International Journal of Molecular Biology & Biochemistry.
Volume 1, Number 1 (2013), pp. 9-22
© International Research Publication House
http://www.irphouse.com

Computational Sequence Analysis and Functional Annotation of KGM_05782 Protein of Danaus Plexippus

Meenu Sharma[1]*, Meenu Saini[2] and Maneet Singh[3]

[1]Deptt. of Bioinformatics, ADI Biosolution, SCF 25,
Phase 3b2, Mohali, Punjab, INDIA
[2]Deptt. of Bioinformatics, Kurukshtra University, Haryana, INDIA
[3]ADI Backoffice professionals Pvt Ltd, E-47, Phase-8, Mohali, Punjab, INDIA
**Corresponding Author e-mail: ms.meenu@ymail.com*

Abstract

Background: Orthology properties suggest that the Lepidoptera are the top evolving insect order so far observed. Among them Monarch butterflies are well known for their amazing long distance annual migration similar to vertebrates to reach their overwintering basis of central Mexico by expedition of nearly 4000km. The study is focused to identify the main fuel and energy unit in this insect responsible for such amazing behavior in addition to biochemical effects.

A hypothetical protein constitutes a fraction of proteome but provides bulk of information. The work is an effort put forward to assign functions on the basis of homology and comparative studies to unknown segments. The Utilization of Computational techniques (tools and biological databases) aided in understanding the proteome in a much effective way as compared to the traditional approaches that are majorly time consuming, costly and unpredictable. Current annotation work involves prediction of conserved domain like YjgF_YER057c_UK that belongs to Adenine nucleotide alpha hydrolases superfamily, motifs, ligand Binding site, Homotrimer interaction site, protein structure at various levels and assigning structure, functions, etc that will facilitate further research and correlate these facts with its amazing flight capacity.

Index Terms— Hypothetical Proteins, Comparative analysis, Scientific Data Mining, Sub-cellular location, Interactive Data exploration and discovery, modeling and structure prediction, Pattern Analysis.

1 Introduction

Till date various genome projects have been compiled and some are still in the pipe line.[1] The completion of the genome sequences provides a platform for understanding genetic complexity and elucidating genetic variations contributing to diverse traits and diseases. The Danaus Plexippus has currently become the center of interest with numerous reasons. Monarchs are milkweed specialists, and their evolved chemical defense mechanism has led to the monarch’s widely known involvement in a mimicry complex with the viceroy butterfly [2] and arose the research interests worldwide. The High-throughput biology technologies have yield complete genome sequences and functional genomics data for several organisms.[3] Till date various genome projects have been compiled including humans, animals and plants.[4] Danaus Plexippus also belong to one of the hot genome sequencing project in pipeline [5]. However, it has been found that nearly 50% of genes are often labeled hypothetical, unknown, uncharacterized, unnamed adding up to the hurdle in scientific study and understanding. Inspite of the fact that biological properties, structure and function of proteins encoded by such genes are unknown but they can be predicted with various comparative approaches. [6]

Proteins being the ultimate effecter molecules perform a wide range of functions that may involve transportation, structural components, stimulation, molecular scissors, cascades, etc [7][8]suggesting that these un-annotated and uncharacterized proteins may also play some lethal roles essential for organism’s survivability. The Danaus plexippus belongs to rapidly evolving Lepidoptera Order that contains few hits against our considered protein sequence[5] but still share 70.8% average amino acid identity with Bombyx that belongs to Diptera.[2][5] The integration of wide range of data from the related and similar sequences aids in the characterization of hypothetical proteins and facilitates the functional annotation. The features of the monarch genome and its proteome provide a treasure trove for furthering our understanding of monarch butterfly migration, a solid background for population genetic analyses between migratory and non-migratory populations, and a basis for future genetic comparison of the genes involved in navigation yet to be discovered in other long-distance migrating species, including vertebrates like migratory birds. [2]

2 Methodology

2.1 Sequence retrieval

Danaus Plexippus a well known Monarch butterfly comprises of 272 mb genome which contain 16433 plus proteins, among which approximately 9627 protein are hypothetical or unknown whose function are not clearly defined yet. The work involve first finding of some hypothetical protein sequences from Danaus Plexippus proteiom. The hypothetical protein secquences against Danaus plexippus were retrieved from ftp NCBI database [9] using entrez search engine and were studied. These sequences were explored using NCBI Blast tools namely Blastp. Conservation of various domains was studied and out of these protein sequences one is selected (protein named KGM_05782) and studied further. To analyze the hypothetical protein and to assign their physicochemical, structural and functional properties various bioinformatics tools and databases were cross referred.

