Comparative and evolutionary analysis of Growth Hormone (GH) protein of Cyprinids using computational approach

Dibyashree Swain, Kiran Dashrath Rasal, Pranati Swain, Prajna Pradhan, Jitendra Kumar Sundaray, Hirak Kumar Barman and Pallipuram Jayasankar

Abstract
Growth hormone (GH) is an important polypeptide necessary for normal growth as well as development in fishes. Even though, GH has been well characterized in several fishes, lack of information on the structure GH protein. Recent advances of computational tools are helpful for developing structure using protein sequences using either homology modelling or de-novo modelling. In this study, we have used computational tools for three-dimensional structure prediction and analysis evolutionary profile of GH from 12 fishes belonging to cyprinids. The sequence analysis predicted GH polypeptide of 204-210 amino-acids, including a putative signal peptide of 1-22 hydrophobic residues and mature protein containing 183 amino-acids among cyprinid fishes. We could identify well conserved region among all the species, which states, that region maintained by evolution despite speciation. The evolutionary analysis was performed using CLC genomic workbench followed by a bootstrap test (100 replicates). A phylogenetic tree was depicting the relationship of various cyprinids GH protein sequences. We have generated 3D structure of GH protein of rohu using I-TASSER, which was further validated using SAVES server. The Procheck analysis depicted the Ramachandran plot consisting of 76.8% residues lies in the most favored regions. For understanding global network of GH, STRING9.1 speculated that, this protein interacting with several proteins, strong association with two proteins such as GHRa and GHRb. A comparative study of different sequence features among them. Currently with the wide availability of sequence data (Nucleotide/Protein), we can gain previously impossible insights into evolution of protein universe, on the domain, network, and genome levels. Network level studies allow us to understand the forces of evolution that affect changes and conservation of protein networks.

Keywords: Growth hormone, Modelling, I-TASSER, Phylogeny, Three dimensional structure.

1. Introduction
Over the period of decades, various features of growth hormone (GH) gene physiology along with molecular aspects have been studied in several fish species. Preferably, the growth rate of fish is being an important economic trait which needs to be enhanced with help of traditional selective breeding practices as well as gene manipulation techniques for mitigating growing fish food demand. The growth rate of fish has been controlled by various genes such as growth hormone (GH), Insulin-like growth factor-I (IGF-I) and Growth hormone receptors (GHRa/b), which had explored with help of molecular techniques [1, 2, 3] The growth hormone, which mainly regulates the growth and development as well as significant role in osmoregulation, steroidogenesis, anti-freeze protein (AFP) production, reproduction in fishes [2, 4, 5, 6, 7]. In fishes, increased attention pertains to growth has led to investigation of gene transcripts encoding GH, IGF-I, IGF-II, and their receptors along with IGF-binding proteins [3, 8]. It has been demonstrated that two sub-types of the GH receptor (GHR1 and GHR2) are present in most of the teleost [2, 9]. In addition, transgenic approach has been utilized in several fish species such as Atlantic salmon [5], catfish [10], carp [11], tilapia [12], loach [13]. Arctic char [14] for improvement production and performance traits such as growth, disease resistance. Recent knowledge of the complete sequences of some farmed carp species genes has led us to a comparative study of different sequence features among them. Currently with the wide availability of sequence data (Nucleotide/Protein), we can gain previously impossible insights into evolution of protein universe, on the domain, network, and genome levels. Network level studies allow us to understand the forces of evolution that affect changes and conservation of protein networks.
Understanding the evolutionary pressures that affect protein networks provides insight into their functional importance for various organisms, including how they affect the organism's competitiveness. The study of protein evolution has also become a key area of scientific research with discoveries such as the coordinated changes of key residues and substitution leading to a phenotype change [15]. The theory and practice of phylogenetic tree construction matured into the PHYLIP program, and phylogenetic analysis yielded significant discoveries in genome evolution, such as the relationships between life forms and the dynamics of genome structure [16]. Therefore, this wealth of DNA/protein sequence data in the database provided the perfect resources to gain further information into protein composition and their function, protein-protein interaction network and evolution. The improvement in the computational algorithms/tools made new avenue for studying the genome/proteome at wide scale.

