RESEARCH ARTICLE **(Open Access)**

In silico analysis of osteocalcin gene in Cyprinidae

XHILIOLA BIXHEKU[1], ANILA HODA[2*], LORENA HYSI[2],VILSON BOZGO[2]

[1]Quality Assurance Agency for Higher Education,

[2]Agricultural University of Tirana,

*Corresponding author email: ahoda@ubt.edu.al;

Abstract

Cyprinidae family includes a large number of fish species, where the most popular are zebrafish (*Danio rerio*), grass carp (*Ctenopharyngodon idella*) and common carp (*Cyprinus carpio*). In this study, osteocalcin in *Cyprinus carpio* was investigated, regarding physicochemical characteristics, structural properties using several available bioinformatic tools. Sequences were retrieved from Genbank. Sequence alignment of the gene and protein sequence were done with ClustalW. The 3D structure of protein was predicted by SWISS-Model softare andwas checked by Z-Score using Qmean server, ERRAT, and Rampage Ramachandran plot analysis. The information provided here is a theoretical overview that will help to get an idea about the predicted protein structure.

Keywords: osteocalcin, bioinformatic tools, sequence alignment

Introduction

The Cyprinidae is the second largest fish family in the world and one of the most widespread in freshwater. [1]. Cyprinidae family includes a large number of fish species, where the most popular are zebrafish (*Danio rerio*), grass carp (*Ctenopharyngodon idella*) and common carp (*Cyprinus carpio*). In this study, osteocalcin in *Cyprinus carpio* was investigated by the use of bioinformatic tools. Osteocalcin (bone Gla protein) is an extracellular matrix protein synthesized by osteoblasts that is a marker of bone. Nishimoto et al., 1995 [2] have characterized osteocalcin from Cyprinus carpio for aminoacid sequence and extent of secondary structure. Carp osteocalcin is a polypeptide of 45 amino acids and an abundant component of carp rib bone comprising over 35% of the total extractable proteins [2]. Actually a large number of computational tools are available from different sources and help researchers to analyze the properties of different proteins. The aim of this study is to select homologous sequences related to bone Gla protein [Cyprinus carpio 'color'] (AIT51847.1) and analyze the physico-chemicals properties of selected proteins by *in silico* methods.

Material and methods

The nucleotide and protein NCBI databases were used to retrieve in FASTA format the target nucleotide and protein sequences of *Cyprinus carpio*. The Basic Local Alignment Search Tool (BLAST), P-suite was used to find the regions of similarity between sequences. Multiple sequence alignment (MSA) is carried out by Clustal Omega.

Primary sequence analysis

Pepstats analysis tool (http://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/), available at EMBl-EBI website is used to calculate the statistics of protein properties. ProtParam [3] online tool (http://web.expasy.org/protparam/) available at ExPASy server compute physicochemical properties like theoretical isoelectric point (pI), molecular weight, total number of positive and negative residues, extinction coefficient, instability index, aliphatic index and grand average hydropathy (GRAVY).

Secondary sequence prediction

In order to calculate the secondary structural features Self Optimized Prediction Model (SOPMA) online tool [4] was used. It predict the quantitative values of alpha helix, beta sheet and coils within the protein sequence.

PSIPRED (Protein Sequence Analysis Workbench) (http://bioinf.cs.ucl.ac.uk/psipred/) predict secondary structure from primary sequence and a graphical representation of the secondary structure is obtained. CYS_REC (http://www.softberry.com/berry.phtml) online tool is used to identifies the position of cysteins, total number of cysteins present and pattern, if present, of pairs in the protein sequence.

Subcellular localization was predicted by the use of CELLO v.2.5 [5] 1.1 server (http://www.cbs.dtu.dk). Motif Scan [6] server (http://myhits.isb-sib.ch/cgibin/motif_scan) was used to identify known motifs in the sequence. Furthermore, Pfam server (http://www.sanger.ac.uk/software/pfam/search.html) was used for domain analysis.

Tertiary structure prediction and model evaluation template.

