Structure Prediction and Modeling of Endochitinase Protein of *Trichoderma harzianum* *(Th* Azad*)* to improve Its Stability and Efficiency

Antima Sharma, Mukesh Srivastava, Sonika Pandey, Anuradha Singh, Vipul Kumar and Mohammad Shahid

Abstract—The study aims at exploring the genome of *Trichoderma harzianum* (*Th* Azad) and the genes involved in inhibiting the growth of soil borne pathogens using bioinformatics tools and techniques. It involves in silico homology modeling of the endochitinase protein of *T. harzianum* and employing the concept of Circular Permutation in an attempt to make it more thermostable and liable at vivid temperature conditions. Therefore, the specific strain (*Th* Azad) of *T. harzianum* species under study is isolated from the soil sample, identified at molecular level using especially *in silico* designed primers, and then using the sequence for further analysis. Conserved motif in the protein sequence of endochitinase is searched in the PROSITE database. It also helps us in determining the protein families and domains that have remain conserved throughout evolution. This particular fragment (or conserved motif) is then subjected to circular permutation at random in the protein sequence in an attempt to increase its thermostability and also enable an improved homology modeling, protein folding and protein design. Permutation by duplication has been proposed in this study to compare the original structure with the artificially engineered endochitinase model. It can be concluded that this study may result in creating a thermodynamically stable and biologically functional protein for protecting crops from soil borne pathogenic fungi. It may also help in elucidating the evolution of endochitinase protein found in *Trichoderma* species and in other pathogenic fungi such as *Fusarium.*

Index Terms—Circular Permutation; Endochitinase; Homology Modeling; Motif.

1 Introduction

It is well known among the agricultural fraternity that *Trichoderma* is an efficient Biological Control Agent that attacks the soil borne pathogens through mycoparasitism. Cell wall degrading enzymes such as chitinases, glucanases, xylanases and proteinases are the key enzymes involved in this mechanism. Though each enzyme has a specific role in the interaction with the pathogen but research on endochitinase gene present in chitinase enzyme has proved beneficial in improving the *Trichoderma* strains. Carsolio *et al.* [1] and Viera [2] studied the role of *Trichoderma harzianum* endochitinase (Ech42) in mycoparasitism by genetically manipulating the gene that encodes ech42. Several transgenic strains were prepared by inserting multiple copies of the gene ech42 that even resulted in an increased biocontrol activity. Lieckfeldt *et al.* [3] carried out a complete phylogenetic analysis of *Trichoderma* based on endochitinase gene stating that this gene can act as a molecular marker for reconstructing phylogeny as compared to ITS-1 and ITS-2 rDNA sequencing. Thus, the importance of endochitinase gene in *Trichoderma* has been exploited and explored in this study in an attempt to improvise and produce a more competent gene product.

Molecular identification of the *Trichoderma* strain under study is done with the help of a universal set of primers i.e., Internal Transcribed Spacer regions. The polymerase chain reaction is fundamental to molecular biology and is the most important practical molecular technique for the research laboratory. Primer design is one of the key steps for successful PCR. For PCR applications, primers are usually 18–35 bases in length and should be designed such that they have complete sequence identity to the desired target fragment to be amplified [4].

Motif discovery in the endochitinase gene of a particular strain of *Trichoderma* is the principle behind protein engineering and develop an improved protein using an uncommon method known as circular permutation. Protein structures involving gene duplication and gene fusion

- *Antima Sharma is currently working in the project running in the Biocontrol Laboratory, Department of Plant Pathology, Chandra Shekhar Azad University of Agriculture & Technology, Kanpur-208002, Uttar Pradesh, India, E-mail:antimasharma11@gmail.com*
- *Mukesh Srivastava is the Principal Investigator of the project running in the Biocontrol Laboratory, Department of Plant Pathology, Chandra Shekhar Azad University of Agriculture & Technology, Kanpur-208002, Uttar Pradesh, India, E-mail:mukeshcsau@rediffmail.com*
- *All other co-authors are also working in the same project.*

events can be highly illuminated, and researchers have long sought fundamental explanations for evolutionary origins of duplicated protein structures. Circular permutations are a frequent event in molecular evolution, and they have been observed in many protein families and superfamilies [5]. However, circular permutation can perturb local tertiary structure, resulting in improved protein catalytic activity. Protein engineering has benefits of reorganizing the polypeptide chain of a protein by circular permutation [6].