In this study, although, there is availability of GH sequences and its encoded protein sequences in the database for cyprinids, but lack of species-specific structural information. This led to absence of knowledge of protein network and evolution in fishes. In this study, we have retrieved protein sequences of cyprinids, used for comparative as well as evolutionary analysis. We have predicted the three dimensional structure and its protein-interaction network. Also, we have used molecular docking approach to investigate receptor binding activities. Our results provide meaningful information for further studies on the cyprinids and their phylogeny.

2.0 Material and Methods
We have used different bioinformatics tools for studying the growth hormone protein, which listed along with specified purpose.

2.1 Retrieval of GH protein sequences
The UniProt is easily accessible database of protein sequence (http://www.uniprot.org/). We have retrieved total 12 protein sequences from different fishes belongs to cyprinids for our study (Table1). We have retrieved all protein sequences in a FASTA format with accession numbers.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>UniProtKB accession number</th>
<th>Number of amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danio rerio</td>
<td>Q571R1</td>
<td>210</td>
</tr>
<tr>
<td>Labeo rohita</td>
<td>Q9W6J7</td>
<td>207</td>
</tr>
<tr>
<td>Labeo calbasu</td>
<td>Q5Y4C2</td>
<td>210</td>
</tr>
<tr>
<td>Labeo bata</td>
<td>Q5Y4C1</td>
<td>210</td>
</tr>
<tr>
<td>Labeo fimbriatus</td>
<td>Q5Y4C8</td>
<td>210</td>
</tr>
<tr>
<td>Labeo koius</td>
<td>Q5Y4C6</td>
<td>210</td>
</tr>
<tr>
<td>Cyprinus carpio</td>
<td>P10298</td>
<td>210</td>
</tr>
<tr>
<td>Carassius auratus</td>
<td>Q93359</td>
<td>210</td>
</tr>
<tr>
<td>Hypophthalmichthys molitrix</td>
<td>P69159</td>
<td>210</td>
</tr>
<tr>
<td>Ctenopharyngodon idella</td>
<td>P69158</td>
<td>210</td>
</tr>
<tr>
<td>Catla catla</td>
<td>Q90WV7</td>
<td>210</td>
</tr>
<tr>
<td>Cirrhinus miragala</td>
<td>Q90W30</td>
<td>210</td>
</tr>
<tr>
<td>Cyprinus tinca</td>
<td>D7RPP4</td>
<td>210</td>
</tr>
</tbody>
</table>

2.2 Physico-chemical Characterization and domain, motif analysis of GH proteins
The properties of GH proteins was determined and characterized from the protein sequences. Thus, ProtParam (http://web.expasy.org/protoparam/) expasy tool, which is useful for computation of physical and chemical parameters of given protein based on sequence information. We have calculated several physic-chemical properties such as theoretical isoelectric point (pI), molecular weight, and total number of positive and negative residues, extinction coefficient, half-life, instability index, aliphatic index and grand average hydrophathy (GRAVY) of all 12 retrieved protein sequences. The PROSITE server was used to determine specific profiles as well as patterns related to the GH proteins. The signal peptide site were predicted using SignalP server (version 4.0), which is based on neural network method (http://www.cbs.dtu.dk/services/SignalP/).

2.3 Multiple Sequence Alignment and Phylogenetic study of GH
In order to determine comparison among different protein sequences, we have used global multiple sequence alignment (MSA) program for analysis of GH protein sequences from different fishes. Now a day, multiple sequence alignment (MSA) method is widely used for assessing sequence conservation and conservation of protein domains in protein study. In this step, CLC genomic workbench 7.5.1 tool was used for MSA analysis. Understanding phylogenetic relationship among different protein sequences, we have observed the evolutionary relationship of these sequences by cladogram. The Evolutionary analysis of GH protein among 12 species was conducted CLC genomic workbench 7.5.1. The NAPplot software was utilized for obtaining a graphical view of the phylogenetic tree of GH sequences.

2.4 Analysis of Gene ontology and protein-protein interaction network of GH
We further studied the gene ontology of GH protein obtaining biological, molecular and cellular functions, which recognized using Uniprot server (http://www.uniprot.org/). The STRING algorithm (Search Tool for the Retrieval of Interacting Proteins) was used for studying the protein-protein interaction network of the GH (http://string-db.org/).

