Computational interpretation and comparative annotation of lipocalin-2 protein having the candidate for antibacterial agent

Neelesh Sharma1,2, Mrinmoy Ghosh1, Sung Jin Lee3* and Dong-Kee Jeong1*

1Department of Animal Biotechnology, Faculty of Biotechnology, Jeju National University, Jeju 690 756, Republic of Korea
2Division of Veterinary Medicine, Faculty of Veterinary Sciences and Animal Husbandry
Sher-e-Kashmir University of Agricultural Sciences and Technology, Jammu 181 102, India
3College of Animal Life Sciences, Kangwon National University, Chunchon 200 600, Republic of Korea

Received 23 October 2014; revised 17 May 2015; accepted 29 May 2015

Recent studies have demonstrated that lipocalin-2 (LCN2) binds specifically with enterobactin to resist the bacterial infection in mammalian cells. The present study focuses on three key aspects, viz., diversity of bovine LCN2 in content with other mammalian groups, its structural variation and docking of bovine LCN2 with the scaffold proteins. Twenty nine sequences of LCN2 were selected for comparative annotation study using computational server and bioinformatic tools. Significant variations were found in physico-chemical properties and diversity in antigen binding sites between organisms. The homology model of the LCN2 from human and bovine illustrated the conformation of the best homology model. An attempt was made to determine binding affinities of bovine LCN2 with 1ZOK and 1PN5 through molecular modeling and docking. The aim of the study was not only to do comparative analysis, but also to find out how far LCN2 in bovine could be used as model for LCN2 proteins from other mammals. This may provide framework to develop the promising novel class of biopharmaceutical agents.

Keywords: Bacteriostatic agent, docking, homologus modeling, lipocalin, motif, siderophores

Introduction

Bovine mastitis is the most prevalent disease affecting dairy cow worldwide1. Economically, it is considered the major cause of the financial recession in dairy industry. The onset of mastitis is associated with different bacterial infections, mostly *Staphylococcus aureus* and *Escherichia coli*2. It has been suggested that intramammary infection with *E. coli* often results in acute mastitis with severe clinical manifestations, whereas symptoms induced by *S. aureus* infection are usually less severe3. The current treatment strategies involve the practices of hygiene at the farm and administration of the antibiotics. However, some of the infections are asymptomatic and, in such cases, the use of antibiotics needs to be minimized3.

Antibiotic resistance is a worldwide problem. Indiscriminate use of antibiotics in cattle for mastitis treatment is a main cause of antibiotic resistance in bacteria4,5. The surveillance data has shown that resistance in *E. coli* is consistently higher for those antimicrobial agents that have been in use for the longest time in humans and in veterinary medicine6. The significant increase in the emergence and spread of multidrug-resistant to newer compounds is caused by the genetic diversity within *E. coli* that promotes the acquisition of virulence. This antibacterial resistance pattern demands urgent need to find out some alternative antibacterial proteins or peptides to manage bacterial infections in livestock and humans.

Antibacterial peptides are an important component of the innate defenses of all living species7. Some antibacterial peptides show both antifungal and antiviral activity, and are often referred to as ‘antimicrobial peptides’ (AMPs). These peptides are potent, broad-spectrum antibiotics that demonstrate a potential as novel therapeutic agents. Some of the antimicrobial peptides have multifunctional properties, and an attempt has been made to predict this property from the hydrophobicity of all amino acid side chains. Because of broad range of activity, lesser toxicity and decreased resistance development by the target cells, the AMPs have been considered as a promising and potential drug candidate for future. More than 2,000 AMPs have been reported in the

*Author for Correspondence:
Tel: +82-64-754-3331; Fax: +82-64-725-2403
dkjeong@jejunu.ac.kr; sjlee@kangwon.ac.kr

*For Supplementary Data, see website:
www.nopr.niscair.res.in, or www.niscair.res.in
antimicrobial peptide database. It has been well established that many of AMPs have an excellent direct antimicrobial activity and thus provide templates for the commercial development of a new generation of antibiotics.

In mammals, a siderocalin called as lipocalin-2 (LCN2) has the ability to bind specifically with enterobactin (a type of the bacterial siderocalin). The bacterial cells contain enterobactin, the small molecules of siderophores, having high affinity for ferric iron, much of which is acquired from the host by scavenging iron. For survival during infection, the bacteria secrete these siderophores, which sequester iron from the host proteins. LCN2 inhibits bacterial growth by binding with bacterial siderophores and sequestering iron. In response to bacterial infection, innate immune cells over express and secrete lipocalin, which can resist the bacterial growth by preventing enterobactin-mediated iron acquisition. Thus, it acts as a siderocalin and could constitute an innate immune defense during bacterial infection.

