Research Article

Open Access

Ab-initio prediction of sequence and structural biology of fish muscle proteins using homology modeling, phylogeny and different computational approaches

Abstract

Fish is a diverse group of organisms living in different aquatic environment and containing almost all essential amino acids. Fourteen muscle proteins including titin, dystrophin, filamin, myosin heavy chain, spectrin, M1/M2, nebulin, alpha-actinin, gelsolin, actin, tropomyosin, troponin, thymosin and plastin3 were chosen for in-silico characterization. Sequence analyses were performed using BindN, Conseq, DIANNA, PROFEAT and ProtFun for exploiting structural and functional importance. Homology modeling technique was applied for predicting 3D structure which will assist in future for searching catalytic role of proteins in metabolic pathway. 3D Structure of eight muscle proteins was predicted using Protein Structure Prediction Server (PS2) based on MODELLER algorithm. Phylogenetic relationship was inferred by sequence alignment through CLUSTAL X and furthermore phylogenetic tree was constructed by using MEGA which was statistically evaluated by DIVIEN. From structural analyses, these muscle proteins were inferred to contain functional domains, number of motifs, beta turns with important secondary structural features. Furthermore sequence study suggested, these proteins have important biochemical features such as number of cysteines, disulphide bonds, DNA and RNA binding sites, functionally conserved amino acid residues and were characterized as non-allergen proteins which can be used for designing effective vaccines. Overall, evidence from computational study revealed that these muscle proteins have structural and functional significance, which can play important role in drug designing and in exploring gene diversity. This novel approach to study muscle proteins would be beneficial for human since both vertebrates and invertebrates have muscle proteins in common.

Keywords: sequence analyses, homology modeling, structural analyses, vertebrates, invertebrates

Volume 2 Issue 3 - 2015

Sana Khalid,[1] Sobia Idrees,[1] Hina Khalid,[1] Bilal Hussain,[1] Sandeep Tiwari,[2] Syed Shah Hassan,[2] Artur Silva,[3] Vasco Azevedo,[2] Syed Babar Jamal[2]

[1]Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Pakistan
[2]Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Brazil
[3]Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil

Correspondence: Syed Babar Jamal, Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, Tel 005531 3409 2610, Fax 005531 3409 2610, Email syedbabar.jamal@gmail.com

Received: April 09, 2015 | **Published:** May 05, 2015

Introduction

Protein with its amino acid are important for maintaining structure of cells, making antibodies to work properly, regulate the growth of hormones with enzymes and contributes to the repairing mechanisms. Organism living in marine and fresh water consists of protein with high amino acid proportion. Fish is a diverse group of organisms that habituates in different aquatic environment and holds prime importance in food industry. Biologically, fish muscle proteins contain all essential nutrients like milk, meat and egg protein. This protein varies in amount from species to species. Globally the consumption of fish production by human is about 77 percent. Fish was chosen as a sample source because there are many different verities of fish and source of protein for many fish species are readily available. Furthermore, fish is very nutritious part of man's diet since it is rich in vitamins, minerals and all essential amino acids in right proportions. Study of muscle genes and proteins will be beneficial for human for *in silico* drug designing. Fish skeletal muscle is known to be the perfect model to explore the structure and function of muscle, due to perfect arrangement of different types of fibers which is present in axial and pectoral fin muscles.[1] Thus, computational study will allow muscle genes and proteins to be studied at greater level of detail. A variety of bioinformatics tools are available for detailed comparative study and visualization of amino acid sequences, which provides knowledge about molecular evolution and variety of information related to structure and function of protein. Detection of conserved regions in protein and nucleic acid sequences are of great importance, because it gives knowledge about structure and function.[2] Then *in silico* study of fish muscle proteins was performed to analyze its structural and functional importance with amino acid properties.

The objective of present study was to perform sequence analysis of fish muscle proteins, using different computational tools, study the amino acid composition and secondary structure features, using homology-modeling approach to find the 3D structure of muscle proteins. In addition, illustrate physiochemical properties by ensuring the quality of the predicted model and finally predicting the evolutionary relationship of various proteins to get knowledge about biodiversity of different species with homologous sequences.

Materials and methods

Protein retrieval and sequence analysis

Protein sequences of fish muscle were retrieved from Uniprot Knowledgebase database and NCBI using accession no. G1ERR8, Q9PV76, E6ZGD0, Q9PRF1, F8K8N3, Q1L5K3, E6ZHF3, gi|5726351, Q8AW95, gi|59858543, Q58HZ6, Q9NAS5, gi|185132813, Q8UVF6 and gi|49901349. These muscle proteins selected were titin, dystrophin, filamin, myosin, spectrin, M1, M2, nebulin, acitinin, gelsolin, actin, tropomyosin, troponin, thymosin and plastin 3. A detailed sequence analysis of selected proteins was performed to understand its structure and function with significant features. DNA and RNA binding sites were predicted using BindN[3] in order to understand the function of DNA and RNA binding protein. ConSurf[4] was used for predicting position of evolutionary conserved amino acids. The degree of conservation of amino acid depicts the structural and functional importance. The positions that evolve rapidlyare considered as variable while positions that evolve slowly are known to be conserved. Cysteine classification and disulfide connectivity prediction was carried out using DiANNA[5] tool. This knowledge helps us to understand secondary structure of protein since disulfide bonds bridges play important role for stabilizing the folding process in protein. In addition knowledge of disulfide bond with cysteine also provides information related to genome annotation. Structural and physicochemical features of proteins and peptides were computed using PROFEAT.[6] These features were predicted by machine learning methods, which contribute to structural and functional knowledge. ProtParam[7] was used for evaluation of physico chemical features of protein. Parameters computed by ProtParam were molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half life, aliphatic index and grand average of hydropathicity (GRAVY).

Prediction of secondary structure

Secondary structure of muscle proteins were computed using SWISS PDB Viewer,[8] PSIPRED,[9] NetTurnP[10] and NetSurfP.[11] Secondary structure features such as helices, strands, coils, acidic and basic residues, domains, transmembrane topology were predicted using Swiss PDB viewer and PSIPRED. NetTurnP and NetSurfP was used for beta turns and protein surface accessibility prediction. Beta turns formation are important in folding, stability of proteins and molecular recognition processes.

Prediction of 3D structure by using homology-modeling approach

Homology modeling approach was used to predict three dimensional structure of fish muscle proteins including actin, actinin, dystrophin, gelsolin, M2 protein, plastin, thymosin and troponin. The 3D model generated by this computational approach has a high resolution with good accuracy. First BLAST database was searched to find the best template of known structure with highest identity. BLAST search with default parameters were performed against PDB to find best template. The template having maximum identity was selected for homology modeling to study the protein of interest. Then 3D model was generated by using template of known structure with the help of protein structure prediction web server (PS2).[12] Protein structure prediction server was selected because it is based on MODELLER algorithm and easy to use.

Validation of 3D structure

After 3D model was constructed evaluation was performed using PSVS and WHAT IF. PSVS was used for assessment of 3D model which integrates information from various structure evaluation software including RPF, PROCHECK, MolProbity, Verify 3D, Prosa II, and other structure validation software. Stereochemistry analyses were performed using WHAT IF. Deep View was used for visualizing 3D structure.[11]

Functional analyses of fish muscle proteins

To study the function of muscle proteins ProtFunc[13] was used. This server utilizes information from other prediction server of DAS annotaion viewer related to post transational modification then finally categorize the information in form of cellular role, enzyme class and gene ontology features. NCBI's Conserved Domain Database (CDD)[14] was used for finding conserved domain in protein sequence.