Circular permutation, wherein the original termini of a protein are concatenated and new termini are generated elsewhere within the sequence, is a general protein engineering strategy to produce full-length, active recombinant protein. It is widely known that terminal residues of proteins (i.e. the N- and C-termini) are predominantly located on the surface of proteins and exposed to the solvent. N-terminal region tend to adopt an extended beta-strand conformation while C-terminal regions are often helical.

In this study, Permutation by duplication [7,8] has been applied where the motif is duplicated and inserted at random in the protein sequence. The structure is modeled in Swiss Model Automated Workspace [9] and the structural features are then determined in UCSF Chimera software [10].

Identifying structurally similar proteins with different chain topologies, including circular permutation, can aid studies in homology modeling, protein folding, and protein design. An algorithm that can structurally align two proteins independent of their backbone topologies would be an important tool [11].

The artificially engineered model of endochitinase protein is compared with the original model to find the differences in terms of stability and occurrence.

2 Materials & Methodology

The nucleotide sequence (chi1) coding for endochitinase enzyme in *Trichoderma harzianum* is taken from NCBI database with the accession number U49455.1 and then is used for *in silico* primer designing using FastPCR program.

The protein sequence of endochitinase gene was then used for searching conserved motifs that play an important role in the degradation of cell walls of fungal pathogens. PROSITE is an annotated collection of motif descriptors used for the identification of protein families and domains [12]. Thus, ScanProsite was used to search for specific patterns in the protein of interest and also to determine their families [13]. The searched motif is then validated for its presence in *Trichoderma* species by performing a similarity search to validate the prediction and check if the discovered motif belongs to all the strains of *Trichoderma* species.

Homology modeling of the proteins was done with the help of Swiss Model Automated Workspace and UCSF Chimera was used for structure visualization and calculating structural properties.

3 Results & Discussion

The gene sequences of the enzymes involved in fungal cell wall degradation are identified first that include chitinases, glucanases, proteases, xylanases etc. but chitinase is considered to be the unique enzyme that is found in a variety of *Trichoderma* strains [2]. Thus, chitinase protein sequences are retrieved from the NCBI protein database and used for further analyses. Primer designing is essential in the first step as it helps in quick and easy identification of a specific gene of interest thus an *in silico* approach is used for designing specific forward and reverse primers. The primer pair that was found to meet all the preferred criteria such as melting temperature, primer quality, length, no self annealing and no hairpin formation was selected for further studies on amplification of specific endochitinase gene present in *Trichoderma* species.

3.1 Primer Pair and Their Melting Temperatures

Forward primer 5′- ACCAACTGGGGCATCTACGA -3′ (Tm=59.0˚C)

Reverse primer 5′- TCCAAGAATCATCGGCATAGTGC -3′ (Tm=57.5˚C)

This primer pair is used for the amplification of the endochitinase gene in *T. harzianum Th* Azad in order to confirm its presence (**Figure 1**).

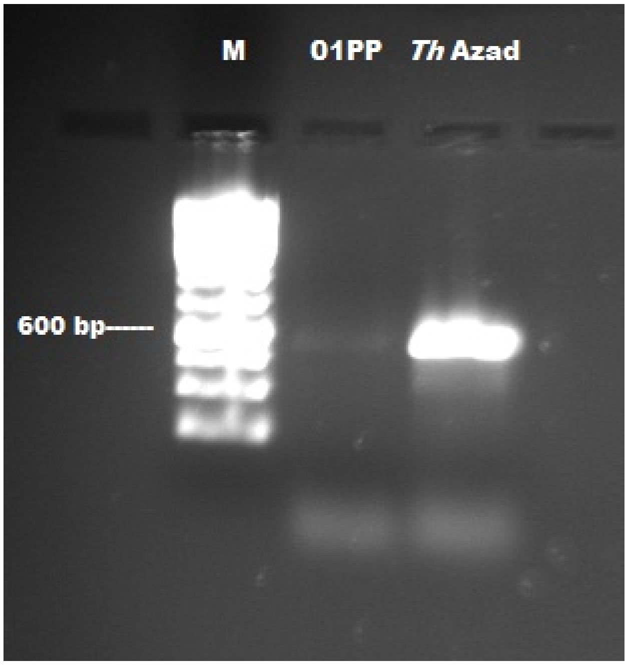


Fig. 1. Amplified fragment of endochitinase present in *T. harzianum Th* Azad (where M denoted Marker)

Thereafter, the primary amino acid sequence of endochitinase protein (Accession No. AAA98644.1) is then proceeded for the secondary structure prediction that is carried out in the Swiss Model Automated Workspace. It performs a step by step procedure of template assessment, domain annotation and structure assessment of the protein sequence entered by the user. Once the model is generated, it is visualized in UCSF Chimera software that gives a complete description of the predicted model such as the numbers of chains, strands, helices and loops in the three-dimensional protein structure.