The sequence-structure problem of LCN2 is yet to be resolved completely and, therefore, the number of known unique structural folds a particular protein adopts is limited. The major drawbacks of wet lab experimental or molecular biophysical methods, which are to be used to characterize the proteins of various organisms, are time consuming and expensive. Moreover these methods are not amendable by high throughput techniques. The computational applications, indeed, could help to sort out these barriers. A number of computational tools have been developed for making structural predictions and to study the physico-chemical properties of the proteins.

The present report consists of two parts based on the computational approaches. The initial study used various in silico tools to study the structural and functional diversity of LCN2, and in the second phase, protein-protein interaction and docking studies were performed. In addition, the report also focuses on the antigenic binding site of bovine LCN2 (bLCN2) and its efficiency towards the MHC-I and -II binding.

**Materials and Methods**

**Retrieval of LCN2 Protein Sequences**

Twenty nine LCN2 protein sequences were retrieved from National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). The search results yielded LCN2 sequences from various species, viz., Mammalian, Actinopterygii, Arachnida, Reptilia, Chondrichthyas, Amphibia and Aves (Table 1). The retrieved sequences were randomly chosen as data sets to understand the diversity in LCN2 protein family and to compare the physicochemical properties of the said protein.

<table>
<thead>
<tr>
<th>Organisms</th>
<th>Species</th>
<th>Acc. no.</th>
<th>Taxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Homo sapiens</td>
<td>gi 49457137</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Cow</td>
<td>Bos taurus</td>
<td>gi 359070499</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Pig</td>
<td>Sus scrofa</td>
<td>gi346986374</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Sheep</td>
<td>Ovis aries</td>
<td>gi42622980</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Dog</td>
<td>Canis lupus familiaris</td>
<td>XP_862322.1</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Rhesus monkey</td>
<td>Macaca mulatta</td>
<td>gi 384943872</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Golden hamster</td>
<td>Mesocricetus auratus</td>
<td>gi524975816</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Gorilla</td>
<td>Gorilla gorilla gorilla</td>
<td>gi 42636183</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Mouse</td>
<td>Mus musculus</td>
<td>gi 124297609</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Rat</td>
<td>Rattus norvegicus</td>
<td>gi 57870616</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Tasmanian devil</td>
<td>Sarcophilus harrisi</td>
<td>gi 39506474</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>Pan troglodytes</td>
<td>gi 55632379</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Squirrel monkey</td>
<td>Saimiri bolivensis bolivensis</td>
<td>gi 403299735</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Common marmoset</td>
<td>Callithrix jacchus</td>
<td>gi 390458396</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Sumatran orangutan</td>
<td>Pongo abelii</td>
<td>gi 297685446</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Degu</td>
<td>Octodon degus</td>
<td>gi 507685901</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Cat</td>
<td>Felis catus</td>
<td>gi 410979338</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Guinea pig</td>
<td>Cavia porcellus</td>
<td>gi 348569741</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Tick</td>
<td>Amblyomma maculatum</td>
<td>gi 346473553</td>
<td>Arachnida</td>
</tr>
<tr>
<td>Rock dove</td>
<td>Columba livia</td>
<td>gi 449266894</td>
<td>Aves</td>
</tr>
<tr>
<td>Chicken</td>
<td>Gallus gallus</td>
<td>gi 19908346</td>
<td>Aves</td>
</tr>
<tr>
<td>Mallard</td>
<td>Anas platyrhynchos</td>
<td>gi 483518920</td>
<td>Aves</td>
</tr>
<tr>
<td>Turkey</td>
<td>Meleagris gallopavo</td>
<td>gi 326930139</td>
<td>Aves</td>
</tr>
<tr>
<td>Zebra finch</td>
<td>Taeniopygia guttata</td>
<td>gi 449477689</td>
<td>Aves</td>
</tr>
<tr>
<td>Salmon fish</td>
<td>Salmo salar</td>
<td>gi 209734482</td>
<td>Actinopterygi</td>
</tr>
<tr>
<td>Zebra mbuna fish</td>
<td>Maylandia zebra</td>
<td>gi 499003266</td>
<td>Actinopterygi</td>
</tr>
<tr>
<td>Frog</td>
<td>Hyla japonica</td>
<td>gi 24899231</td>
<td>Amphibia</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Drosophila melanogaster</td>
<td>gi 9181925</td>
<td>Insecta</td>
</tr>
<tr>
<td>Viper</td>
<td>Azemiops feae</td>
<td>gi 384110780</td>
<td>Reptilia</td>
</tr>
</tbody>
</table>
The multiple sequence alignment study was carried out by using the ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2), MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/) and T-Coffee (http://tcoffee.vital-it.ch/apps/tcoffee/do:regular) programs. These program were preferred owing to their advanced features to evaluate the quality of the alignments and also have the capacity to identify the occurrence of motifs. At the time of alignment, the insertion gaps were allowed in the region of final alignment without affecting the secondary structure.