2.2 Physicochemical and functional characterization

The hypothetical protein in Danaus plexippus was studied for various physico-chemical properties majorly theoretical Isoelectric point (pI), molecular weight, total number of positive and negative residues, aliphatic index [10], extinction coefficient [11], instability index [12] and grand average hydropathy (GRAVY) [13] using the Expasy's protein server [14]. These basic properties provided a foundation and enlightened the path for the future exploration.

2.3 Conserved domain and family

The idea regarding the conserved domains was also drawn form the Blastp results. We also selected some of the well established online tools and databases for this purpose. PFAM [15] [16], InterProScan [17] [18] and CDD-Blast [19] [20] were to name a few that were used to predict the conserved domains and families present in the unknown protein sequence.

2.5 Prediction of transmembrane Region

A transversing protein can function as a channel protein. Taking this fact into account we performed analysis for detection of transmembrane regions in the hypothetical protein. Online tools working for proteins input like SMART [21] [22], TMHMM [23] [24], PSORT [25] [26] and Phobius [27] [28] server were used to characterize whether the protein is soluble or transmembrane in nature [29].

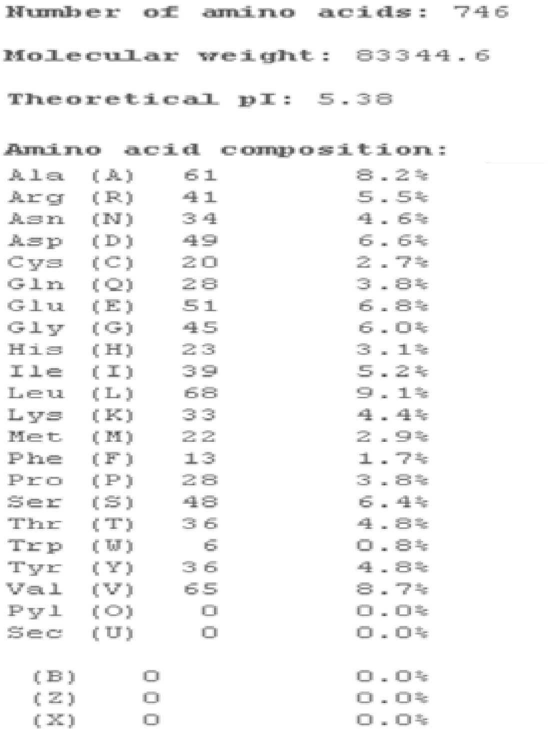
2.6 Protein structure prediction

Analysis is considered incomplete if it cannot exist in the 3D world. For this we proceeded with 3D structure prediction and comparative study with the help of Swiss Model [30] [31] [32], pdbsum [31] [33] [34]. For template based modeling modeler 9v7 [35] [36] was also used. Protein structure visualization is done by SwissPdbviewer [31] [37] and PyMol [38] that revealed some very interesting facts about the considered protein.