2.5 Three dimensional structure analysis
For getting homologous 3D structure for GH protein, BLASTp (Basic local alignment) searches were performed against protein databank (PDB). But, due to absence of homologous 3D structure/template, we have used I-TASSER (Iterative Threading ASSEMBly Refinement) server, which is an on-line platform for protein structure -function prediction (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). It is based on multiple threading alignments and iterative structural assembly simulations. The prediction of accuracy of the model depend upon a confidence score (C-score) based on the quality of the threading alignments and structural assembly refinement simulations [17]. In this server, we have submitted as a query sequence (FASTA) of GH protein for obtaining 3D model with academic id. The best/top 3D model was selected according rank based on C-score, TM score and RMSD value. The 3D structure of GH was visualized by PyMOL (http://www.pymol.org/) and UCSF CHIMERA (https://www.cgl.ucsf.edu/chimera/). The PyMOL is open source molecular visualization tool for interactive visualization and analysis of molecular structures. The structural evaluation along with stereo-chemical quality assessment of predicted model were carried out by using the SAVES (Structural Analysis and Verification Server), which is integrated server (http://nihserver.mbi.ucla.edu/SAVES/). The
ProSA (Protein Structure Analysis) web server (https://prosa.services.came.sbg.ac.at/prosa) was used for refinement and validation of protein structure [18]. The ProSA was used for checking model structural quality with potential errors and program shows a plot of its residue energies and Z-scores which determine overall quality of the model.

3. Results and Discussion
Recently, several bioinformatics tools/algorithms are being used in diverse biological fields, alternatively, improvement in the precision of experiment [19, 20, 21, 22]. In this work, in order to understand evolutionary as well as comparative analysis of the GH protein of fishes belonging to cyprinids, we have used several bioinformatics algorithms. The precise annotations as well as classification of proteins, which deduced from nucleotide sequences, are important for the perfect understanding and study of protein sequence. Thus, information available via the protein- database can be utilized for evaluation and/or prediction of specific characters about a protein of interest. Although, the number of aa residues showed 171 to 217 in human, bovine, and in some species and thus allow identification of those sections of the amino acid sequence most critical to protein function. For this, several advanced and accurate algorithms/tools have been developed such as CLUSTAL-W, T-Coffee, MAFFT, based on the fast Fourier transformation, and MEGA [28, 30, 31, 32]. After performing MSA, we have seen that “FRLLACFKKDMHKVETYLRVANCRRSLDNSCTL” is the conserved region (identical region) in all 12 different protein sequences of GH which specifies that, this peptide sequence may have been maintained by evolution despite speciation (Fig 1). Again, we have made BLASTp of this conserved sequence in NCBI and revealed that, sequence is conserved among all the fishes with more than 92% identity. In our study, phylogenetic tree was constructed CLC genomic workbench UPMGA and NJ methods (Fig 2). The earlier study demonstrated that, bootstrapping method generates reliable trees using re-sampling the given dataset several times [33, 34]. We observed that, Labeo species fishes are more related to catla than cyprinus species at protein level. The previous studies depicted that, phylogenetic analysis and comparative modelling for depiction of growth hormone gene of ornamental fishes [35]. The earlier study revealed, GH sequences cDNAs, and protein sequences have been used as a phylogenetic marker for studies of evolutionary genetics of various fishes [36, 37].

Table 2: Physico-chemical characteristics of growth hormone protein of 13 fishes of cyprinids