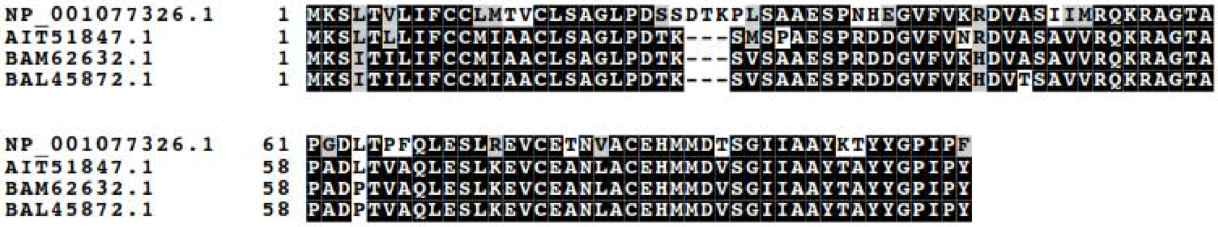
Computational prediction of three dimensional (3D) structure of the protein was performed by using Swiss-Modeler (http://swissmodel.expasy.org/) program [7]. After modeling, the quality and validation of the model was evaluated by several structure assessment methods, containing Z-Score by using QMEAN [8], Rampage Ramachandran plot analysis (http://mordred.bioc.cam.ac.uk), and ERRAT [9].

Result and discussions

The protein sequence (AIT51847.1) was blasted and the list of five protein sequences with significant alignment were retrieved and the detailed information is given in table 1. Multiple Sequence Alignment (MSA) of selected sequences, that was performed by Clustal Omega is given in figure 1, where the shaded regions indicate similar residues.

Table 1: List of protein sequences selected according to species

Organisms	*Accession ID*	*Gene Accession*	*Protein name*
Cyprinus carpio 'color'	AIT51847.1	KF876170.1	bone Gla protein
Carassius auratus	BAM62632.1	AB685220.1	osteocalcin
Carassius auratus	BAL45872.1	AB606419.1	bone gamma-carboxyglutamate protein
Danio rerio	NP_001077326.1	NM_001083857.3	osteocalcin precursor



```
NP_001077326.1     1 MKSLTVLIFCCLMTVCLSAGLPDSSDTKPLSAAESPNHEGVFVKRDVASIIMRQKRAGTA
AIT51847.1         1 MKSLTLLIFCCMIAACLSAGLPDTK---SMSPAESPRDDGVFVNRDVASAVVRQKRAGTA
BAM62632.1         1 MKSITILIFCCMIAACLSAGLPDTK---SVSAAESPRDDGVFVKHDVASAVVRQKRAGTA
BAL45872.1         1 MKSITILIFCCMIAACLSAGLPDTK---SVSAAESPRDDGVFVKHDVTSAVVRQKRAGTA

NP_001077326.1    61 PGDLTPFQLESLREVCETNVACEHMMDTSGIIAAYKTYYGPIPF
AIT51847.1        58 PADLTVAQLESLKEVCEANLACEHMMDVSGIIAAYTAYYGPIPY
BAM62632.1        58 PADPTVAQLESLKEVCEANLACEHMMDVSGIIAAYTAYYGPIPY
BAL45872.1        58 PADPTVAQLESLKEVCEANLACEHMMDVSGIIAAYTAYYGPIPY
```

Figure 1: Multiple Sequence Alignment of selected protein sequences.

Aminoacid composition for each of four selected proteins computed by PEPSTATS is given in table 2. The results show that the most abundant aminoacid is Ala. The least common aminoacid are His, Asn, Gln. The cysteine residues take part in disulphide bonds formation, that play an important role in folding and stability of a protein [10]. The results of table show a low level of cysteine residues, therefore the stability of the protein is not dedicated to the formation of disulphide bonds. The physical and chemical parameters of selected proteins are computed by ProtParam and are shown in table 3. The isoelectric point (pI) varies from 4.16 to 5.56, indicating that proteins are considered as acidic. The Instability index (Ii) is lower than 40, except of *Cyprinus carpio*, therefore these proteins are considered as stable. If the instability index is higher than 40 the protein may be unstable [11]. The GRAVY scores are positive, indicating that all proteins are hydrophobic. Aliphatic Index have very high values in all proteins, indicating that proteins are thermostabile [12].