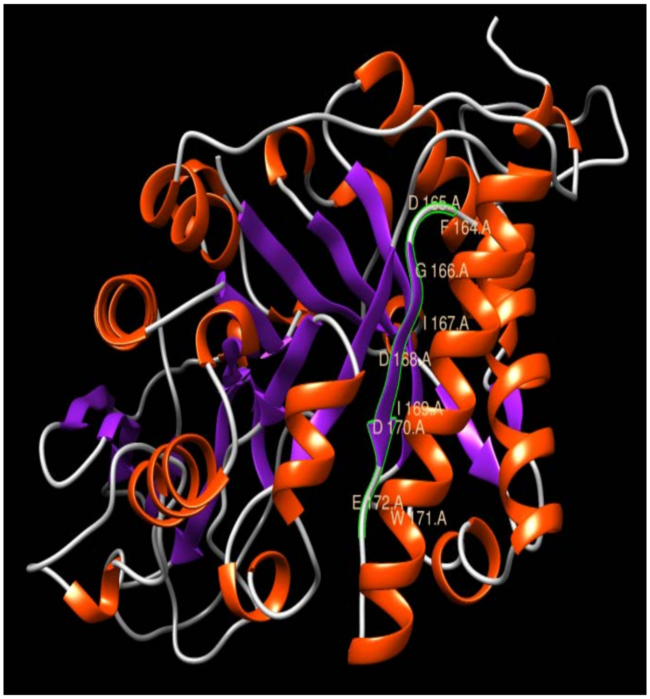


Fig. 2. Modeled endochitinase (as seen in UCSF Chimera) where the predicted motif is highlighted in green outline with labels. The structure is colored based on the secondary structures coloring scheme ('helix' in orange, 'strand' in magenta and 'coil' in grey).

Motifs are targeted in this study as they are believed to remain conserved throughout evolution of a particular organism. Regulatory motifs are short DNA sequences that are used to control the expression of genes, dictating the conditions under which a gene will be turned on or off. Each motif is typically recognized by a specific DNA-binding protein called a transcription factor (TF). A transcription factor binds precise sites in the promoter region of target genes in a sequence-specific way, but this contact can tolerate some degree of sequence variation. Thus, different binding sites may contain slight variations of the same underlying motif, and the definition of a regulatory motif should capture these variations while remaining as specific as possible.

The direct identification of regulatory motifs presents numerous challenges. By their nature, they are very short (6 to 15 bp), frequently degenerate and can appear at varying distances and orientations upstream of target genes. Unlike genes that contain clear start and stop codons, as well as well-defined splicing signals, motifs have no detectable sequence features and they are indistinguishable from random sequences of the same length. Their identification has thus relied heavily on experimental intervention, such as mutational analysis of promoter regions, or genomewide gene expression studies under various environmental cell perturbations.

The protein sequence of the endochitinase gene is used as an input to search for motifs against PROSITE database. This resulted in the prediction of a peptide (or motif) of 9 amino acids length that was found to be similar to the PROSITE hit **PS01095**. The chitinase gene entered in the PROSITE database matched with a pattern "**FDGIDIDwE**" (position 164-172) that belongs to Chitinases family 18 active site and has a Glutamate (E) residue at position 172 in the entered gene sequence.

The amino acid sequence of *Trichoderma harzianum* endochitinase gene displaying the predicted motif (highlighted in yellow color and '**E**' is the active site) is shown below:

MLSFLGKSVALLAALQATLSSASPLATEERSVEKRANGY
ANSVYFTNWGIYDRNFQPADLVASDVTHVIYSFMNLQ
ADGTVISGDTYADYEKHYADDSWNDVGTNAYGCVKQ
LFKVKKANRGLKVLLSIGGWTWSTNFPSAASTDANRK
NFAKTAITFMKDWGFDGIDIDWEYPADATQASNMILLL
KEVRSQLDAYAAQYAPGYHFLLTIAAPAGKDNYSNVR
LADLGQVLDYINLMAYDYAGSFSPLTGHDANLFNNPS
NPNATPFNTDSAVKDYINGGVPANKIVLGMPIYGRSFQ
NTAGIGQTYNGVGSGSWEAGIWDYKALPKAGATVQY
DSVAKGYYSYNSATKELISFDTPDMINTKVAYLKSLGLG
GSMFWEASADKKGADSLIGTSHRALGGLDTTQNLLSYP
NSKYDNIKNGLN

The PROSITE description describes that Chitinases (EC 3.2.1.14) are enzymes that catalyze the hydrolysis of the β-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases.