Submission of the model in protein model database (PMDB)

The models generated for actin, actinin, dystrophin, gelsolin, M2 protein, plastin 3, thymosin, troponin was successfully submitted in Protein model database (PMDB)[15] having PMID: PM0078304, PM0078303, PM0078298, PM0078299, PM0078300, PM0078301, PM0078302 and PM0078305.

Phylogenetic analysis of fish muscle proteins

This section includes multiple sequence alignment of proteins, phylogenetic tree construction and its evaluation, performed using following computational approach. Phylogenetic trees of 10 fish muscle proteins including actin, actinin, dystrophin, fimbrin, gelsolin, myosin heavy chain, spectrin, titin, tropomyosin and troponin were made. BLAST analysis of selected proteins was performed against non redundant databases by setting parameters on default. Then sequences with highest identity greater than 70% were collected for multiple sequence alignment. The same strategy was repeated for each selected protein and step by step sequences were collected for multiple sequence alignment. Computational tools including Clustal X,[16] MEGA[17] and DIVEIN[18] were used for understanding the evolutionary significance of fish muscle proteins.

Multiple Sequence Alignment through Clustal X

Clustal X[16] is a widely used multiple sequence alignment tool which is completely coded in C++. Clustal X, which is desktop version of Clutal W was used for multiple sequence alignment in order to get knowledge about structure, function, location, stability and origin of protein. FASTA formatted file containing amino acid sequences was loaded to Clustal X as given by opening file menu. These amino acid sequences were selected by performing BLAST analysis of fish muscle proteins against non redundant protein sequence databases. The sequences with lower E-value and identity greater than 70% were chosen for multiple sequence alignment. The alignment was performed in Clustal X by setting parameters as gap opening 30, gap extension 20, delay divergent sequences 30, negative matrix off and protein weight matrix used was Gonnet series. Nexus, Clustal and FASTA was marked for an output.

Construction of Phylogenetic tree by using MEGA

MEGA[17] stands for Molecular Evolutionary Genetics Analysis used for evolutionary study of DNA and protein sequences. It is a desktop application which was used for comparative study of homologous sequences belonging to different species and different gene families. MEGA 4 was used for constructing phylogenetic trees. The Molecular Evolutionary Genetics Analysis was downloaded and saved on desktop. Multiple sequence alignment was loaded and newick trees were constructed, then tree image was displayed. Bootstrap analysis was performed on 1000 replicates using maximum likelihood algorithm and phylogenetic tree was constructed for each relevant protein to understand the origin and evolution of species.

Statistical evaluation of phylogenetic trees using DIVEIN

DIVEIN[18] which stands for Divergence, diversity, informative sites and phylogenetic was used for computing the statistical measurements related to diversity and divergence from pairwise distance. It was also used for graphical visualization of phylogenetic trees. DIVEIN requires Apache server to run. Maximum likelihood approach is implemented using PhyML which uses Hill climbing algorithm for adjusting branch length and topology of tree. Nexus file in aligned format was used as an input, for evaluating phylogenetic trees.

Results and discussion

The present study was to perform sequence and structure analysis of fish muscle proteins. The protein sequences were retrieved from Uniprot database and NCBI with accession number as G1ERR8, Q9PV76, E6ZGD0, Q9PRF1, F8K8N3, Q1L5K3, E6ZHF3, gi|5726351, Q8AW95, gi|59858543, Q58HZ6, Q9NAS5, gi|185132813, Q8UVF6 and gi|49901349.

Protein sequence analysis

BindN was used for predicting DNA and RNA binding residues for fish muscle proteins which is useful for understanding protein-nucleic acid interaction. The degree of conservation of amino acid depicts the structural and functional importance. The positions which evolve rapidly are considered as variable while positions which evolve slowly are known to be conserved. This tool was used for identification of functional region in fish muscle proteins. ConSurf was explored for estimation of evolutionary conserved amino acids in protein which was based on phylogenetic relationship inferred from homologous sequences (Table 1).

Table 1 Binding residues with conserved amino acids predicted by BindN and ConSurf

Protein	Total no. of residues	No. of exposed residues according to neural network algorithm	No. of buried residues according to neural network algorithm	No. of functional residues (highly conserved and exposed)	No. of structural residues (highly conserved and buried)	Predicted DNA binding residues	Predicted RNA binding residues
Actin	103	52	51	21	4	22	22
Actinin	110	64	46	20	12	21	24
Dystrophin	40	29	11	6	4	15	16
Filamin	1343	0	0	0	0	262	303
Gelsolin	730	458	235	112	47	147	186
M1	196	0	0	0	0	37	31
M2	190	115	75	25	15	40	33
Myosin	43	28	15	17	8	7	9
Nebulin	57	43	14	14	7	29	33
Plastin	627	405	221	83	46	103	122
Spectrin	220	154	66	32	8	40	49
Thymosin	42	38	4	7	0	12	13
Titin	129	80	49	33	17	21	39
Tropomyosin	284	213	73	56	7	45	69
Troponin	223	186	37	39	5	81	121

PROFEAT is a bioinformatics server used for calculating structural and chemical features of protein from primary sequence data. These features provides knowledge about biological properties of proteins and peptides. Thus in order to compute the structural and physicochemical features of proteins and peptides PROFEAT was used. All fish muscle proteins were found as non allergen (Table 2).

Table 2 Protein family name predicted by PROFEAT

Protein	Protein functional family prediction
Titin	All lipid binding protein, ion binding, chlorophyll biosynthesis, calcium binding, TC 3A 1 ATP binding cassette (ABC) family, motor protein, actin binding, magnesium binding.
Filamin	Cell adhesion, zinc binding, all lipid binding proteins, virulence, metal binding, antigen, actin binding, and DNA repair.
Spectrin	All lipid binding proteins, metal binding, actin binding, calcium binding.
M1	Iron binding, transferases, alkyl or aryl groups, all lipid binding proteins, zinc binding, structural protein (matrix protein, core protein, viral occlusion body, keratcin), oxidoreductases acting on CH-CH group of donors, lipid metabolism, transferases including acyl transferases, all DNA binding, metal binding, lyases including carbon oxygen lyases, DNA repair.
M2	Transmembrane, transferases are including glycotransferases, iron binding, copper binding, oxidoreductases acting on heme group of donors, magnesium binding.

Table Continued....

Protein	Protein functional family prediction
Actinin	rRNA binding protein, zinc binding, DNA repair, calcium binding, magnesium binding, TC 3A 1 ATP binding cassette (ABC) family.
Gelsolin	Zinc binding, actin capping, tranferases including glycotranferases, all lipid binding protein, metal binding, actin binding, photosystem 1, calcium binding.
Actin	Zinc binding, all DNA binding, actin binding.
Tropomyosin	All lipid binding protein, actin binding, copper binding.
Troponin	Copper binding.
Plastin 3	Zinc binding, trasnferases transferring phosphorous containing groups, glycotransferases, metal binding, all lipid binding protein, actin binding, calcium binding, pore forming toxins (proteins and peptides), transferases transferring one carbon groups, photosystem 1, carbon binding.