Phylogenetic Diversity of LCN2 and Search for Motif
Phylogenetic analysis based on protein sequences was carried out using the maximum-likelihood method. The Phylogeny.fr program was used for the construction of phylogenetic tree following neighbor-joining method. Phylogenetic analysis of sequence data was performed using the PHYLIP software package.11

Analysis of domain was done by Multiple EM for Motif Elicitation (MEME) (http://meme.nbcr.net/meme/cgi-bin/meme.cgi). The MEME suite web server provides a unified portal for online discovery and analysis of sequence motifs representing features, such as, DNA binding sites and protein interaction domains. Once a collection of motifs was found, additional analyses were performed to further characterize those motifs. Motif Alignment and Search Tool (MAST) (http://meme.nbcr.net/meme/cgi-bin/mast.cgi) was used for searching biological sequence databases for sequences that contain one or more of a group of known motifs. The output from MAST is a list of sequences, for which the E-value is less than a user-specified threshold. MAST works by calculating match scores for each sequence in the database compared with each of the motifs in the group of provided motifs.

Computational Tools Used for Analysis of Physico-chemical Properties
Computational tools were applied to predict the physico-chemical properties of lipocalin sequences of the LCN2 from the selected species. Protein Calculator v3.3 tool was applied to calculate the average number of atoms present in the residues and charge of the protein molecules at pH 7. The number of residues and their frequency was checked by CLC free Workbench (version 6.8.2). Moreover, the theoretical isoelectric point (pI) value, number of negatively (NCR) and positively charged residues (PCR), extinction coefficient (EC), pI, instability index (II), alphabetic index (AI) and grand average of hydropathicity (GRAVY) were predicated by ExPaSy’s ProtParam program (http://web.expasy.org/protparam/). The instability index provides an estimate of the stability of the protein. A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable.

The absorbance of a protein at 280 nm depends on the content of Trp, Tyr and cystine (disulfide bonds). The molar extinction coefficient of lipocalin proteins from various species was calculated using the following equation:

\[ E = Numb(Tyr) \times \text{Ext}(Tyr) + Numb(Trp) \times \text{Ext}(Trp) + Numb(Cystine) \times \text{Ext}(Cystine) \] … (1)

Where (for proteins in water measured at 280 nm):

\[ \text{Ext}(Tyr) = 1490, \text{Ext}(Trp) = 5500, \text{Ext}(Cystine) = 125 \]

The AI of the protein is defined as the relative volume occupied by aliphatic side chains, i.e., alanine, valine, iso-leucine and leucine. The aliphatic index of a protein was calculated according to the following formula proposed by Ikai et al.12:

\[ \text{AI} = X(\text{Ala}) + a \times X(\text{Val}) + b \times (X(\text{Ile}) + X(\text{Leu})) \] … (2)

Where, X(Ala), X(Val), X(Ile) and X(Leu) are mole percent of alanine, valine, iso-leucine and leucine. The coefficients a and b are mole percent (100 × mole fraction) of alanine, valine, iso-leucine and leucine. The coefficients a and b are the relative volume of valine side chain (a = 2.9) and of Leu/Ile side chains (b = 3.9) to the side chain of alanine.

Local hydrophobicity is found to play a dominant role in the stabilization of secondary structures in proteins. According to Bhattacharjee and Biswas,13 the average hydrophobicity (H) of the i-th amino acid residues at position in β-strand is calculated as:

\[ H_i = \sum_{j=1}^{20} \Delta G_{ij}^{\text{corr}} \frac{n_{ij}}{n_j} \] … (3)

\( \Delta G_{ij}^{\text{corr}} \) experimentally measured free energy change, where \( n_{ij} \) and \( n_j \) are the number of i-th residues at \( j \)-th position and the number of total residues at \( j \)-th position, respectively.

Prediction of Antigenic Binding Site in Lipocalin Residues
Antigenic predication tools could help to predict the potential antigenic sites of the protein sequence. The residue with antigenic sites were estimated by EMBOSS antigenic tool following the method proposed by Kolaskar and Tongaonkar14 and the plots were drawn by JaMBW 3.1.7 (http://www.bioinformatics.org/jambw/3/1/7/) computational program14.
Classical major histocompatibility complex (MHC) molecules play a pivotal role in the adaptive immunity mediated by the T cells. Peptide binding to MHC is a prerequisite for T cell recognition and is, in addition, the event that most selectively defines T cell epitopes\(^\text{15}\). Position-specific scoring matrix (PSSM) derived from a set of aligned peptides known to bind to a given MHC molecule can be used as a predictor of both peptide-MHC bindings. The MHC binding sites for bLCN2 were predicted. For the determination of potential MHC binders, an elegant machine learning technique SVM was applied.