Site directed mutagenesis experiments and crystallographic data have shown that a conserved **glutamate (E)** is involved in the catalytic mechanism and probably acts as a **proton donor**. This glutamate is at the extremity of the best conserved region in these proteins.

Consensus pattern: [LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E

E is the active site residue.

The predicted motif is then searched for all similar *Trichoderma/Hypocrea* species using PSI-BLAST program in order to confirm the prediction of motif and also to assign putative function to unknown motifs if found apart from the prediction. The predicted motif was found to be present not only in *Trichoderma harzianum* but also in all closely related strains of *Trichoderma* and *Hypocrea* such as *Trichoderma longibrachiatum, T. atroviride, T. koningii, T. viride and T. virens*, but at different locations.

The 3D model of the reference endochitinase protein is generated in Swiss Model and its Root Mean Square Deviation (RMSD) is checked. Lower the RMSD value more is the stability of the 3D structure. The RMSD of the engineered protein is compared to the original endochitinase model to check for its stability and dynamics. Circular permutation by duplication is applied on the protein sequence of the endochitinase of *Trichoderma harzianum* where the predicted motif is duplicated first and then inserted at a random location in the protein sequence. The amino acid sequence below shows the duplication of the motif and insertion of the same at random locations:

MLSFLGKSVALLAALQATLSSASPLATEERSVEKRANGY
ANSVYFTNWGIYDRNFQPADLVASDVTHVIYSFMNLQ
ADGTVISGDTYADYEKHYADDSWNDVGTNAYGCVKQ
LFKVKKANRGLKVLLSIGGWTWSTNFPSAASTDANRK
NFAKTAITFMKDWGYPADATQASNMILLLKEVRSQLD
AYAAFDGIDIDWEFDGIDIDWEQYAPGYHFLLTIAAPA
GKDNYSNVRLADLGQVLDYINLMAYDYAGSFSPLTGH
DANLFNNPSNPNATPFNTDSAVKDYINGGVPANKIVL
GMPIYGRSFQNTAGIGQTYNGVGSGSWEAGIWDYKAL
PKAGATVQYDSVAKGYYSYNSATKELISFDTPDMINTK
VAYLKSLGLGGSMFWEASADKKGADSLIGTSHRALGGL
DTTQNLLSYPNSKYDNIKNGLN

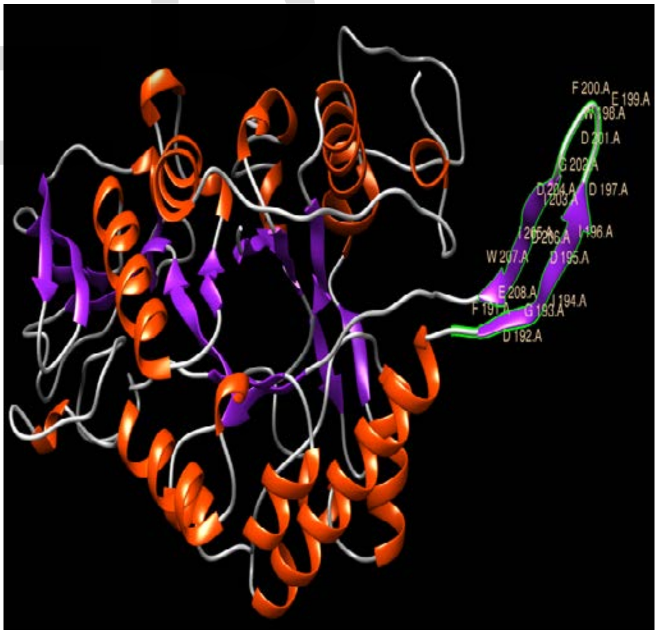


Fig. 3. Modeled circularly permuted endochitinase (as seen in UCSF Chimera). The structure is colored based on the secondary structures coloring scheme ('helix' in orange, 'strand' in magenta and 'coil' in grey). The permuted region (labeled and highlighted) is seen to be on the exposed surface of the protein.