Prediction of 3D structure by using homology-modeling approach

An important term used in structure prediction is homology modeling which refers to prediction of three-dimensional structure of protein by using template of known 3D structure. The 3D structure of protein provides knowledge about function of protein and activity of an enzyme. Structure prediction also plays key role in bioinformatics in terms of medicine and biotechnology. First BLAST database was searched to find the best template of known structure with highest identity. BLAST search with default parameters were performed against PDB to find best template. The template having maximum identity was selected for homology modeling to study the protein of interest. Then 3D model was generated by using template of known structure with the help of protein structure prediction web server (PS^2). Template used for predicting 3D model was 1D4X_A for actin, ITJT_A for actinin, 1DXX_A for dystrophin, 2FGH_A for gelsolin, 2JDF_A for M2 protein, 1AOA_A for plastin 3, 1HJO_A for thymosin and 1JID_E for tropnin (Figure 1-8).

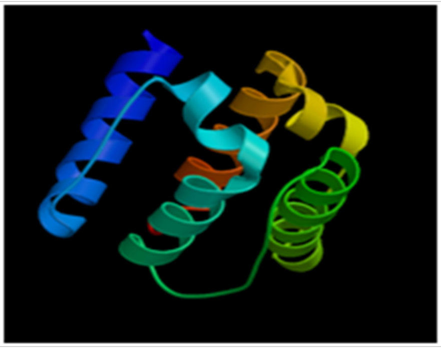

Figure 1 Actinin 3D structure.

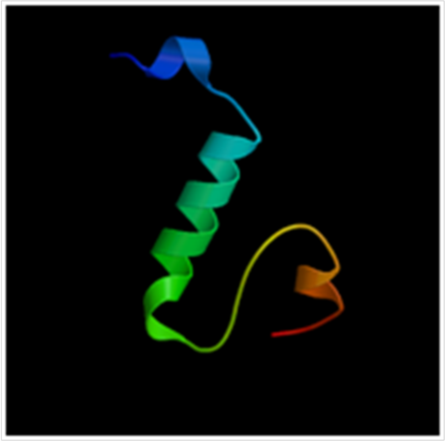

Figure 2 Dystrophin 3D structure

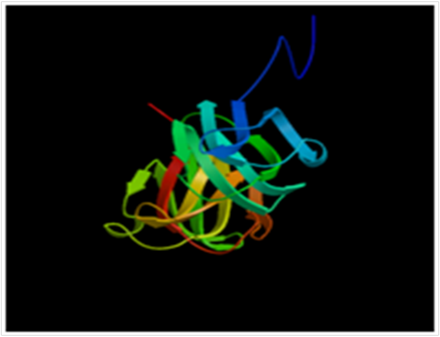

Figure 3 M2 protein 3D structure.

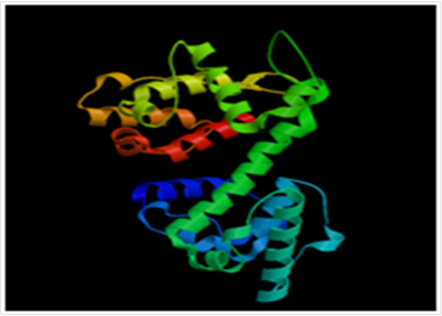

Figure 4 Plastin3D structure.

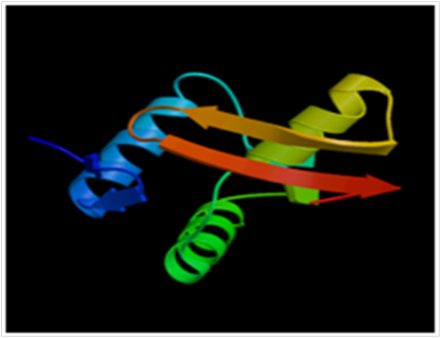

Figure 5 Actin 3D structure.

After construction of 3D model evaluation was performed using PSVS and WHAT IF. PSVS was used to determine the Ramachandran plot to assure the quality of the model. The result of the Ramachandran plot of all predicted models showed greater than 90% residues in favorable region representing that it is a reliable and good quality model (Table 3). A model having more than 90% residues in favorable region is considered as good quality model. 3D model was further evaluated by WHAT IF, which after performing stereo chemical analysis indicated that predicted models are correct.

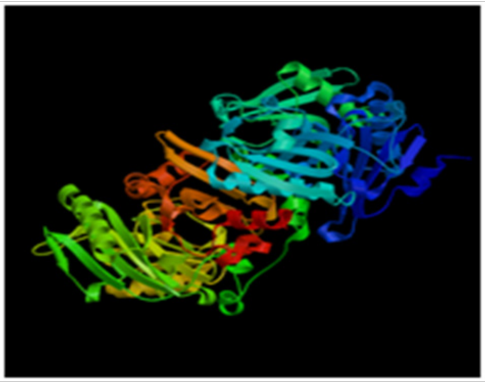

Figure 6 Gelsolin 3D structure

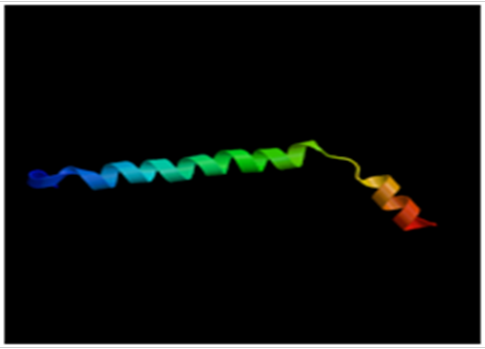

Figure 7 Thymosin 3D structure.

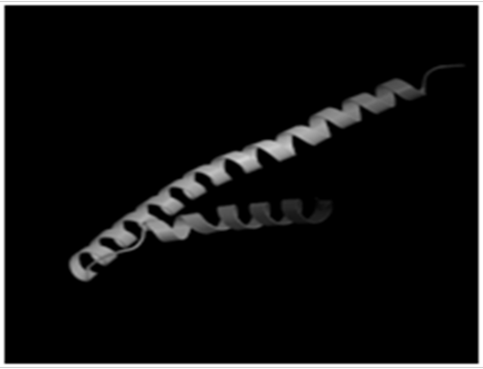

Figure 8 Troponin 3D structure.

Visualization of 3D structures was performed using DEEP VIEW. Secondary structure of muscle proteins were computed using SWISS PDB Viewer and PSIPRED. NCBI's Conserved Domain Database (CDD) was used for finding conserved domain in protein sequence. Secondary structure features (Table 4) such as helices, strands, coils, acidic and basic residues, domains, transmembrane topology were predicted using Swiss PDB viewer, CDD and PSIPRED.

Secondary structure of protein plays important role in protein classification, predicting structural changes and function of protein.

NetTurnP and NetSurfP was used for beta turns (Table 5) and protein surface accessibility prediction. Beta turns are non repetetitive structures. Beta turns formation are important in folding, stability of proteins and molecular recognition processes. DiANNA[5] was used for cysteine classification and prediction of disulfide connectivity, which provides useful information related to secondary structure since disulphide bonds, helps in stabilizing the folding of protein.