**Prediction of Secondary Structure and Subcellular Localization of the Human (hLCN2) and Bovine LCN2 (bLCN2)**

The secondary structures of these proteins were predicted with FASTA sequences using SOPMA (psa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=NPSA/npsa_sopma.html). This method calculates the content of α-helix, β-sheets, turns, random coils and extended strands. Prediction of protein localization can provide important clues to function and can help to identify drug and vaccine targets against pathogenic bacteria. In principle, the localization site of proteins is determined by signals in their amino acid sequence, which in some cases can be recognized without structural information. CELLO combines SVM-based prediction and homology search to predict localization for eukaryotes and prokaryotes also. The disulphide bridges (SS bonds) in all the sequences were analyzed by identifying the position of cysteines in primary structure. The CYS_REC server (http://www.softberry.com/berri.phtml?topic=cys_rec&group=programs&subgroup=prote) was used to identify the total number of cysteines present and to predict the most probable SS bond pattern of pairs in the protein sequence.

**Homologous Modeling of hLCN2 and bLCN2 and Validation of Model**

An attempt was made to find a suitable template protein for the modeling of the target protein. The template protein was searched through GeneThrader and selected for 3D modeling of the target protein. The 3D structure of targeted LCN2 was generated by homology modeling using MODELLER 9v8 program. The Swiss PDB viewer was used to visualize and refine the models, and PyMOL was used to generate publishable images of the PEPCK models. Backbone conformation was evaluated by the inspection of the Psi/Phi Ramachandran plot obtained from PROCHECK analysis\(^\text{16}\).

**Prediction of Active Binding Site in bLCN2 and hLCN2 Proteins**

Q-SiteFinder is a new energy-based method for predicting protein-ligand binding sites. Q-SiteFinder has been designed to meet two main requirements. First, it is intended for identification of ligand binding sites for virtual screening and de novo drug design. Second, protein residues within a suitable range of the probe clusters can be identified, which may be used for functional site identification and comparison. The predicted 3D structure of the LCN2 was scanned against all the proteins of known 3D structure in the PDB and the details of the possible ligands with the binding residues were analyzed. In the present annotation, the iron binding site and the responsible residues were predicted.

**Protein-Protein Docking Study of bLCN2**

The docking was performed using AutoDock 4.2.5 software with the reference of the template complex with the two scaffold protein molecule (PDB i.d., 1ZOK & 1PN5) for the docking with the bLCN2 protein. Docking simulations were run using Lamarckian Genetic Algorithm (LGA) with the grid points of 44 × 44 × 40 Å and the active site residues at the center of the grid box. The docking parameters were set at LGA for 100 runs. The energy evaluations were set to 27,000 and 1,500,000 generations. Population size was set to 150 and the rate of gene mutation, and the rate of gene crossover were set to 0.02 and 0.8, respectively. Then, obtained confirmations were summarized, collected and extracted using AutoDock tool. Before docking, the form of water molecules and bound ligands were removed. The electrostatic potential calculation, model visualization and image generation were performed by PyMOL software.

**Results**

**Multiple Sequence Alignment and Phylogeny Relation**

The present study was carried out for in silico analysis of LCN2 protein. The multiple sequence alignment of selected protein sequences revealed a stretch of protein with homology (Fig. 1). It was found that *Homo sapiens* lipocalins are closely spaced with *Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii* and *Macaca mulatta*. Seven clades were observed in the phylogenetic tree. The maximum similarity of *Bos taurus* was found with *Ovis aries* (share 97% of similarity), whereas *H. sapiens* and *G. gorilla* showed 82% of the similarity in LCN2 gene. However,
Fig. 1—Multiple sequence alignment of lipocalin proteins from the selected group of organisms.
Azemiops feae was represented in a separate clade. Overall, it was observed that LCN2 genes had similarity in their sequences, but variations were there among the different taxon. The amino acid sequences of LCN2 were quite divergent and showed low levels of sequence identity, exhibiting great structural and functional diversity, both within and between species\(^{17-19}\).

To find out the evolutionary relationship between the different taxon sources, the protein sequences belonging to lipocalin family were subjected to phylogenetic tree construction. The phylogenetic tree showed five distinct clusters (Fig. 2). The multiple sequence and the cladogram illustrated high degree of divergence between the species for LCN2 protein.
Motif Sequences of LCN2 Identified Using MEME Tools
The motif analyses among the sequences were carried out to know the conserved domain for the functional annotation of proteins. The size of the lipocalin family has grown significantly to encompass a large corpus of protein sequences, which overall showed the less number of conserved regions. The motif were analysed by MEME (Multiple EM for Motif Elicitation). The principle of MEME is based on by searching for repeated, ungapped sequence patterns that are present in protein sequences. In the present study, three motif regions were found in LCN2 sequences (Table 2) and these findings are in agreement of earlier report 20.