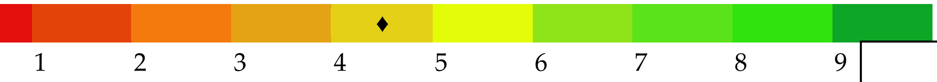
The two modeled structures were then superimposed on each other to calculate the overall RMSD values. When the circularly permuted model was matched with the reference model of endochitinase, its RMSD decreased from 0.369 to 0.108 Angstroms and their percent identity came out to be 95.35. Thus, it is quite clear from the experimental data that the engineered protein created out of circular permutation is thermodynamically more stable and favorable as well. Peisajovich *et al.* [14] also demonstrated the evolutionary feasibility of permutation via duplication by creating functional intermediates at each step of the permutation by duplication model for DNA methyltransferases. Proteins resulting from gene rearrangements and circular permutations have been reported to show high catalytic activity and with different topologies that can provide new insights to protein evolution and protein folding.

PROSESS (Protein Structure Evaluation Suite & Server) is a web server designed to evaluate and validate protein structures [15]. Thus, the global structure assessment of the permuted protein is carried out in this server that displays the following results:

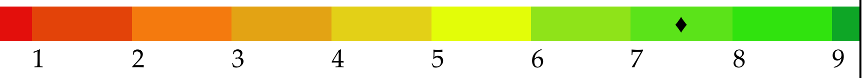
Chain	:	A
Helix%	:	27
Beta-Strand	:	31
Turn%	:	24
Coil%	:	42
Protein Length	:	396

3.2 Global Structure Assessment

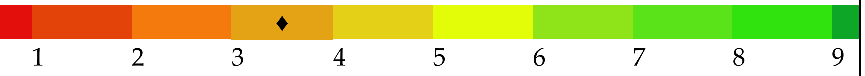
Overall Quality = 4.5



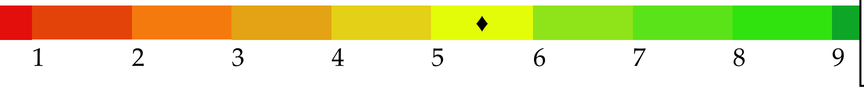
Covalent Bond Quality = 7.5



Non-Covalent/Packing Quality = 3.5



Torsion Angle Quality = 5.5



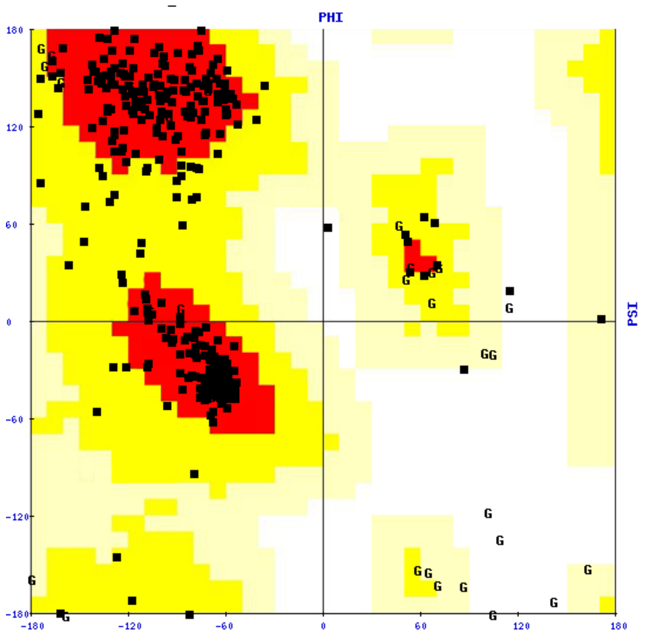


Fig. 4. Ramachandran plot of the permuted protein where most of the amino acids are found to be present in the allowed region and only Glycine (G) residues are present in the disallowed region

The Ramachandran plot above shows the residues present in the protein where red color denotes that there is no steric hindrance, yellow color denotes some steric hindrance and white color denotes the forbidden zone (where only glycine "G" is found as it has no side chains). The table below enlists the number of outliers found in the protein where it is clearly seen that outliers are minimum thus making the protein more stable and thermodynamically favorable.