Functional analyses of fish muscle proteins

To study the function of muscle proteins ProtFunc (Table 6) was used. This study predicted that all muscle proteins have functional importance and were found to be involved in different body functions. Titin and Dystrophin was found to play role in translation, were classified as an enzyme, helps in immune response and acts as lyases. Filamin was functionally categorized as purines and pyrimidines, was classified as an enzyme, acts as lyases and important structural protein. Spectrin was known to be involved in regulatory functions, was classified as nonenzyme and acts as an important growth factor. M1 was found to play role in amino acid biosynthesis, was classified as an enzyme and acts as ligase. M2 was found to play role in energy metabolism, acts as an enzyme and helps in transcription regulation. Nebulin was known to be involved in regulatory functions, was classified as non enzyme and plays role in transcription. Actinin was found to play role in translation, was classified as nonenzyme and acts as an important growth factor. Gelsolin was found essential in central intermediary metabolism, was classified as an enzyme and acts as hydrolases. Actin was found to play role in energy metabolism, was classified as an enzyme and acts as an important growth factor. Tropomyosin and Troponin was found to play role in translation, was classified as nonenzyme and helps in transcription regulation. Thymosin acts as an important hormone. Plastin was found to play role in amino acid biosynthesis, was classified as an enzyme and acts as ligase.

Table 3 Tabulated form of predicted structure of fish muscle proteins Illustrating template and target used with some physiochemical properties predicted by ProtParam

PMDB ID	Protein ID	Target protein	PDB template	Ramachandron Plot % score	Lengh of a.a	Molecular weight	Theoreticl PI
PM0078304	Q58HZ6	Actin	1D4X_A	96.7%	103	11630	5.71
PM0078303	Q8AW95	Actinin	1TJT_A	98%	110	12470	9.47
PM0078298	Q9PV76	Dystrophin	1DXX_A	91.7%	40	4532	8.36
PM0078299	gi|59858543	Gelsolin	2FGH_A	91.7%	730	81360.5	5.54
PM0078300	E6ZHF3	M2 protein	2JDF_A	93.2%	190	23107.3	7.56
PM0078301	gi|49901349	Plastin 3	1AOA_A	93.2%	190	76149.5	5.95
PM0078302	Q8UVF6	Thymosin	1HJO_A	97.3%	42	4851.5	5.31
PM0078305	gi|185132813	Troponin	1JID_E	100%	75	9256	9.86

Table 4 Prediction of secondary structure features of fish muscle proteins

PMDB ID	Helices	Strands	Coils	Acidic residues	Basic residues	Domains	Motif
PM0078304	37	28	39	12	7	1	16
PM0078303	71	0	40	12	17	1	148
PM0078298	15	0	26	2	3	1	17
PM0078299	158	252	321	100	83	6	102
PM0078300	12	86	93	18	19	2	147
PM0078301	139	0	117	86	78	6	108
PM0078302	33	0	10	10	9	1	26
PM0078305	67	0	10	10	9	0	125

Table 5 Summarized table of total number of Beta turns, cysteines, disulphide bond predicted by Net turn P and DIANNA

Protein name	No. of beta turns	No. of predicted cysteines	No. of predicted disulfide bonds
Actin	21	4	0
Actinin	0	2	1
Filamin	766	21	10
Gelsolin	0	9	4
M1	30	5	2
M2	0	11	5
Plastin	170	8	4
Spectrin	32	2	1
Titin	47	3	1
Dystrophin	10	0	0
Thymosin	1	0	0

Table 6 Protein function predicted by ProtFunc

Protein	Protein function predicted by protfunc
Titin	Play role in translation, classified as an enzyme, help in immune response, and acts as lyases.
Dystrophin	Play role in translation, classified as an enzyme, help in immune response, and acts as lyases.
Filamin	Functionally categorized as purines and pyrimidines, classified as an enzyme, acts as lyases and important structural protein.
Spectrin	Known to be involved in regulatory functions, classified as nonenzyme, acts as an important growth factor.
M1	Play role in amino acid biosynthesis, classified as an enzyme, act as a ligase.
M2	Play role in energy metabolism, acts as an enzyme, known to be involved in transcription regulation.
Nebulin	Known to be involved in regulatory functions, classified as non enzyme, play role in transcription.
Actinin	Play role in translation, classified as nonenzyme, acts as a growth factor.
Gelsolin	Play role in central intermediary metabolism, classified as an enzyme, acts as hydrolases.
Actin	Play role in energy metabolism, classified as an enzyme and acts as an important growth factor.
Tropomyosin	Play role in translation, classified as nonenzyme.
Troponin	Play role in translation, classified as nonenzyme, known to be involved in transcription regulation.
Thymosin	Play role in translation, classified as nonenzyme, acts as an important hormone.

Protein function predicted by ProtFunc

Submission of the model in protein model database (PMDB)

The models generated for actin, actinin, dystrophin, gelsolin, M2 protein, plastin 3, thymosin, troponin was successfully submitted in Protein model database (PMDB) and can be find using PM0078304, PM0078303, PM0078298, PM0078299, PM0078300, PM0078301, PM0078302 and PM0078305.

Phylogenetic analysis of fish muscle proteins

By inferring phylogeny novel type of relationship was predicted among species including *Amphistichus koelzi, Oryzias latipes, Dicentrarchus labrax, Plecoglossus altivelis, Daniorerio, Salmosalar, Macrobrachium rosenbergii* and *Anisakis simplex*. Comparative study of actin, actinin, plastin3 or fimbrin, gelsolin, myosin, spectrin, tropomyosin and troponin fish protein revealed the genetic divergence in to two major lineages. Phylogenetic topology of titin and dystrophin muscle protein revealed the genetic divergence into four lineages (Figure 9-18).

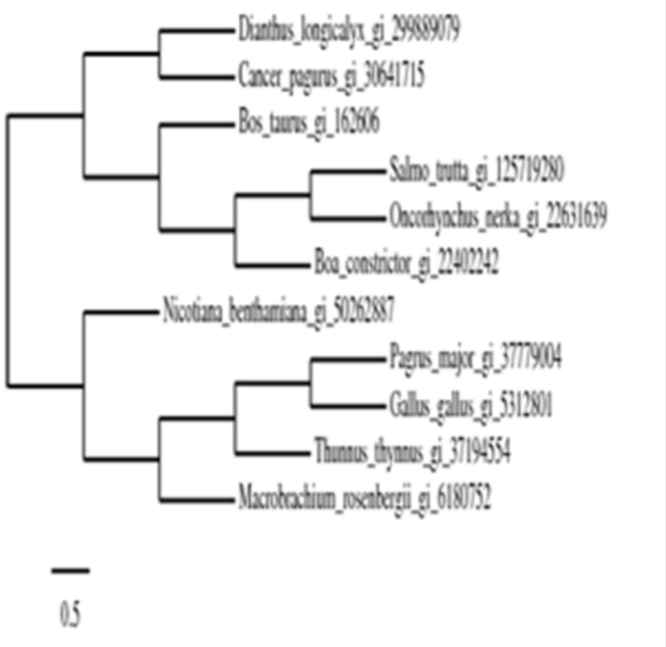
Dianthus_longicalyx_gi_299889079
Cancer_pagurus_gi_30641715
Bos_taurus_gi_162606
Salmo_trutta_gi_125719280
Oncorhynchus_nerka_gi_22631639
Boa_constrictor_gi_22402242
Nicotiana_benthamiana_gi_50262887
Pagrus_major_gi_37779004
Gallus_gallus_gi_5312801
Thunnus_thynnus_gi_37194554
Macrobrachium_rosenbergii_gi_6180752
0.5