Table 2: Aminoacid composition obtained by PEPSTATS analysis tools

Organisms	*Most abundant aminoacid*	*Mole (%)*	*Least common amino acid*	*Mole (%)*	*Mole (%) of Cys residues*
Cyprinus carpio 'color'	Ala	14,851	His	0.990	4.950
Carassius auratus	Ala	15.842	Asn	0.990	4.950
Carassius auratus	Ala	14,851	Asn	0.990	4.950
Danio rerio	Ala	8.654	His, Asn, Gln	1,923	4.808

Table 3: Physical and chemical parameters of selected proteins computed by ProtParam

Accession Number	*Sequence length*	*M. W.*	*pI*	*-R*	*+R*	*EC (Cys residues non reduced)*	*EC (Cys residues reduced)*	*II*	*AI*	*Gravy*
Cyprinus carpio 'color'	101	10745.48	4.90	11	8	6210	5960	45.24	91.88	0.286
Carassius auratus	101	10666.36	5.16	11	8	6210	5960	36.26	91.88	0.312
Carassius auratus	101	10696.39	5.16	11	8	6210	5960	35.42	90.89	0.287
Danio rerio	101	11288.12	5.56	11	8	4720	4470	37.24	84.42	0.128

M. W. : molecular weight; -R negatively charged residues; +R positively charged residues; EC. extinction coefficients; II Instability Index; AI Aliphatic Index; GRAVY: Grand average of hydropathicity

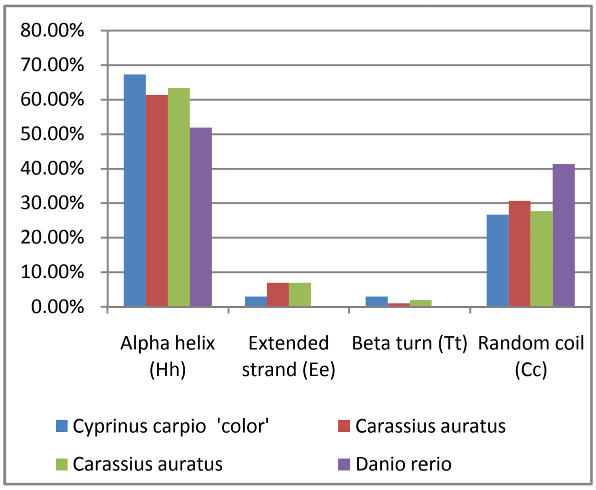


Figure 2: Graphical representation of percentage of helixes, sheets, turns and coils of osteocalcin of different cyprinidae species.

The analysis of disulphide bonds/bridges showed the absence of disulphide bonds between cysteine residues. The results given in figure show that alpha helix dominated followed by random coil, extended strand and beta turn for all four sequences. Alpha helices and beta sheets are secondary structure elements. Hydrogen bonding is a well known feature of alpha helix, that play role in folding and stabilization [10]. Beta sheets play role in biological activities of proteins [10]. Also, the beta turns play an important role in protein stability. [10]. Random coils have important functions like for flexibility and conformational changes of proteins [13]. As it is shown in the figure 3, the confidence of prediction observed throughout the predicted secondary structure was quite high.

```
                         10                  20                  30                  40                  50
  1   M K S L T L L I F C C M I A A C L S A G L P D T K S M S P A E S P R D D G V F V N R D V A S A V V R    50
 51   Q K R A G T A P A D L T V A Q L E S L K E V C E A N L A C E H M M D V S G I I A A Y T A Y Y G P I P   100
101   Y                                                                                                   101
                         10                  20                  30                  40                  50
```