3 Results

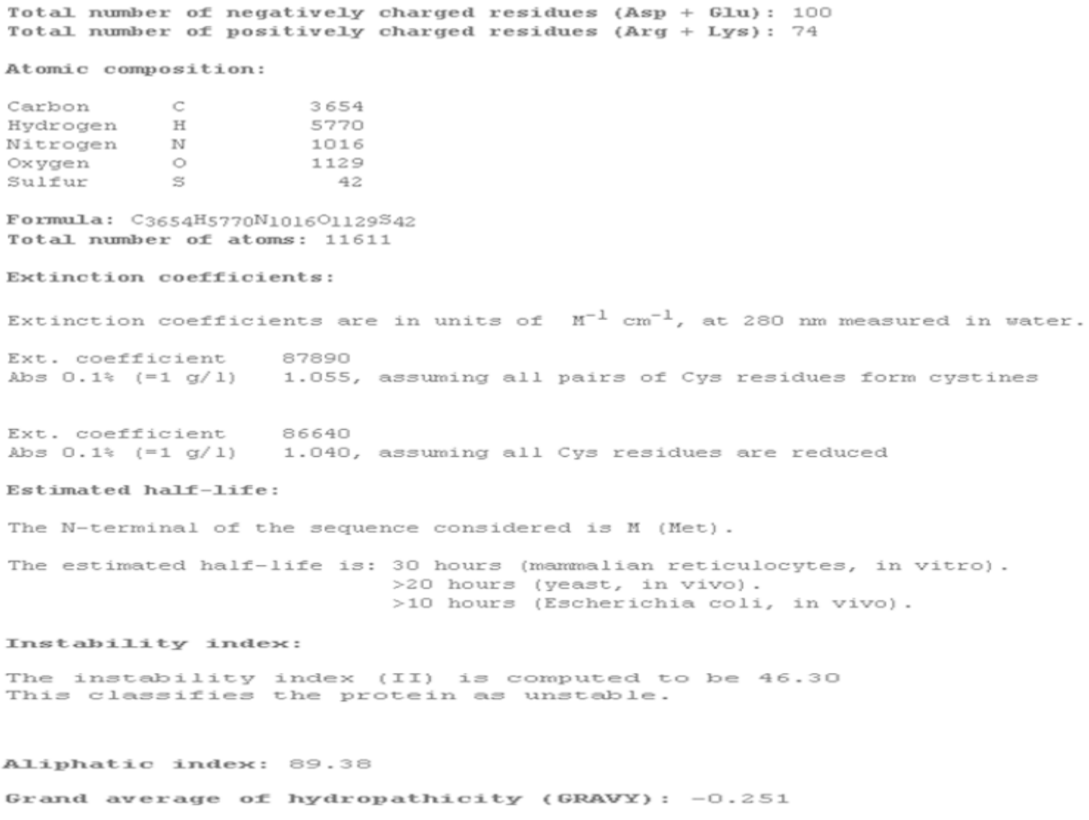
3.1 Physicochemical Analysis

The hypothetical protein KGM_05782 has sequence length of 746aa with all natural occurring amino acids with molecular weight of 83344.6 Da. [Fig1(a)] The isoelectic point 5.38 can be further used for the isolation during wetlab studies. It was also established that the protein had 2.7% cystein residues accounting for 42 sulphur atoms. [Fig1(b)]. It should be noted that disulfide bridges formed by cysteine residues are permanent component of protein primary structure and cysteine is at the center of catalytic site of thiol enzymes. This is further studied and aromatic-sulphur interactions which were found between TYR_{482} - CYS_{495} accounting for 4.049Å and between TRP_{584} - CYS_{544} accounting for 4.26Å using PIC Web browser [Fig1(c), Fig1(d)].



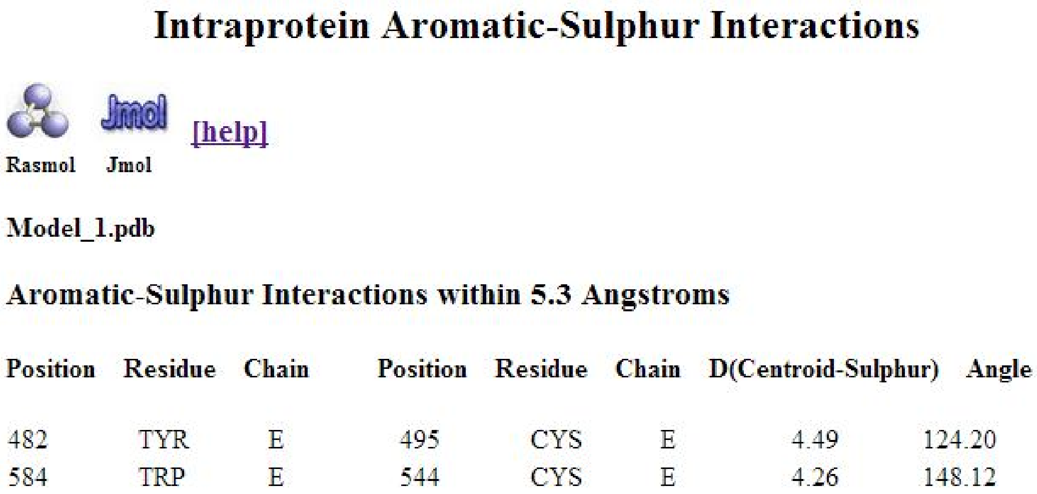
```
Number of amino acids: 746

Molecular weight: 83344.6

Theoretical pI: 5.38

Amino acid composition:
Ala (A)   61          8.2%
Arg (R)   41          5.5%
Asn (N)   34          4.6%
Asp (D)   49          6.6%
Cys (C)   20          2.7%
Gln (Q)   28          3.8%
Glu (E)   51          6.8%
Gly (G)   45          6.0%
His (H)   23          3.1%
Ile (I)   39          5.2%
Leu (L)   68          9.1%
Lys (K)   33          4.4%
Met (M)   22          2.9%
Phe (F)   13          1.7%
Pro (P)   28          3.8%
Ser (S)   48          6.4%
Thr (T)   36          4.8%
Trp (W)    6          0.8%
Tyr (Y)   36          4.8%
Val (V)   65          8.7%
Pyl (O)    0          0.0%
Sec (U)    0          0.0%

 (B)    0             0.0%
 (Z)    0             0.0%
 (X)    0             0.0%
```

