In silico searches for putative growth hormone family homologs in invertebrates

Daniel **Ocampo Daza** and Dan **Larhammar**
Department of Neuroscience, Science for Life Laboratory, Uppsala Universitet, Box 593, SE-75124 Uppsala, Sweden

E-mail: Daniel.Ocampo-Daza@neuro.uu.se

Background

The somatotropin hormone protein family includes growth hormone (GH), prolactin (PRL) and somatolactin (SL), as well as lineage-specific duplicates of these. The proteins are characterized by a conserved amino-acid motif of four cysteine residues that form two disulfide bridges. GH/PRL/SL-family sequences were sought in diverse invertebrate genome and reference sequence databases using different available strategies. None of these search strategies could identify any putative invertebrate GH/PRL/SL homologs with the characteristic amino acid motif of the family, or indeed any significant sequence similarity. One sequence with high sequence identity to human PRL was identified from a tapeworm species; however this sequence is likely the result of horizontal gene transfer or contamination.

Methods and results

The genomes that were analyzed are given here together with assembly information and database address: *Ciona intestinalis* (vase tunicate), assembly JGI2 accessed from the Ensembl genome browser (version 65, Dec 2011, http://www.ensembl.org); *Oikopleura* dioica (free-swimming tunicate) assembly v3 accessed from Genoscope (http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura); Branchiostoma floridae (lancelet) assembly v2.0 accessed from the JGI genome portal (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html); Saccoglossus kowalevskii (acorn worm) build 1.1 accessed from NCBI's Genome Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/); Strongylocentrotus purpuratus (sea urchin) database versions 0.5 and 3.1 accessed from SpBase (http://www.spbase.org/SpBase/). Where possible, tblastn searches (1) were made using human and zebrafish GH/PRL/SL sequences as queries (see table below). The only exception was the free-swimming tunicate genome database where BLAT-searches were made using the same queries. None of the searches in these databases produced significant hits, even with suboptimal expect values.

Becausee of the relatively low sequence conservation within the GH/PRL/SL-family, as well as the large evolutionary distance between the identified vertebrate sequences and any putative invertebrate homologs, the tblastn or BLAT algorithms might not identify enough sequence similarity to produce significant hits. Therefore Pattern Hit Initiated BLAST (PHI-BLAST) searches (2) were carried out in the *Saccoglossus kowalevskii* (acorn worm) genome build 1.1, *Drosophila melanogaster* (fruit fly) genome build 5.30 and *Apis mellifera* (honey bee) genome build 4.5 accessed from NCBI's Genome Map Viewer (*http://www.ncbi.nlm.nih.gov/mapview/*) as a strategy to find more divergent sequences that

still carry the characteristic motif of four conserved cysteine residues. Two different consensus patterns were used as PHI-patterns and the human PRL sequence (Ensembl database ID ENSP00000302150) was used as query sequence for every search. The two consensus patterns were identified in the ProSite database of protein domains, families and functional sites (http://prosite.expasy.org) as somatotropin (growth hormone), prolactin and related hormones signatures with the accession numbers PS00266 and PS00338. These profiles match a large diversity of protein sequence entries in the UniProtKB and Swiss-Prot databases for growth hormones, prolactins, somatolactins and related lineage-specific proteins such as placental growth hormone variants, proliferins and placental prolactin-like proteins. Primarily, these profiles identify the four conserved cysteine residues that are considered a characteristic amino-acid motif for the GH/PRL/SL-family. The human PRL sequence was used as query, and the fruitfly and honey bee genomes were targeted, because the presence of prolactin-like peptides in these species has been suggested by *in silico* analysis (3) and immunohistochemical methods (original study cited in reference (4)). Our PHI-BLAST searches for GH/PRL/SL-like sequences in acorn worm, fruit fly and honey bee produced no hits. The putative prolactin-like sequences identified by Liu et. al. (2006) (3) are identified in both NCBI's reference sequence database and the UniProtKB database (http://www.uniprot.org) with the accession numbers CG9358 and BK003312. Based on protein domain and functional site identifications through InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) these sequences do not display the significant pattern of amino acid conservation to be considered prolactin homologs. Indeed, CG9358 is identified as pherokine 3, a completely unrelated protein with kinase activity, in both NCBI RefSeq and UniProt, and as an insect pheromone-binding protein through InterProScan.

Query sequences for BLAST searches

	Reference IDs
hGH1	Ensembl: ENSG00000259384
h*PRL*	Ensembl: ENSP00000302150
z*fgh*	Ensembl: ENSDARG00000038185
z*fprl*	Ensembl: ENSDARG00000037946
z*fprl2*	NCBI GenBank: NM_001162854.1
z*fsmtla*	NCBI Gene: 553408

Because the PHI-BLAST searches of three invertebrate genome builds were unsuccessful, putative invertebrate homologs were also sought through profile-Hidden Markov protein sequence similarity searches in the HMMER web server (http://hmmer.janelia.org/) (5) using the phmmer algorithm with standard settings and the human prolactin sequence as query. These searches could only identify one invertebrate sequence; in the canine tapeworm *Taenia hydatigena*. It is identified by the accession number Q8T110_9CEST in the UniProtKB database, where it is annotated as a prolactin fragment (http://www.uniprot.org/uniprot/Q8T110). This amino acid sequence was aligned to an alignment of vertebrate GH/PRL/SL sequences using the profile alignment function in ClustalX version 2.1 (http://www.clustal.org) and a neighbor joining (NJ) phylogenetic tree was made in ClustalX (NJ clustering algorithm, tree supported by 1000 bootstrap iterations).

The phylogenetic analysis (see data files below) shows that the identified *Taenia hydatigena* sequence is most similar to the human prolactin sequence and nests within the mammalian branch of prolactins (PRL) with good branch support. The amino acid sequence identity between this sequence and the human prolactin sequence is approx. 78% in a

pairwise alignment (see data files below), an unlikely degree of sequence similarity considering the large evolutionary distance between the two species. The pairwise alignment was made in Jalview 2.8 (http://www.jalview.org) applying the Smith-Waterman algorithm. This tapeworm sequence is likely the result of a horizontal gene transfer event or a contamination. In any case it is unlikely to represent an ancestor of the vertebrate GH, PRL and SL sequences.

Data files

130723_Taenia_prl-hit_NJtree.phb
Neighbor joining tree file in Newick format including the identified tapeworm sequence. Vertebrate species names are abbreviated to the first letter of the genus followed by the first two letters of the species (Hsa = *Homo sapiens* et. c.). The tapeworm sequence is identified with the UniProt ID Q8T110_9CEST.

130723_Taenia_prl-hit_NJtree.pdf
Image of the neighbor joining tree (PDF). The tapeworm sequence is colored green.

Pairwise_align_hPRL+Taenia_prl-hit.txt
Text file with the results of the pairwise alignment between the human PRL sequence and the identified tapeworm sequence.

HsaPRL+Taenia_prl-hit.aln
Clustal alignment file containing the human PRL sequence and the identified tapeworm sequence.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215:403–10.

2. Sayers EW et al. (2010) Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 38:D5–16.

3. Liu F et al. (2006) In silico identification of new secretory peptide genes in Drosophila melanogaster. *Molecular & cellular proteomics : MCP* 5:510–22.

4. Quintanar JL et al. (2007) Prolactin-like hormone in the nematode Trichinella spiralis larvae. *Experimental parasitology* 116:137–41.

5. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic acids research* 39 Suppl 2:W29–37.