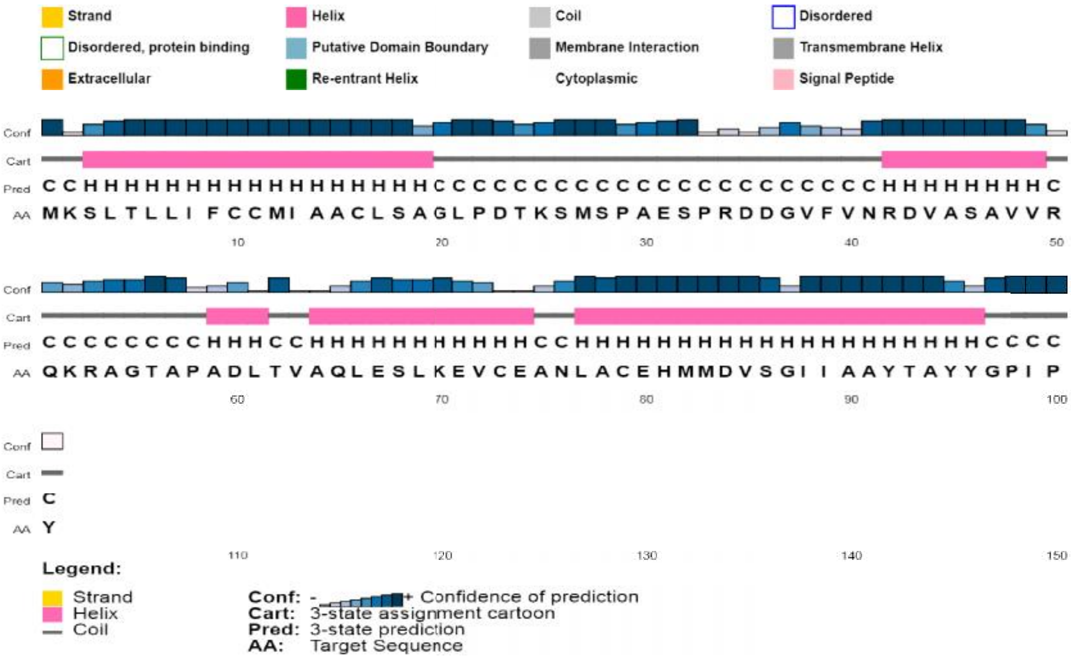

Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide
Conf
Cart
Pred CCHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCHHHHHHHHC
AA MKSLTLLIFCCMIAACLSAGLPDTKSMSPAESPRDDGVFVNRDVASAVVR
10 20 30 40 50
Conf
Cart
Pred CCCCCCCCHHHCCHHHHHHHHHHHCCHHHHHHHHHHHHHHHHHHHHCCCC
AA QKRAGTAPADLTVAQLESLKEVCEANLACEHMMDVSGIIAAYTAYYGPIP
60 70 80 90 100
Conf
Cart
Pred C
AA Y
110 120 130 140 150
Legend:
Strand
Helix
Coil
Conf: - + Confidence of prediction
Cart: 3-state assignment cartoon
Pred: 3-state prediction
AA: Target Sequence

Figure 3: Graphical representation of the predicted secondary structures present within the target protein osteocalcin (AIT51847.1)

Subcellular localization prediction was performed by the use of CELLO v 2.5 and osteocalcin was localized in the extracellular matrix. According to Nishimoto et al., 2003, [2] Osteocalcin (bone Gla protein) is an extracellular matrix protein synthesized by osteoblasts,

There were determined three types of motifs, by Motif scan tools (Table 4). The highest number of motifs was N-myristoylation site, with four times. Protein N-myristoylation is a cotranslational lipidic modification of many eukaryotic proteins consisting on the attachment of myristic acid to the N-terminus that is catalyzed by the ubiquitous eukaryotic enzyme, N-myristoyltransferase (NMT) [14].

Table 4 : The motifs of osteocalcin in C. carpio by motif Scan

Motif information	*No of sites*	*Aminoacid residues*
Casein Kinase II pohosphorylation site	3	28 - 31; 32 - 35; 8 - 71
N-myristoylation site	4	20 - 25; 37 - 42; 55 - 60; 87 - 92;
Protein kinase C phosphorylation site	2	32 - 34; 68 - 70

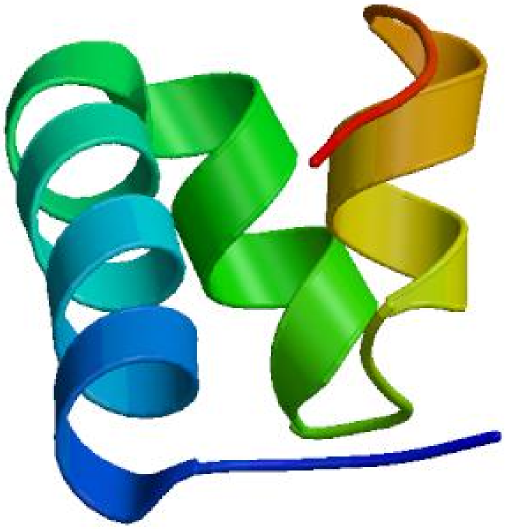


Figure 4. The predicted three-dimensional structure of osteocalcin by modelled SWISSMODEL