TABLE 1
DESCRIPTION OF OUTLIERS

Residue class	Percent of outliers	Number of outliers	Total number
General (non-Gly, non-Pro, non-pre-Pro)	0.91	3	330
Glycine	0.00	0	36
Proline	14.29	2	14
Pre-Proline	0.00	0	14

4 Conclusion

Trichoderma harzianum is, however, an effective biocontrol agent against *F. oxysporum* on several crops. The lack of mycoparasitic interaction between *T. harzianum* and *F. oxysporum* indicates that this mechanism is unimportant in this specific system. Hence, other mode of action such as identifying the cell wall degrading enzymes and engineering them using *in silico* approach to improve their potential aspects seems to be a challenging task in the present scenario. Making the protein more thermostable and liable at vivid environmental conditions through protein engineering has paved way for the scientists to develop better strains of *Trichoderma* species containing such kind of proteins that would not only enhance the potency of the strain but would also add innovation to the strain in an attempt to protect crops from fungal diseases. It can also be concluded that the permuted protein endochitinase of *Trichoderma harzianum* can undergo gene rearrangements and chain topologies leading to divergence of new favorable proteins.

Circular Permutation in proteins helps in better understanding of the protein evolution and functionality. Inserting multiple copies of the gene is a common and well adapted procedure to enhance the biological activity of a protein but the work described in this paper has enunciated a novel approach in the area of agricultural bioinformatics. The principle of this research is based upon protein engineering by playing with the functional and conserved regions (or motifs) as they are believed to be the binding sites of many inducers that help in catalyzing or suppressing the biological activity of the protein.

5 ACKNOWLEDGEMENT

The authors are grateful to the financial support granted by the ICAR under the Niche Area of Excellence on "**Exploration and Exploitation of *Trichoderma* as an antagonist against soil borne pathogen**" running in Biocontrol Laboratory, Department of Plant Pathology, Chandra Shekhar Azad University of Agriculture and Technology, Kanpur-208002, Uttar Pradesh, India.

6 References

[1] Carsolio C, Benhamou N, Haran S, Cortés C, Gutiérrez A, Chet I, Herrera-Estrella A (1999) Role of the *Trichoderma harzianum* Endochitinase Gene, ech42, in Mycoparasitism. Appl. Environ. Microbiol. 65(3): 929.

[2] Vieira PM, Coelho AS, Steindorff AS, de Siqueira SJ, Silva Rdo N, Ulhoa CJ (2013) Identification of differentially expressed genes from *Trichoderma harzianum* during growth on cell wall of *Fusarium solani* as a tool for biotechnological application. BMC Genomics 14: 177.

[3] Lieckfeldt E, Cavignac Y, Fekete C, Borner T (2000) Endochitinase gene-based phylogenetic analysis of *Trichoderma*. Microbiol. Res. 155: 7-15.

[4] Kalendar R, Lee D, Schulman AH (2011) Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. Genomics 98: 137-144.

[5] Lindqvist Y, Schneider G (1997) Circular permutations of natural protein sequences: structural evidence. Curr Opin Struct Biol. 7 (3): 422-427.

[6] Yu Y, Lutz S (2011) Circular permutation: a different way to engineer enzyme structure and function. Trends Biotechnol 29(1): 18-25.

[7] Ponting RB, Russell RB (1995) Swaposins: circular permutations within genes encoding saposin homologues, Trends Biochem Sci. 20: 179-180.

[8] Lin J, Feng L (2013). The Structural Consequences after Protein Domain Duplication Events. Life Science Journal 10(3): 610-614.

[9] Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling. Bioinformatics 22: 195-201.

[10] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25(13): 1605-1612.

[11] Dundas J, Binkowski TA, DasGupta B, Liang J (2007) Topology independent protein structural alignment. BMC Bioinformatics 8: 388.

[12] Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 3: 265-274.

[13] De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 1: 34 (Web Server issue): W362-5.

[14] Peisajovich SG, Rockah L, Tawfik DS (2006) Evolution of new protein topologies through multistep gene rearrangements. Nature Genetics 38: 168-173.

[15] Berjanskii M, Liang Y, Zhou J, Tang P, Stothard P, Zhou Y, Cruz J, Macdonell C, Lin G, Lu P, Wishart DS (2010) PROSESS: a protein structure evaluation suite and server. Nucleic Acids Res. Webserver Edition.