Annotation of the Physico-chemical Properties
The comparative study of the twenty nine LCN2 proteins showed massive differences in their physico-chemical properties. The computed pI value higher than 7 indicates the base characteristic of protein. In the present study, LCN2 analysis observed pI >7, which demonstrates base nature of LCN2 protein. The computed extinction coefficient (ECs) of proteins ranged from 11710 to 36900 M⁻¹ cm⁻¹. The high EC of LCN2 indicates presence of high concentration of Cys, Trp and Tyr. The negative GRAVY values of LCN2 indicate the possibility of better interaction with water (Table 3). The structural analysis found low hydrophobic residue contents in LCN2 (8Suppl Tables 1 & 2). The net hydrophobic residues in hLCN2 and bLCN2 were found to be 48.48 and

<table>
<thead>
<tr>
<th>Organisms</th>
<th>Types and number of atoms</th>
<th>MW (kDa)</th>
<th>pI</th>
<th>Charge at pH 7</th>
<th>C</th>
<th>H</th>
<th>N</th>
<th>O</th>
<th>S</th>
<th>EC</th>
<th>AI</th>
<th>II</th>
<th>GRAVY</th>
<th>NCR</th>
<th>PCR</th>
<th>Organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>O. degus</td>
<td></td>
<td>16799.2</td>
<td>9.19</td>
<td>5.1</td>
<td>757</td>
<td>1176</td>
<td>204</td>
<td>219</td>
<td>5</td>
<td>17420</td>
<td>80.98</td>
<td>12.46</td>
<td>-0.388</td>
<td>14</td>
<td>18</td>
<td>O. degus</td>
</tr>
<tr>
<td>F. catus</td>
<td></td>
<td>16787.2</td>
<td>9.28</td>
<td>6.1</td>
<td>780</td>
<td>1165</td>
<td>198</td>
<td>211</td>
<td>3</td>
<td>36900</td>
<td>76.81</td>
<td>25.57</td>
<td>-0.385</td>
<td>12</td>
<td>17</td>
<td>F. catus</td>
</tr>
<tr>
<td>C. Porcellus</td>
<td></td>
<td>16732.2</td>
<td>9.19</td>
<td>5.6</td>
<td>758</td>
<td>1163</td>
<td>197</td>
<td>215</td>
<td>8</td>
<td>22015</td>
<td>75.42</td>
<td>25.00</td>
<td>-0.345</td>
<td>13</td>
<td>18</td>
<td>C. Porcellus</td>
</tr>
<tr>
<td>A. Maculatum</td>
<td></td>
<td>22604.5</td>
<td>8.16</td>
<td>2.7</td>
<td>1002</td>
<td>1545</td>
<td>271</td>
<td>304</td>
<td>11</td>
<td>25815</td>
<td>66.87</td>
<td>48.62</td>
<td>-0.415</td>
<td>22</td>
<td>24</td>
<td>A. Maculatum</td>
</tr>
<tr>
<td>C. livia</td>
<td></td>
<td>19669.33</td>
<td>8.57</td>
<td>3.6</td>
<td>868</td>
<td>1369</td>
<td>237</td>
<td>266</td>
<td>9</td>
<td>27055</td>
<td>70.56</td>
<td>56.07</td>
<td>-0.441</td>
<td>16</td>
<td>19</td>
<td>C. livia</td>
</tr>
<tr>
<td>G. gallus</td>
<td></td>
<td>16473.52</td>
<td>7.28</td>
<td>0.4</td>
<td>725</td>
<td>1135</td>
<td>203</td>
<td>229</td>
<td>6</td>
<td>27055</td>
<td>67.71</td>
<td>39.54</td>
<td>-0.720</td>
<td>18</td>
<td>18</td>
<td>G. gallus</td>
</tr>
<tr>
<td>A. platyrynchos</td>
<td></td>
<td>16473.51</td>
<td>7.94</td>
<td>1.4</td>
<td>727</td>
<td>1145</td>
<td>195</td>
<td>231</td>
<td>5</td>
<td>27055</td>
<td>68.33</td>
<td>40.14</td>
<td>-0.677</td>
<td>18</td>
<td>19</td>
<td>A. platyrynchos</td>
</tr>
<tr>
<td>M. gallopavo</td>
<td></td>
<td>20173.76</td>
<td>6.77</td>
<td>-0.4</td>
<td>886</td>
<td>1404</td>
<td>245</td>
<td>277</td>
<td>8</td>
<td>27055</td>
<td>78.93</td>
<td>38.60</td>
<td>-0.496</td>
<td>20</td>
<td>19</td>
<td>M. gallopavo</td>
</tr>
<tr>
<td>T. guttata</td>
<td></td>
<td>16470.44</td>
<td>7.98</td>
<td>1.6</td>
<td>722</td>
<td>1132</td>
<td>200</td>
<td>229</td>
<td>6</td>
<td>27055</td>
<td>66.32</td>
<td>41.50</td>
<td>-0.635</td>
<td>15</td>
<td>16</td>
<td>T. guttata</td>
</tr>
<tr>
<td>S. salar</td>
<td></td>
<td>16118.27</td>
<td>8.95</td>
<td>3.9</td>
<td>723</td>
<td>1116</td>
<td>198</td>
<td>211</td>
<td>5</td>
<td>30940</td>
<td>77.83</td>
<td>31.72</td>
<td>-0.386</td>
<td>14</td>
<td>17</td>
<td>S. salar</td>
</tr>
<tr>
<td>M. zebra</td>
<td></td>
<td>15437.30</td>
<td>4.87</td>
<td>-5.6</td>
<td>698</td>
<td>174</td>
<td>217</td>
<td>2</td>
<td>1072</td>
<td>24075</td>
<td>87.79</td>
<td>57.92</td>
<td>-0.158</td>
<td>16</td>
<td>10</td>
<td>M. zebra</td>
</tr>
<tr>
<td>H. japonica</td>
<td></td>
<td>15496.60</td>
<td>8.73</td>
<td>3.6</td>
<td>682</td>
<td>1097</td>
<td>189</td>
<td>212</td>
<td>5</td>
<td>13075</td>
<td>76.13</td>
<td>48.41</td>
<td>-0.605</td>
<td>17</td>
<td>20</td>
<td>H. japonica</td>
</tr>
<tr>
<td>M. melanogaster</td>
<td></td>
<td>17739.1</td>
<td>6.39</td>
<td>0.4</td>
<td>804</td>
<td>1233</td>
<td>213</td>
<td>235</td>
<td>3</td>
<td>28545</td>
<td>86.77</td>
<td>44.43</td>
<td>-0.308</td>
<td>18</td>
<td>18</td>
<td>M. melanogaster</td>
</tr>
<tr>
<td>A. feae</td>
<td></td>
<td>15706.03</td>
<td>5.17</td>
<td>-4.7</td>
<td>698</td>
<td>1108</td>
<td>178</td>
<td>214</td>
<td>9</td>
<td>11710</td>
<td>86.64</td>
<td>46.53</td>
<td>-0.229</td>
<td>21</td>
<td>21</td>
<td>A. feae</td>
</tr>
</tbody>
</table>