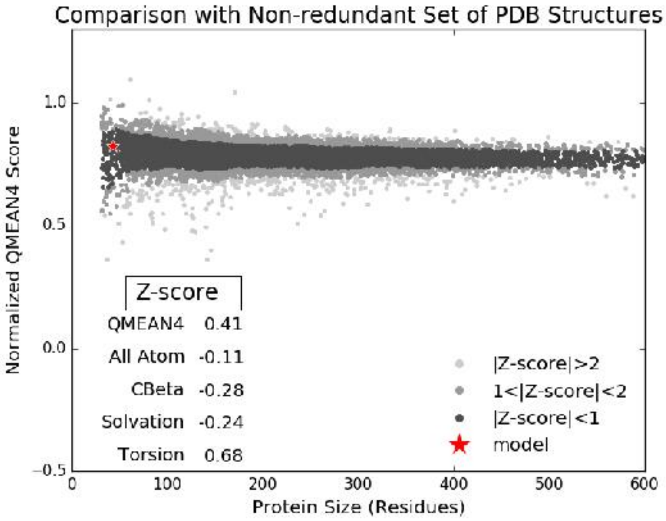
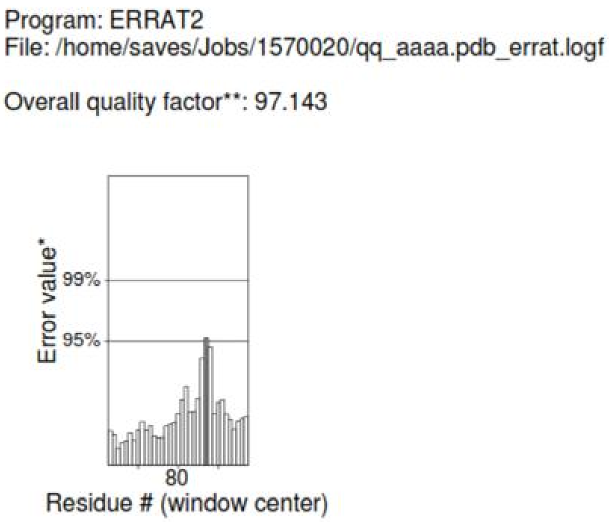


Figure 5:Z-score of query protein using QMEAN server

The three dimensional structure of osteocalcin was predicted by SWISS MODEL homology modeling program. The results are shown in figure. PDB 1vzm.1.A was selected as template with 61.36% sequence identity to query sequence (AIT51847.1).

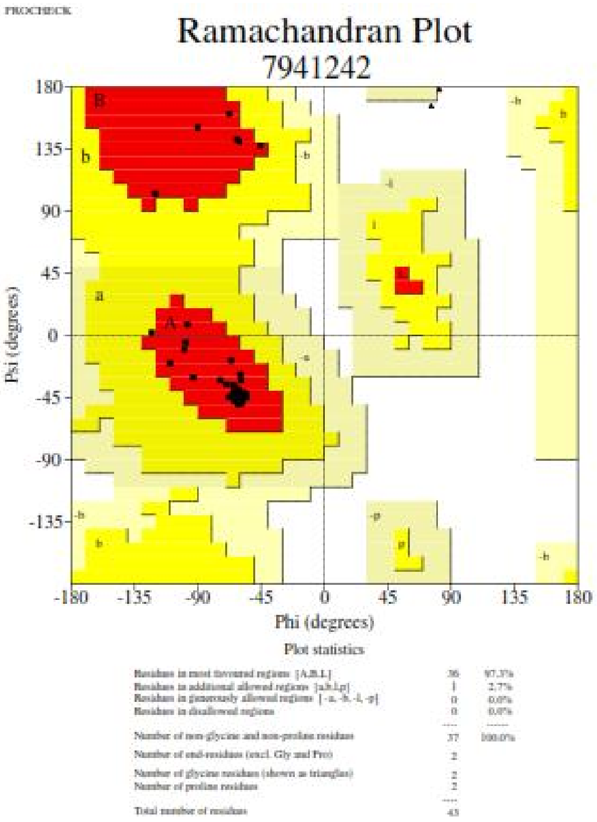
The structure was validated validated through energy minimization with Z-Score by the use of Qmean server, ERRAT, and Rampage Ramachandran plot analysis. The Z-score was found as 0.41 (Figure 5). The overall quality factor evaluated by ERRAT was found as 97.143 which is very good quality (Figure 6) Ramachandran plot analysis (Figure 7) showed 97.3% residues in most favored region, 2.7% residues in additonal allowed region. More than 90% of residues reside in the favored region, which imply a good quality model [15].



```
Program: ERRAT2
File: /home/saves/Jobs/1570020/qq_aaaa.pdb_errat.logf

Overall quality factor**: 97.143
```