Figure 9 Phylogenetic tree of Actin

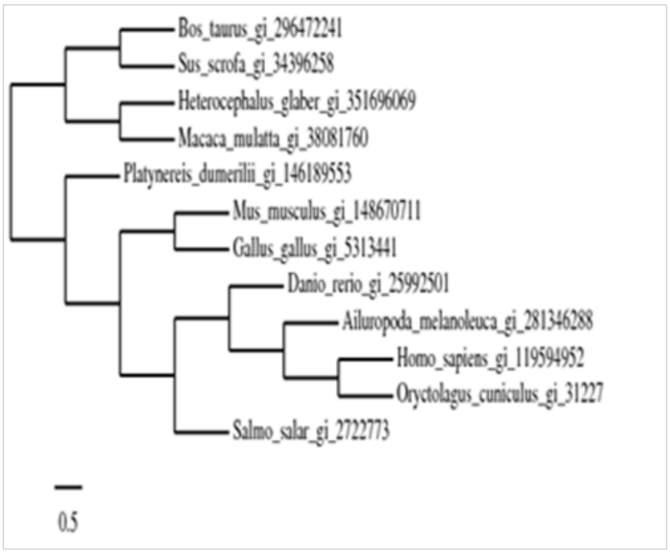
Bos_taurus_gi_296472241
Sus_scrofa_gi_34396258
Heterocephalus_glaber_gi_351696069
Macaca_mulatta_gi_38081760
Platynereis_dumerilii_gi_146189553
Mus_musculus_gi_148670711
Gallus_gallus_gi_5313441
Danio_rerio_gi_25992501
Ailuropoda_melanoleuca_gi_281346288
Homo_sapiens_gi_119594952
Oryctolagus_cuniculus_gi_31227
Salmo_salar_gi_2722773
0.5

Figure 10 Phylogenetic tree of Actinin.

Statistical evaluation of phylogenetic tree

To computes the statistical measurements related to diversity and divergence from pairwise distance DIVEIN (Table 7) was used. It allows graphical visualization of phylogenetic trees. DIVEIN requires Apache server to run. Maximum likelihood approach was implemented using PhyML which applies Hill climbing algorithm for adjusting branch length and topology of tree.

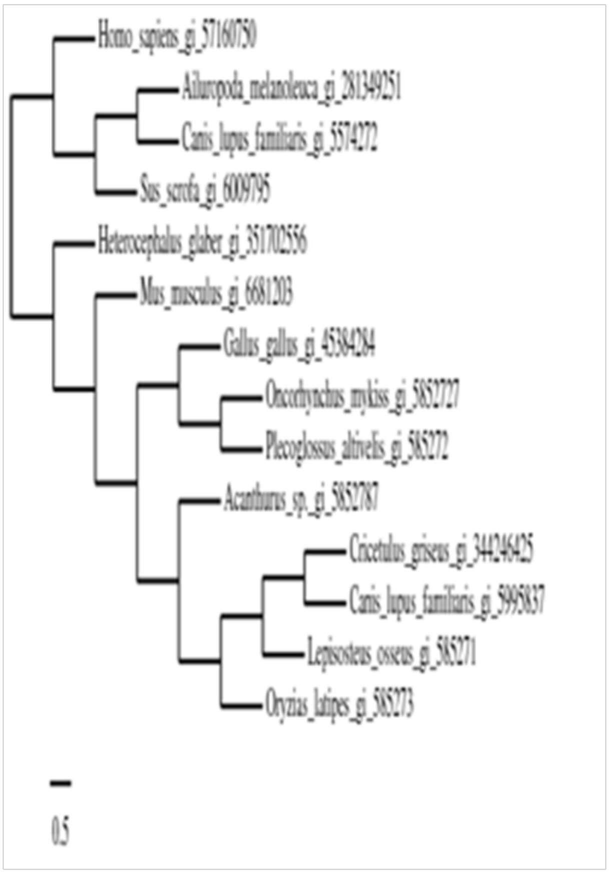
Homo_sapiens_gi_57160750
Ailuropoda_melanoleuca_gi_281349251
Canis_lupus_familiaris_gi_5574272
Sus_scrofa_gi_6009795
Heterocephalus_glaber_gi_351702556
Mus_musculus_gi_6681203
Gallus_gallus_gi_45384284
Oncorhynchus_mykiss_gi_5852727
Plecoglossus_altivelis_gi_585272
Acanthurus_sp._gi_5852787
Cricetulus_griseus_gi_344246425
Canis_lupus_familiaris_gi_5995837
Lepisosteus_osseus_gi_585271
Oryzias_latipes_gi_585273
0.5

Figure 11 Phylogenetic tree of Dystrophin.

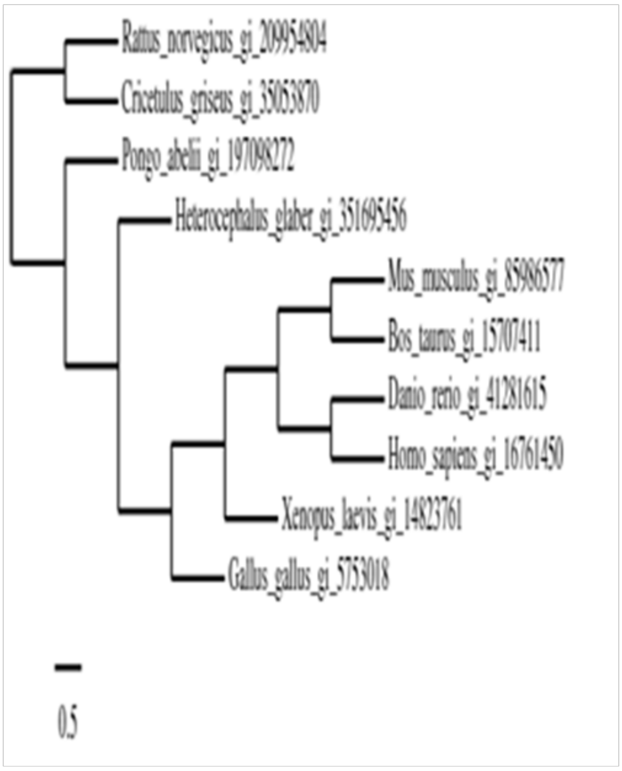
Rattus_norvegicus_gi_209954804
Cricetulus_griseus_gi_35053870
Pongo_abelii_gi_197098272
Heterocephalus_glaber_gi_351695456
Mus_musculus_gi_85986577
Bos_taurus_gi_15707411
Danio_rerio_gi_41281615
Homo_sapiens_gi_16761450
Xenopus_laevis_gi_14823761
Gallus_gallus_gi_5753018
0.5

Figure 12 Phylogenetic tree of Fimbrin.

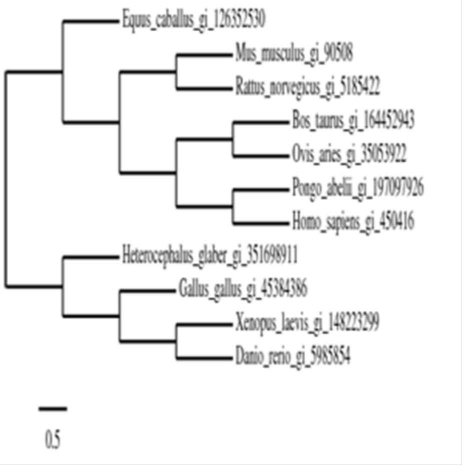
Equus_caballus_gi_126352530
Mus_musculus_gi_90508
Rattus_norvegicus_gi_5185422
Bos_taurus_gi_164452943
Ovis_aries_gi_35053922
Pongo_abelii_gi_197097926
Homo_sapiens_gi_450416
Heterocephalus_glaber_gi_351698911
Gallus_gallus_gi_45384386
Xenopus_laevis_gi_148223299
Danio_rerio_gi_5985854
0.5

Figure 13 Phylogenetic tree of Gelsolin.