<table>
<thead>
<tr>
<th>Taxon</th>
<th>MW (Da)</th>
<th>pI</th>
<th>Negative charged residues</th>
<th>Positive charged residues</th>
<th>Formula</th>
<th>AI</th>
<th>GRAVY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danio rerio</td>
<td>23768.4</td>
<td>6.32</td>
<td>25</td>
<td>24</td>
<td>C1047H1703N287O317S12</td>
<td>100.24</td>
<td>-0.139</td>
</tr>
<tr>
<td>Labeo rohita</td>
<td>23521.1</td>
<td>6.31</td>
<td>26</td>
<td>25</td>
<td>C1004H1682N290O317S12</td>
<td>105.02</td>
<td>-0.142</td>
</tr>
<tr>
<td>Labeo calbasu</td>
<td>23804.4</td>
<td>6.31</td>
<td>26</td>
<td>25</td>
<td>C1047H1703N287O317S12</td>
<td>104.90</td>
<td>-0.140</td>
</tr>
<tr>
<td>Labeo bata</td>
<td>23805.5</td>
<td>6.31</td>
<td>26</td>
<td>25</td>
<td>C1047H1703N287O317S12</td>
<td>105.38</td>
<td>-0.125</td>
</tr>
<tr>
<td>Labeo fimbriatus</td>
<td>23790.4</td>
<td>6.31</td>
<td>26</td>
<td>25</td>
<td>C1047H1703N287O317S12</td>
<td>104.43</td>
<td>-0.141</td>
</tr>
<tr>
<td>Labeo kontius</td>
<td>23790.4</td>
<td>6.31</td>
<td>26</td>
<td>25</td>
<td>C1047H1703N287O317S12</td>
<td>104.43</td>
<td>-0.141</td>
</tr>
<tr>
<td>Cyprinus carpio</td>
<td>23765.3</td>
<td>5.82</td>
<td>25</td>
<td>22</td>
<td>C1048H1682N317O317S12</td>
<td>99.33</td>
<td>-0.159</td>
</tr>
<tr>
<td>Carassius auratus</td>
<td>23759.3</td>
<td>5.95</td>
<td>25</td>
<td>23</td>
<td>C1048H1682N317O317S12</td>
<td>97.48</td>
<td>-0.173</td>
</tr>
<tr>
<td>Hypophthalmichthys molitrix</td>
<td>23580.1</td>
<td>5.96</td>
<td>25</td>
<td>23</td>
<td>C1048H1682N317O317S12</td>
<td>99.33</td>
<td>-0.140</td>
</tr>
<tr>
<td>Chelodonyx catla</td>
<td>23790.4</td>
<td>6.31</td>
<td>26</td>
<td>25</td>
<td>C1047H1703N287O317S12</td>
<td>104.43</td>
<td>-0.141</td>
</tr>
<tr>
<td>Cirrhinus mesembrinus</td>
<td>23791.4</td>
<td>6.31</td>
<td>26</td>
<td>25</td>
<td>C1047H1703N287O317S12</td>
<td>104.90</td>
<td>-0.127</td>
</tr>
<tr>
<td>Cyprinus sinca</td>
<td>23699.2</td>
<td>6.32</td>
<td>25</td>
<td>24</td>
<td>C1048H1682N317O317S12</td>
<td>97.00</td>
<td>-0.175</td>
</tr>
</tbody>
</table>

The PROSITE server is one of the most widely utilized for abundant pattern and profile analysis of a given protein [29]. Thus, functions of the GH proteins were analyzed by submitting the aa sequences to the PROSITE server. In subsequent analysis, Multiple Sequence Alignment (MSA) can give insight into sequence conservation across several species and thus allow identification of those sections of the sequence most critical to protein function. For this, the amino acid leucine (L=16%) and serine (S=12.3%) has been found mainly abundant in the GH of these 12 fish species. The aliphatic index (AI) for GH was calculated in the range of 97 to 105 and predicted as stable. Because, if the instability index of a given protein is < 40, it is predicted as a stable, while > 40 indicated as unstable protein [24]. In our case, the value of an instability index of all protein was above 40 which indicate that all 12 proteins were unstable. The Extinction coefficients (EC) value of GH was calculated, which helpful in the protein- ligand and protein-protein interaction study. The pl values below 7 of all proteins were acidic in character. The GRAVY value or index for given proteins is calculated as the sum of the hydropathy values of all the aa divided by the number of residues in the protein sequence [29]. GRAVY index > 0 indicates a protein is hydrophilic and reveals hydrophobicity of the protein. All the GH protein sequences of 12 different fishes having hydrophilic in nature. We also analyzed the presence of cysteine residues; which form disulfide bonds, significant for protein folding and stability [26]. The earlier report predicted that, the GH amino acid sequence has around four cysteine residues with highly conserved positions in all vertebrates [27], though the residues:...
Further, in order to understand the structure level of GH protein, the modeling of protein of rohu was done by I-TASSER an automated protein structure prediction algorithm. The best 3D structure with high confidence score (C-score) was selected and used for further investigations. The top model of C-score of 0.46 was selected with estimated TM-score was 0.65±0.13 and estimated RMSD value 7.7±4.3Å (Fig 3). The selected model validated and Ramachandran plot showed good quality of model with most favored region 76.8% (Fig 4). The ProSA analysis showed that, Z score value, which indicates its quality assessment.