Fig. 2—The nonlined up cladogram phylogenetic tree of the LCN2 protein from the selected species analyzed by NJ method. The branch length values are displayed on the node.
45.73%, respectively. hLCN2 have higher percentage of charged, acidic amino acids and lower percentage of aliphatic and aromatic amino acids than bLCN2. Hence, hLCN2 is relatively more hydrophobic and more soluble in water than bLCN2.

**Prediction of Antigenic Binding Site and Efficiency towards MHC-I and -II Binding**

Five antigenic sites were determined in hLCN2, and four antigenic sites in bLCN2. The highest peak was found at start position 118 to 132; whereas in bLCN2, 110 to 122 residues showed maximum length of the antigenic site ([Suppl Table 3]). Moreover, the comparative study shows that the site of antigenicity varied between the species ([Suppl Fig. 1]).

The prediction on binding sites was based on support vector machine using amino acid sequences. In this test, we predicated the MHC-I and MHC-II binding regions of bLCN2 ([Suppl Tables 4 & 5]). SVM-based immunodominace classifiers trained on residue properties and amino acid sequence were able to discriminate the immunogenic peptides from non-immunogenic peptides with an accuracy of 60% and threshold of 0.5.

In the present study, we have predicted MHC-I binding peptides for 8mer_H2_Db allele, the optimal score (OS) was found 18.847 in the peptide sequence of SSQPGQFT; 9mer_H2_Db allele, OS 18.194 in the peptide sequence of NTDYNQFAI; 10mer_H2_Db allele, OS 20.09 in the sequence of ADQFQGKWT; and 11mer_H2_Db allele, OS 19.765 in the sequence of VRVVNVTDNQF. We have also predicted the SVM based MHCII-IAb (OS 12.165); MHCII-IAd (OS 13.309); and MHCII-IAg7 (OS 15.278), which show potential binders from bLCN2 protein.