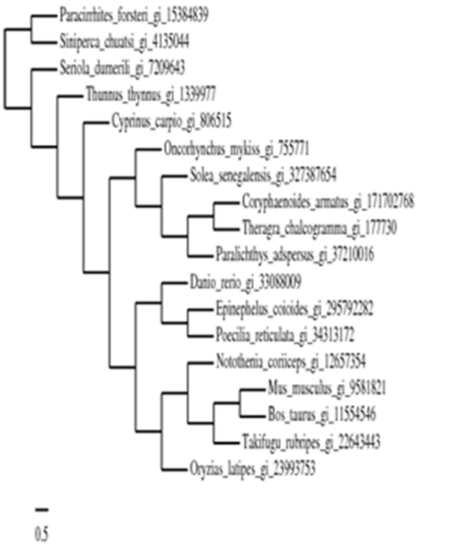
Paracirrhites_forsteri_gi_15384839
Siniperca_chuatsi_gi_4135044
Seriola_dumerili_gi_7209643
Thunnus_thynnus_gi_1339977
Cyprinus_carpio_gi_806515
Oncorhynchus_mykiss_gi_755771
Solea_senegalensis_gi_327387654
Coryphaenoides_armatus_gi_171702768
Theragra_chalcogramma_gi_177730
Paralichthys_adspersus_gi_37210016
Danio_rerio_gi_33088009
Epinephelus_coicides_gi_295792282
Poecilia_reticulata_gi_34313172
Notothenia_coriiceps_gi_12657354
Mus_musculus_gi_9581821
Bos_taurus_gi_11554546
Takifugu_rubripes_gi_22643443
Oryzias_latipes_gi_23993753
0.5

Figure 14 Phylogenetic tree of Myosin.

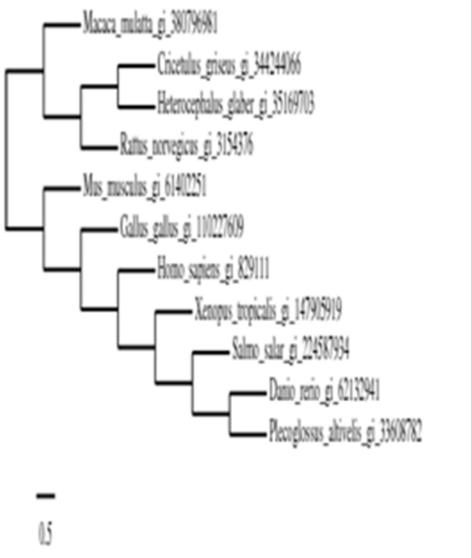
Macaca_mulatta_gi_380796981
Cricetulus_griseus_gi_344244066
Heterocephalus_glaber_gi_35169703
Rattus_norvegicus_gi_3154376
Mus_musculus_gi_61402251
Gallus_gallus_gi_110227609
Homo_sapiens_gi_829111
Xenopus_tropicalis_gi_147905919
Salmo_salar_gi_224587934
Danio_rerio_gi_62132941
Plecoglossus_altivelis_gi_33608782
0.5

Figure 15 Phylogenetic tree of Spectrin.

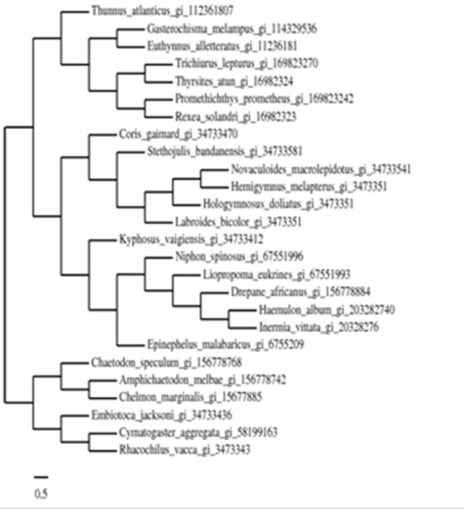
Thunnus_atlanticus_gi_112361807
Gasterochisma_melampus_gi_114329536
Euthynnus_alletteratus_gi_11236181
Trichiurus_lepturus_gi_169823270
Thyrsites_atun_gi_16982324
Promethichthys_prometheus_gi_169823242
Rexea_solandri_gi_16982323
Coris_gaimard_gi_34733470
Stethojulis_bandanensis_gi_34733581
Novaculoides_macrolepidotus_gi_34733541
Hemigymnus_melapterus_gi_3473351
Hologymnosus_doliatus_gi_3473351
Labroides_bicolor_gi_3473351
Kyphosus_vaigiensis_gi_34733412
Niphon_spinosus_gi_67551996
Liopropoma_eukrines_gi_67551993
Drepane_africanus_gi_156778884
Haemulon_album_gi_203282740
Inermia_vittata_gi_20328276
Epinephelus_malabaricus_gi_6755209
Chaetodon_speculum_gi_156778768
Amphichaetodon_melbae_gi_156778742
Chelmon_marginalis_gi_15677885
Embiotoca_jacksoni_gi_34733436
Cymatogaster_aggregata_gi_58199163
Rhacochilus_vacca_gi_3473343
0.5

Figure 16 Phylogenetic tree of Ttitin.

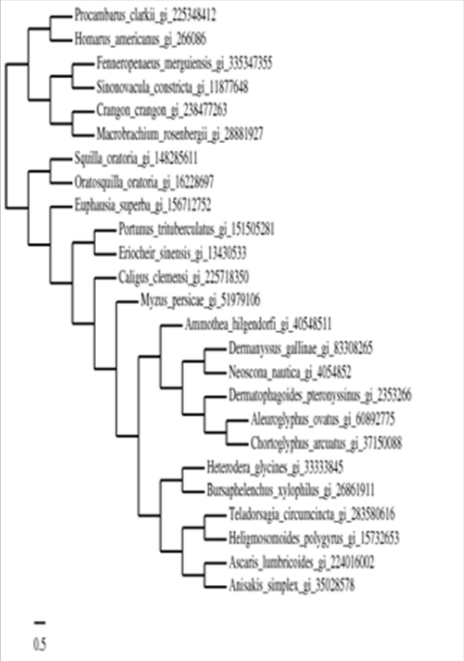
Procambarus_clarkii_gi_225348412
Homarus_americanus_gi_266086
Fenneropenaeus_merguiensis_gi_335347355
Sinonovacula_constricta_gi_11877648
Crangon_crangon_gi_238477263
Macrobrachium_rosenbergii_gi_28881927
Squilla_oratoria_gi_148285611
Oratosquilla_oratoria_gi_16228697
Euphausia_superba_gi_156712752
Portunus_trituberculatus_gi_151505281
Eriocheir_sinensis_gi_13430533
Caligus_clemensi_gi_225718350
Myzus_persicae_gi_51979106
Ammothea_hilgendorfi_gi_40548511
Dermanyssus_gallinae_gi_83308265
Neoscona_nautica_gi_4054852
Dermatophagoides_pteronyssinus_gi_2353266
Aleuroglyphus_ovatus_gi_60892775
Chortoglyphus_arcuatus_gi_37150088
Heterodera_glycines_gi_33333845
Bursaphelenchus_xylophilus_gi_26861911
Teladorsagia_circumcincta_gi_283580616
Heligmosomoides_polygyrus_gi_15732653
Ascaris_lumbricoides_gi_224016002
Anisakis_simplex_gi_35028578
0.5

Figure 17 Phylogenetic tree of Tropomyosin.