Fig 1: Comparison of GH peptide sequence of cyprinids and showing conserved residues among sequences found.

Fig 2: Phylogenetic tree and Cladogram for growth hormone (GH) constructed by the UPMGA and neighbor-joining method with 100 replicates of bootstrap.

Fig 3: Growth hormones of rohu consensus three dimensional structure generated using I-TASSER
The protein-protein interaction exploration is a comprehensive approach to identify the organization of desire proteome. The functional network of protein study will be helpful for drug discovery, to understand metabolic pathways and to predict or develop genotype-phenotype associations [38, 39]. In order to understand network of GH protein, we performed analysis with submitting FASTA sequences to STRING 9.1. In STRING, the functional interaction was analyzed by using confidence score. Interactions with score < 0.3 are considered as low confidence, scores ranging from 0.3 to 0.7 are classified as medium confidence and scores > 0.7 yield high confidence [40, 41]. The GH protein network showing functional association with 10 proteins such as GHRa, GHRb, prolactin receptor a, prolactin receptor b, Janus kinase 2, Janus kinase 2a, prolactin, signal transducer and activator of transcription 5.1, suppressor of cytokine signaling 2, and signal transduction and activation of transcription 3 with confidence score 0.975, 0.975, 0.953, 0.952, 0.943, 0.943, 0.932, 0.919, 0.913, and 0.911, respectively (Fig 5). We have depicted that, the receptor proteins are showing interacting more with the high confidence score as compare to other protein. The earlier study depicted, growth hormone gene structure is not conserved among teleosts, presenting a mammalian like organization (5 exons and 4 introns) in carps [42]. We could identify that, GH is more conserved at the protein level in fish species belonging to cyprinids. But, still questions remains, why their growth is differ from each other.

4. Conclusion
This is first comprehensive study on comparative as well as evolutionary study of growth hormone protein sequences among 12 different important fish species of cyprinids. In this study, we have investigated that most of physico-chemical properties were almost same in all the protein sequences. After performing multiple sequence alignment, we have depicted that “FRLLACFKKDMHKVETYLRVANCRRSLDSNCTL” is the conserved (identical) motif which present in all 12 protein sequences of GH. We also analyzed overall 100 hit from the NCBI protein database, which indicated the importance of this region among a diverse group of fish species with more than 91-100% identity. Moreover, protein-protein interaction pathway of this GH protein has helped us to understand the roles and associated proteins in various cellular pathways. We also generated 3D structure of GH protein using de novo modelling and subsequently validated using Ramachandran plot. The GH at protein level tends to be highly conserved across cyprinids. Thus finally, present work will support to understand more about GH proteins of fish species among cyprinids and their network.

5. Acknowledgment
We are thankful to Indian Council of Agriculture Research, New Delhi.

6. References
3. Filby AL, Tyler CR. Cloning and characterization of cDNAs for hormones and/or receptors of growth hormone, insulin-like growth factor-I, thyroid hormone, and corticosteroid and the gender-, tissue-, and developmental-specific expression of their mRNA transcripts in fathead minnow (Pimephales promelas). General and comparative endocrinology 2007; 150(1):151-63.


36. Vaz BS, Cerqueira GM, Silva JC, Manke VH, Moreira CG, Moreira HL. Sequence analysis of the growth hormone gene of the South American catfish Rhamdia


42. Hong Y, Schartl M. Sequence of the growth hormone (GH) gene from the silver carp (Hypophthalmichthys molitrix) and evolution of GH genes in vertebrates. Biochimica et biophysica acta 1993; 1174(3):285-8.