**Analysis of Secondary Structure Variations and Predication of 3D Homologous Structure of bLCN2 and hLCN2**

The comparative study between secondary structures of bLCN2 and hLCN2 was carried out and results showed the key differences between both of them (Table 4). The maximum percentage of α-helix was found in hLCN2, while maximum percentage of β-turn was observed in bLCN2.

The target protein sequences were obtained from the NCBI protein database. And the suitable template protein structure of the target was selected by performing protein sequence similarity search against 3D structure from PDB. The blast was used for protein sequence similarity search. Further, the Swiss-Model server was used to model protein structure by using 1X89C and 1dfvB as the templates for bLCN2 and hLCN2, respectively (Figs 3A & B).

**Table 4—Physical properties of secondary structure of hLCN2 and bLCN2**

<table>
<thead>
<tr>
<th>Secondary key features</th>
<th>hLCN2</th>
<th>bLCN2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha helix (Hh)</td>
<td>20.00%</td>
<td>17.81%</td>
</tr>
<tr>
<td>Extended strand (Ee)</td>
<td>30.32%</td>
<td>34.93%</td>
</tr>
<tr>
<td>Beta turn (Tt)</td>
<td>3.87%</td>
<td>6.16%</td>
</tr>
<tr>
<td>Random coil (Cc)</td>
<td>45.81%</td>
<td>41.10%</td>
</tr>
</tbody>
</table>

Evaluation of model quality is a crucial step in homology modeling. The target protein structures were validated through ProCheck using Ramachandran plots. The darkest areas correspond to the “core” regions and represented the most favorable combinations of phi-psi values. The DOPE (Discrete Optimized Protein Energy) score values for all the models of target were generated. Among all, the model having the lowest DOPE score was considered as the final model. Ideally, one would hope to have over 90% of the residues in these “core” regions. It was observed that 93% of residues were in the favored region in hLCN2 and 90% residues in bLCN2 model. The predicted structures were well-conformed and indicated reasonably good quality. The Z-score of the bLCN2 was found to be 1.19 (Figs 4A & B), whereas hLCN2 model had a score of 0.42 (Figs 5A & B).

**Computational Annotation of Ligand Binding Site (LBS) and Docking Study of bLCN2**

Computational method was used to find out the LBS of the protein of interest. Identifying the location of LBSs on a protein has important applications including molecular docking, de novo drug design and structural identification and comparison of functional sites. The Q-SiteFinder depicts 10 major binding sites...
of the protein (Table 5). The LBS of bLCN2 has shown that the Gly at the C-terminal efficiently bound with the iron. Hence binding with bacterial siderophores and sequestering iron, LCN2 could resist the bacterial proliferation.

Studies were also conducted to generate high accuracy structures of protein-protein complexes from the structures of unbound component proteins for docking. The two scaffold proteins, viz., DLG1 and NLRP (PDB i.d.: 1ZOK & 1PN5, respectively) were used for binding with the predicated 3D model of bLCN2. The selected 1ZOK and 1PN5 proteins from PDB were associated with the assembly and localization of T-Cell Receptor (TCR) signaling molecules, and activation of p38 and assembly of the inflammasome, respectively. The best model was chosen based on the best dockpose. After docking the RSMD deviation, the models had deviated from 1.15 to 1.18 Å and 1.11 to 1.14 Å, respectively. The electrostatic, van der Waals, and interaction energies were observed to be −324.2, −75.8, and −942.4 kcal mol⁻¹ at DLG1-bLCN2 complex, whereas corresponding figures were −424.5, −80.1,
Table 5—Analysis of ligands binding site with predicates by Q site finder

<table>
<thead>
<tr>
<th>Ligand binding sites</th>
<th>Residues</th>
<th>Site volume in cubic Å</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASN 59, ALA 60, ILE 61, TYR 72, SER 88, LEU 90, TRP 99, ARG 101, TYR 126, PHE 143, LYS 144, LYS 145, GLN 147, TYR 152, PHE 153, LYS 154</td>
<td>326</td>
</tr>
<tr>
<td></td>
<td>VAL 54, GLY 55, MET 71, LEU 157, LYS 162, GLU 163, LEU 164, THR 165, ARG 169, PHE 188, VAL 190, PRO 191</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>MET 1, SER 2, VAL 3, GLY 4, LEU 5, LEU 6, TRP 7, LEU 8, GLY 9, PHE 10, THR 11.</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>TRP 99, ARG 101, PHE 103, LEU 114, TYR 120, ILE 123, TYR 126, LYS 145, GLN 147</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>LYS 50, TRP 51, SER 73, SER 74, ASN 75, THR 87, SER 88, ILE 89, TYR 98, CYS 195, ILE 196, ASP 198, GLN 199</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>ASN 132, PHE 138, ALA 139, ILE 140, LEU 157, VAL 168, ASN 171, PHE 172, PHE 175</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>PHE 42, ALA 44, LEU 78, GLY 82, TYR 84, PRO 105, SER 106, SER 107, GLN 108, PRO 109, GLY 110, GLN 111, PHE 112</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>PHE 42, PHE 47, TYR 76, TYR 84, PHE 103, PHE 112, VAL 128, VAL 130, VAL 141, PHE 143, TYR 158</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>THR 53, VAL 56, ILE 61, LEU 69, LYS 70, MET 71, TYR 72, LEU 90, VAL 187, THR 189</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>ASP 134, GLN 137, PHE 138, LEU 164, THR 165, GLU 167, VAL 168</td>
<td>84</td>
</tr>
</tbody>
</table>
and ~824.6 kcal mol\(^{-1}\) between NLRP-bLCN2 complex. Thus, it was understandable that the mode of protein binding with the respective binding site was feasible and correct. The protein-protein docking study would be the framework for the further analysis of the pattern of coexpression of bLCN2 on the antibacterial and anti-inflammatory response.