BindN[3] was used for prediction of DNA and RNA binding residues in order to understand the function of DNA and RNA binding proteins. Filamin protein was found to have greater number of DNA and RNA binding residues. In filamin 262 DNA residues with 303 RNA residues were predicted. In plastin 3 protein 103 DNA and 122 RNA residues were found. In troponin predicted DNA residues were 81 and RNA residues were 121 in number. Thus BindN showed that selected fish muscle proteins are good binding proteins. ConSurf[4] was explored for estimation of evolutionary conserved amino acids in protein which was based on phylogenetic relationship inferred from homologous sequences. In actin number of functional residue predicted was 21 whereas in myosin 17 residues, in dystrophin 6, in titin 33, in spectrin 32, in M2 protein 26 amino acids were highly conserved and exposed. Filamin protein was found to have high number of functionally conserved amino acids with 225 residues. Study of conserved position of these amino acids contributes to structural and functional knowledge. Thus from ConSurf study it was found these muscle proteins have structural and functional importance.

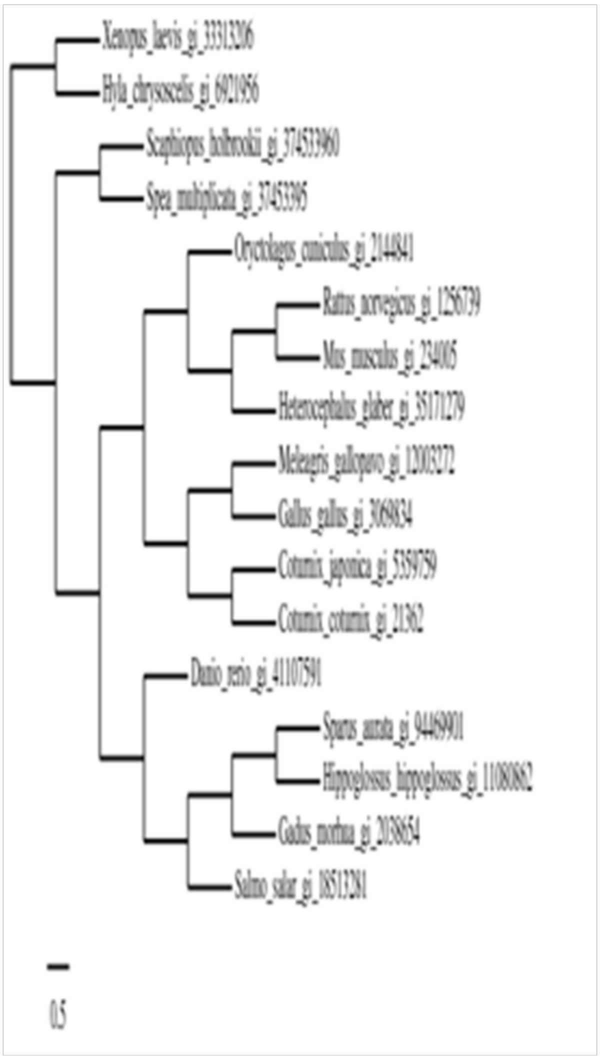
Xenopus_laevis_gi_33313206
Hyla_chrysoscelis_gi_6921956
Scaphiopus_holbrookii_gi_374533960
Spea_multiplicata_gi_37453395
Oryctolagus_cuniculus_gi_2144841
Rattus_norvegicus_gi_1256739
Mus_musculus_gi_234005
Heterocephalus_glaber_gi_35171279
Meleagris_gallopavo_gi_12003272
Gallus_gallus_gi_3069834
Coturnix_japonica_gi_5359759
Coturnix_coturnix_gi_21362
Danio_rerio_gi_41107591
Sparus_aurata_gi_94469901
Hippoglossus_hippoglossus_gi_11080862
Gadus_morhua_gi_2038654
Salmo_salar_gi_18513281
0.5

Figure 18 Phylogenetic tree of Troponin.

DiANNA[5] was used for cysteine classification and prediction of disulfide connectivity. In gelsolin, plastin 3 and M2 protein four disulfide bonds were predicted. In M1 protein 2 disulfide bonds whereas in spectrin 2 and titin 1 disulfide bond was predicted. Filamin protein was found to have greater number of disulfide bond. Two cysteines were predicted in spectrin 2, and actinin. In titin 3, in plastin 38, in M2 protein 11, in M1 protein 5, in gelsolin 9 and in filamin 21cysteines were predicted. This knowledge helps us to understand secondary structure of protein since disulfide bonds bridges play important role for stabilizing the folding process in protein. In addition knowledge of disulfide bond with cysteine also provides information for genome annotation. PROFEAT[6] is a bioinformatics server used for calculating structural and chemical features of protein from primary sequence data. These features provides knowledge about biological properties of proteins and peptides. Thus in order to compute the structural and physicochemical features of proteins and peptides PROFEAT was used. All fish muscle proteins were found as non allergen.

DEEP VIEW[11] was used for analyzing secondary structure features such as coils, ribbons, acidic and basic residues. In gelsolin 158 helices, 252 strands and 321 coils were predicted. In actin 37 helices, 28 strands and 39 coils were predicted. In actinin 71 helices and 40 coils were predicted. Dystrophin was found to contain 15 helices and 26 coils. In M2 protein 12 helices, 86 strands and 93 coils were predicted. 139 helices and 117 coils were predicted from plastin 3D model. In case of thymosin 33 helices where as in troponin 67 helices were predicted with 10 coils in both proteins. PSIPRED integrates several protein structure prediction methods on one platform. PSIPRED[9] was used for prediction of protein structure, transmembrane topology prediction and for recognition of folds and domains. Homology modeling approach was used to predict three dimensional structures. Homology modeling refers to prediction of tertiary structure of protein of interest using template of known 3D structure with homologous sequence. WHAT IF and PSVS[11] was used for structure validation and evaluating stereochemistry of 3D model. The identification of a conserved domain footprint may be the only clue towards cellular or molecular function of a protein, as it indicates local or partial similarity to other proteins, some of which may have been characterized experimentally.[15] Template used for predicting 3D model was 1D4X_A for actin, ITJT_A for actinin, 1DXX_A for dystrophin, 2FGH_A for gelsolin, 2JDF_A for M2 protein, 1AOA_A for plastin 3, 1HJO_A for thymosin and 1JID_E for tropnin. After validation 3D models were successfully submitted to PMDB[15] as PM0078304, PM0078303, PM0078298, PM0078299, PM0078300, PM0078301, PM0078302 and PM0078305. Protein 3D structure is important in understanding protein interactions, function and their localization.[19] Structure prediction refers to the prediction of 3D structure from its amino acid sequence. Number of motifs found in actin was 16, in actinin 148, in dystrophin 17, in gelsolin 102, in M2 protein 147, in plastin 108, in thymosin 26 and in troponin 125. CDD[14] is a large resource which contains manually curates domain models and provides information about sequence, structural and functional relationship. Six domains were predicted in gelsolin and plastin 3. In actin, actinin, dystrophin and thymosin one domain was found. The main objective of this study was to explore the structural and functional importance of novel fish muscle proteins.

