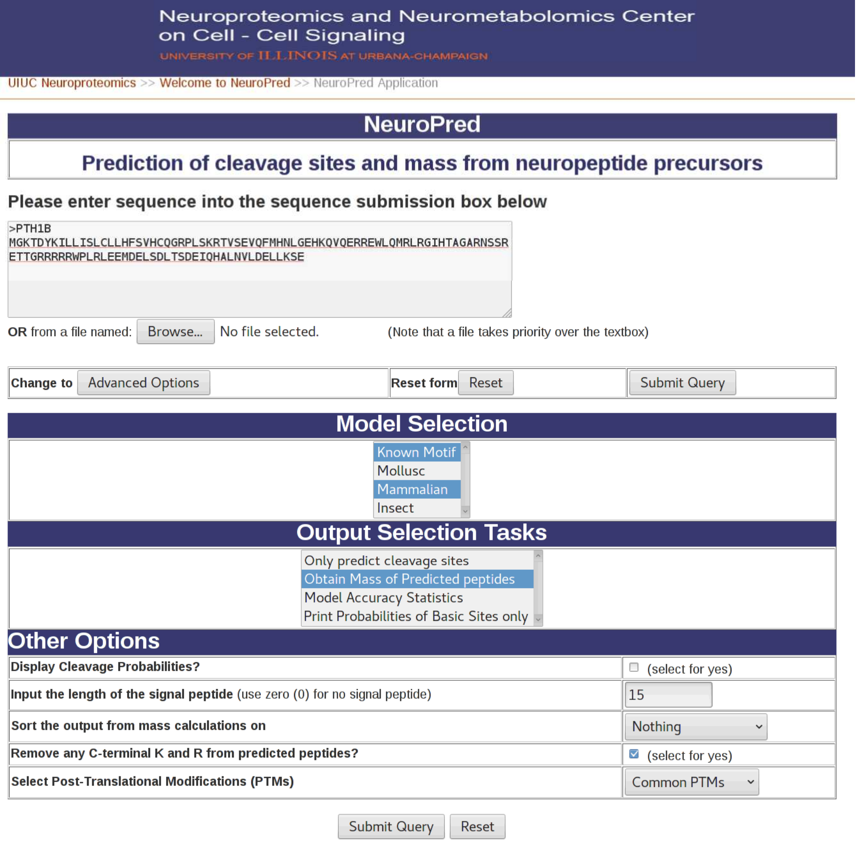
Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00273	269	348	100%	9e-83	84%	NW_005179673.1
□	Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00732	51.2	51.2	18%	1e-06	77%	NW_005180132.1
□	Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00021	43.1	43.1	24%	0.001	48%	NW_005179421.1
□	Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00408	32.7	32.7	71%	2.7	29%	NW_005179808.1
□	Haplochromis burtoni isolate A. burtoni ID#24_Hofmann unplaced genomic scaffold, AstBur1.0 scaffold00276	31.2	31.2	17%	7.1	41%	NW_005179676.1

Type of file: figure
Label: Fig8
Filename: Figure8_SoutheyChapter.tiff


0
0.5
1
1.5
2
2.5
3
3.5
PTHLH3_Dr
PTHLH3_Pn
PTHLH3_Nb
PTHLH3_Ab
PTH2_human
PTH2_Dr
PTH2_Mz
PTH2a_Pn
PTH2a_Ab
PTH2b_Ab
PTH2a_On
PTH2b_On
PTHLH_human
PTHLH_Dr
PTHLH2_On
PTHLH2_Nb
PTHLH2_Mz
PTHLH2_Ab
PTHLHB_Dr
PTHLH1_Nb
PTHLH1_On
PTHLH1_Pn
PTHLH1_Mz
PTHLH1_Ab
PTH1_human
PTH1B_Nb
PTH1B_Ab
PTH1A_Dr
PTH1_Dr
PTH1A_Ab
PTH1A_Nb
PTH1A_Mz
PTH1A_Pn



Neuroproteomics and Neurometabolomics Center on Cell - Cell Signaling

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UIUC Neuroproteomics >> Welcome to NeuroPred >> NeuroPred Application

NeuroPred

Prediction of cleavage sites and mass from neuropeptide precursors

Please enter sequence into the sequence submission box below

```
>PTH1B
MGKTDYKILLISLCLLHFSVHCQGRPLSKRTVSEVQFMHNLGEHKQVQERREWLQMRLRGIHTAGARNSSR
ETTGRRRRRWPLRLEEMDELSDLTSDEIQHALNVLDELLKSE
```

OR from a file named: Browse... No file selected. (Note that a file takes priority over the textbox)

Change to Advanced Options	**Reset form** Reset	Submit Query

Model Selection

Known Motif
Mollusc
Mammalian
Insect

Output Selection Tasks

Only predict cleavage sites
Obtain Mass of Predicted peptides
Model Accuracy Statistics
Print Probabilities of Basic Sites only

Other Options

Display Cleavage Probabilities?	□ (select for yes)
Input the length of the signal peptide (use zero (0) for no signal peptide)	15
Sort the output from mass calculations on	Nothing
Remove any C-terminal K and R from predicted peptides?	☑ (select for yes)
Select Post-Translational Modifications (PTMs)	Common PTMs

Submit Query Reset