**Discussion**

Lipocalins, a functionally diverse group of proteins, are widely found in multiple phyla\(^9\). It has shared a conserved structure even in the absence of significant sequence conservation\(^{19}\). A total of twenty nine full-length protein sequences of lipocalin from different source organisms were considered in the present analysis. The comparison of LCN2 sequences of the selected organisms reports sharing of three main distinct conserved regions despite differences in physico-chemical properties of the expressed proteins. The arrangements of three conserved motifs in LCN2 were also reported in previous reports\(^{10,20}\). The phylogenic analysis among the lipocalin clades, motif arrangements and exons distribution emphasized our finding that lipocalin protein family depicts the diversity among and between the taxon. Our results thus further strengthen the hypothesis of genomic fusion between eukaryotes and inheritance of prokaryotic ancestral lipocalin protein family\(^{21}\).

In silico in-depth physico-chemical characterization of the lipocalin was carried out to understand the functional variation in lipocalins. The results revealed the highest number of Cys, Trp and Tyr in bovine LCN2. The disulfide bridge was directly proportional to the number of residues between the linked cysteines and played an important role in conferring protein rigidity and ligand affinity in LCN2\(^{22,23}\). Trp contributed to the hydrophobicity of LCN2, while Tyr have indicated that the beta bulges region comprised of polar residues closer to the opening of the barrel, suggesting its contribution to the ligand binding site\(^{24}\). In silico studies also identified the iron binding site in the LCN2 protein and supported earlier observation that the presence of Gly at the C-terminal of residue can efficiently bind with the iron and support favorable conformation\(^{25}\). The random coils suggested that LCN2 has an unstable secondary structure with poor hydrophobic packing, which may contribute to poor thermo stability of LCN2 protein and thereby supports earlier observation\(^{26}\).

Prediction of protein surface regions that can be preferentially recognized by antigenic epitopes can help in the design of vaccine components and immuno-diagnostic reagents\(^{27}\). Certain similarities between the bovine and the human LCN2 allow us to use bovine LCN2 as a model to identify the MHC-I and MHC-II binding sites as potential therapeutic and/or diagnostic reagents in several disease areas. The available literature suggests that peptides could be chemically conjugated to a large carrier protein that might improve the immunogenicity and would require further verification and experimental analysis.

The computed protein concentration and extinction coefficients may help in the quantitative study of protein-protein in solution. We found that bovine LCN2 is co-associated with DLG1 and NLRP\(^{28}\). Our docking study could help in better understanding of the co-expression of the LCN2 protein with other antibacterial proteins\(^{10}\). DLG1 protein is associated with immune cells and is important for T-cell activation in the periphery, whereas the large NLRP family proteins are involved in innate immunity.

**Conclusion**

The important features of LCN2 were investigated and analyzed by computational methods. LCN2 is a complex family of proteins and has shown the diversity in their properties. The valid and stable 3D model of LCN2 designed from bovine could further be used in in silico studies to evaluate the structural stability and also to provide a better understanding of topological parameters of the enzyme and molecular basis of antigenicity. The LCN2 binding efficiency holds promise for improving our understanding of ligand binding sites. The docking analysis elucidates that LCN2 exhibited antibacterial property via iron binding activity. The annotation of bovine LCN2 with functionally associated proteins updates our understanding on structural and functional uniqueness of bovine LCN2 as compared to other species. Thus, association of complex LCN2 holds promise for generating novel therapeutic agents.

**Conflict of Interest**

The authors confirm that this article content has no conflicts of interest.

**Acknowledgement**

The present research work was supported by a grant from the Next-Generation Bio Green 21 Program.
References