(a)



```
Total number of negatively charged residues (Asp + Glu): 100
Total number of positively charged residues (Arg + Lys): 74

Atomic composition:

Carbon      C          3654
Hydrogen    H          5770
Nitrogen    N          1016
Oxygen      O          1129
Sulfur      S            42

Formula: C3654H5770N1016O1129S42
Total number of atoms: 11611

Extinction coefficients:

Extinction coefficients are in units of  M-1 cm-1, at 280 nm measured in water.

Ext. coefficient    87890
Abs 0.1% (=1 g/l)   1.055, assuming all pairs of Cys residues form cystines


Ext. coefficient    86640
Abs 0.1% (=1 g/l)   1.040, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is M (Met).

The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
                            >20 hours (yeast, in vivo).
                            >10 hours (Escherichia coli, in vivo).

Instability index:

The instability index (II) is computed to be 46.30
This classifies the protein as unstable.


Aliphatic index: 89.38

Grand average of hydropathicity (GRAVY): -0.251
```

(b)



Intraprotein Aromatic-Sulphur Interactions

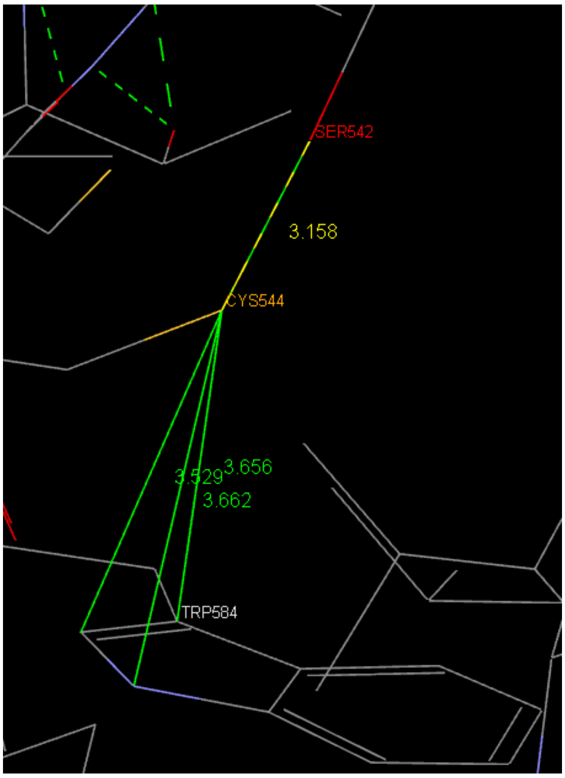
Model_1.pdb

Aromatic-Sulphur Interactions within 5.3 Angstroms

Position	Residue	Chain	Position	Residue	Chain	D(Centroid-Sulphur)	Angle
482	TYR	E	495	CYS	E	4.49	124.20
584	TRP	E	544	CYS	E	4.26	148.12

(c)