Figure 6: Overall quality factor evaluated by ERRAT



```
Plot statistics

Residues in most favoured regions [A,B,L]                    36    97.3%
Residues in additional allowed regions [a,b,l,p]              1     2.7%
Residues in generously allowed regions [~a,~b,~l,~p]          0     0.0%
Residues in disallowed regions                                0     0.0%
                                                             ----  ------
Number of non-glycine and non-proline residues               37   100.0%

Number of end-residues (excl. Gly and Pro)                    2

Number of glycine residues (shown as triangles)               2
Number of proline residues                                    2
                                                             ----
Total number of residues                                     43
```

Figure 7: Graphical representation of Ramachandran Plot by ProCheck

Conclusions

In this paper we tried to analyze structural properties and predict 3D structure of osteocalcin in *Cyprinus carpio*. The analysis by the use of different online softwares showed that it is an extracellular protein, thermostable with a molecular weight of 10.75 kD. The amino acid composition reveals the abundance of Alanine amino acid in all the Cyprinidae studied species. Secondary structure analysis revealed that alpha helix were more abundant towards other secondary structure elements like random coil, extended strand and beta turn.

The modelling of 3D structure of osteocalcin was performed by Swiss Model and was validated with Z-Score by the use of Qmean server, ERRAT, and Rampage Ramachandran plot analysis

References:

1. Durand JD, Tsigenopoulos CS, Unlu: **Phylogeny and biogeography of the family Cyprinidae in the Middle East inferred from cytochrome b DNA—evolutionary significance of this region**. *Molecular Phylogenetics and Evolution* 2002, **22**(1): 91-100.
2. Nishimoto SK, Waite JH, Nishimoto M, Kriwacki RW: **Structure, activity, and distribution of fish osteocalcin**. *Journal of Biological Chemistry* 2003, **278**(14): 11843-11848.
3. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A, others: **Protein identification and analysis tools on the ExPASy server**. *The proteomics protocols handbook* 2005: 571-607.
4. Geourjon C, Deleage G: **SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments**. *Bioinformatics* 1995, **11**(6): 681-684.

5. Yu C-S, Cheng C-W, Su W-C, Chang K-C, Huang S-W, Hwang J-K, Lu C-H: **CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation**. *PloS one* 2014, **9**(6): e99368.

6. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database**. *Nucleic acids research* 2006, **34**(suppl-1): D227-D230.

7. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L *et al*: **SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information**. *Nucleic acids research* 2014, **42**(W1): W252-W258.

8. Benkert P, Kunzli M, Schwede T: **QMEAN server for protein model quality estimation**. *Nucleic acids research* 2009, **37**(suppl-2): W510-W514.

9. Colovos C, Yeates TO: **Verification of protein structures: patterns of nonbonded atomic interactions**. *Protein science* 1993, **2**(9): 1511-1519.

10. Feisal MR, others: **In silico structural analysis, physicochemical characterization and homology modeling of Arabidopsis thaliana Na+/H+ exchanger 1 protein**. : BRAC University; 2015.

11. Guruprasad K, Reddy BB, Pandit MW: **Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence**. *Protein Engineering, Design and Selection* 1990, **4**(2): 155-161.

12. Verma A, Singh VK, Gaur S: **Computational based functional analysis of Bacillus phytases**. *Computational biology and chemistry* 2016, **60**: 53-58.

13. Buxbaum E: **Fundamentals of protein structure and function**, vol. 31: Springer; 2007.

14. Udenwobele DI, Su R-C, Good SV, Ball TB, Varma Shrivastav S, Shrivastav A: **Myristoylation: An important protein modification in the immune response**. *Frontiers in immunology* 2017, **8**: 751.

15. Yadav PK, Singh G, Gautam B, Singh S, Yadav M, Srivastav U, Singh B: **Molecular modeling, dynamics studies and virtual screening of Fructose 1, 6 biphosphate aldolase-II in community acquired-methicillin resistant Staphylococcus aureus (CA-MRSA)**. *Bioinformation* 2013, **9**(3): 158.