Table 7 Summarized table with statistical measurements of phylogenetic tree including protein, number of taxa, likelihood log, parsimony, tree size, gamma Shape parameter, mean, standard deviation and median analyzed by DIVEIN server

Sr.#	Proteins	No. of taxa	Log Likelihood	Parsimony	Tree size	Gamma shape parameter	Mean	S.D	Median
1	Actin	11	-1038.78198	39	0.76396	0.529	0.1401248	0.2126536	0.0354923
2	Actinin	12	-4490.72736	354	1.26628	0.585	0.325363	0.1710674	0.3750267
3	Dystrophin	14	-7484.16288	3435	13.20001	2.06809	0.4517532	0.4039974	0.2799059
4	Fimbrin	10	-4236.3815	448	0.92222	0.799	0.2598686	0.147638	0.3098932
5	Gelsolin	11	-5976.75543	691	1.37231	0.699	0.3301184	0.2391237	0.1970655

Table continued....

Sr.#	Proteins	No. of taxa	Log Likelihood	Parsimony	Tree size	Gamma shape parameter	Mean	S.D	Median
6	Spectrin	11	-9587.44722	363	0.36001	0.153	0.0774005	0.0494607	0.0822327
7	Myosin	18	-11111.8306	984	1.16158	0.784	0.2049471	0.1153865	0.1622767
8	Titin	26	-875.18009	49	0.30294	0.897	0.0445636	0.0248786	0.042512
9	Tropomyosin	25	-2981.71234	400	1.75808	0.471	0.1964246	0.1039309	0.1429878
10	Troponin	17	-3877.24952	490	5.14144	0.352	0.5318019	0.2927219	0.3902863

Fish muscle[1] was found as an excellent model for performing sequence and structural analysis. Sequence analysis was carried out using different bioinformatics tools to understand structure, function and evolution of fish muscle proteins with significant features. Homology modeling technique was applied for predicting 3D structure. This 3D structure is important in understanding protein interaction, function and its localization. Structural knowledge has allowed us to identify functionally important residues and disulfide linkages. Furthermore 3D knowledge of proteins will contribute to design efficient drugs. Phylogenetic analysis of ten fish muscle proteins including actin, actinin, fimbrin, dystrophin, myosin, gelsolin, titin, spectrin, tropomyosin, and troponin were performed. In order to construct phylogenetic trees sequences were aligned by Clustal X using gap opening penalty 30, gap extension penalty 20 and GONNET protein weight matrix.[16] The phylogenetic tree was generated in MEGA 4 using maximum Likelihood approach.[17] The bootstrap was performed using 1000 replications.[20] Thus, novel type of relationship was predicted among species including *Amphistichus koelzi, Oryzias latipes, Dicentrarchus labrax, Plecoglossus altivelis, Danio rerio, Salmo salar, Macrobrachium rosenbergii* and *Anisakis simplex.*

Comparative study of actin, actinin, plastin3 or fimbrin, gelsolin, myosin, spectrin, tropomyosin and troponin fish protein revealed the genetic divergence into two major lineages. Phylogenetic topology of titin and dystrophin muscle protein revealed the genetic divergence into four lineages. The phylogenetic study have application in various fields of biology including systematic, bioinformatics and comparative genomics. Statistically phylogenetic trees were analyzed by DIVEIN predicting number of taxa, values of log likelihood, gamma shape parameter, mean, standard deviation and median. Titin was found to include highest number of taxa with 26 species a and smaller number of taxa was observed in Fimbrin protein with 10 species. This comparative study will be beneficial for predicting the function of individual genes and mechanism of inherited diseases by comparing the genetic material of different species.

Conclusion

Overall evidence from *in silico* approaches revealed that fish muscle proteins have structural and functional significance. Future functional research can be conducted via exploring the proteins of model organisms for using it as a diagnostic tool for designing effective vaccines utilizing structure based drug designing approach.

Acknowledgements

We acknowledge the entire team member specially Prof. Vasco Azevedo and Prof. Bilal Hussain for their support.

Conflict of interest

The author declares no conflict of interest.

References

1. Fernandez DA, Calvo J. Fish muscle: the exceptional case of Notothenioids. *Fish Physiol Biochem*. 2009;35(1):43–52.
2. Campbell AN. Comparative Proteomics Kit 1:Protein Profile Module. *Bio Rad Laboratories*. 2006:1049–1051.
3. Wang L, Brown SJ. BindN: a web based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res*. 2006;34(2):243–248.
4. Ashkenazy H, Erez E, Martz E, et al. ConSurf 2010:calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*. 2010;38(Web Server Issue):W529–W533.
5. Ferrè F, Clote P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res*. 2005;33(Web Server Issue):W230–W232.
6. Rao HB, Zhu F, Yang GB, et al. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequences. *Nucleic Acids Res*. 2011;39(Web Server issue):W385–W390.
7. Gasteiger E, Gattiker A, Hoogland C, et al. ExPASY: the proteomics server for in depth protein knowledge and analysis. *Nucleic Acids Res*. 2003;31(13):3784–3788.
8. Arnold K, Kiefer F, Kopp J, et al. The Protein Model Portal. *J Struct Funct Genomics*. 2009;10(1):1–8.
9. Koussounadis A, Redfern OC, Jones DT. Improving classification in protein structure databases using text mining. *BMC Bioinformatics*. 2009;10:129.
10. Petersen B, Lundegaard C, Petersen TN. NetTurnP: Neural Network Prediction of Beta turns by use of Evolutionary information and Predicted Protein Sequence Features. *PLoS One*. 2010;5(11):11.e15079.
11. Idrees S, Nadeem S, Kanwal S, et al. In silico sequence analysis, homology modeling and function annotation of Ocimumbasilicum hypothetical protein GICT28_OCIBA. *Journal of Bioautomation*. 2012;16(2):111–118.
12. Chen CC, Hwang JK, Yang JM. PS2: protein structure prediction server. *Nucleic Acids Res*. 2006;34(Web Server Issue):152–157.
13. Jensen LJ, Gupta R, Blom N, et al. Prediction of human protein functions from post translational modifications and localization features. *J Mol Biol*. 2002;319(5):1257–1265.
14. Marchler–Bauer A, Lu S, Anderson JB, et al. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res*. 2011;39(Database Issue):D225–D229.
15. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. *Proteins*. 2007;66(4):778–795.
16. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947–2948.

17. Kumar S, Nei M, Dudley J, et al. MEGA: A biologist centric software for evolutionary analysis of DNA and protein sequence. *Brief Bioinform*. 2008;9(4):299–306.

18. Deng W, Maust BS, Nickle DC, et al. DIVEIN: A web server to analyze Phylogenies, Sequence divergence, Diversity and Informative sites. *Biotechniques*. 2010;48(5):405–408.

19. Jones DT. Protein structure prediction in genomics. *Brief Bioinform*. 2001;2(2):111–125.

20. Pavlopoulos GA, Soldatos TG, Barbosa–Silva A, et al. A reference guide for tree analysis and visualization. *BioData Min*. 2010;3(1):1.

Citation: Khalid S, Idrees S, Khalid H, et al. *Ab-initio* prediction of sequence and structural biology of fish muscle proteins using homology modeling, phylogeny and different computational approaches. *MOJ Proteomics Bioinform.* 2015;2(3):81–91. DOI: 10.15406/mojpb.2015.02.00047