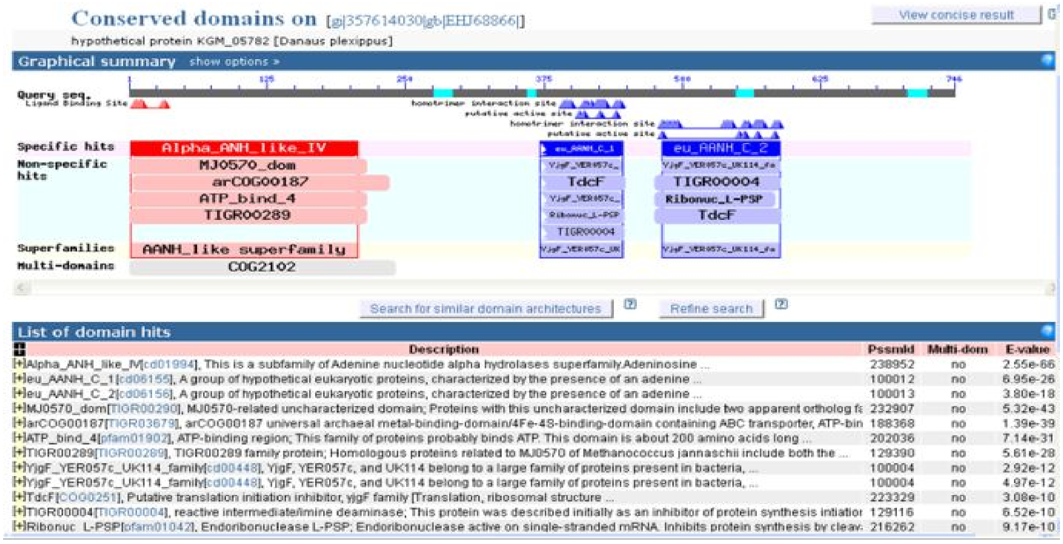
Fig1(a) and (b) shows the physicochemical properties details obtained from expassy and (c) shows the aromatic-sulphur interactions details obtained from PIC Web browser



Fig(d) green lines are interaction links between trp and cys showing aromatic-sulphur interations in a PDB file viewer

3.2 Predicted Domains and Families

For Domain predictions CDD-Blast [Fig 2] along with interproscan was used.



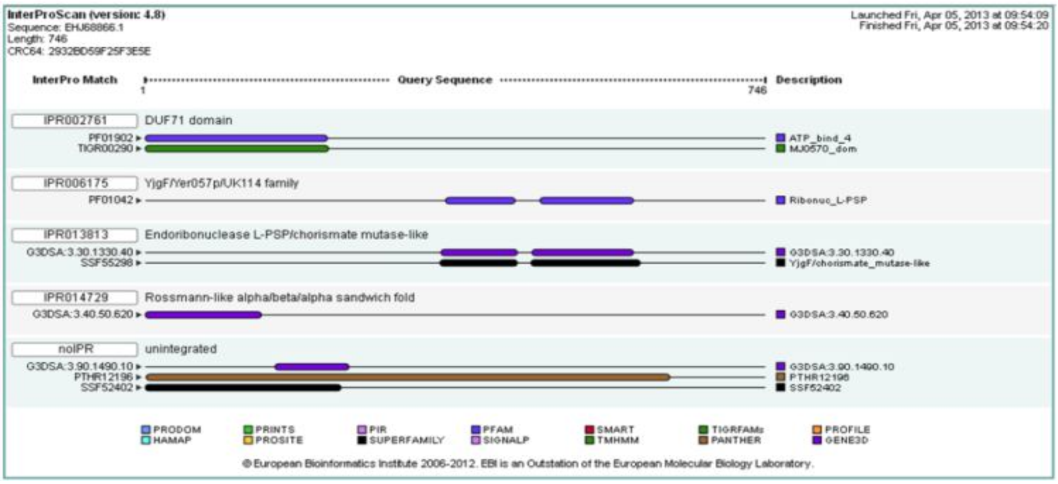
Conserved domains on [gi|357614030|gb|EHJ68866|]

hypothetical protein KGM_05782 [Danaus plexippus]

Description	PssmId	Multi-dom	E-value
[+]Alpha_ANH_like_IV[cd01994], This is a subfamily of Adenine nucleotide alpha hydrolases superfamily.Adeninosine ...	238952	no	2.55e-66
[+]eu_AANH_C_1[cd06155], A group of hypothetical eukaryotic proteins, characterized by the presence of an adenine ...	100012	no	6.95e-26
[+]eu_AANH_C_2[cd06156], A group of hypothetical eukaryotic proteins, characterized by the presence of an adenine ...	100013	no	3.80e-18
[+]MJ0570_dom[TIGR00290], MJ0570-related uncharacterized domain; Proteins with this uncharacterized domain include two apparent ortholog fa	232907	no	5.32e-43
[+]arCOG00187[TIGR03679], arCOG00187 universal archaeal metal-binding-domain/4Fe-4S-binding-domain containing ABC transporter, ATP-bin	188368	no	1.39e-39
[+]ATP_bind_4[pfam01902], ATP-binding region; This family of proteins probably binds ATP. This domain is about 200 amino acids long ...	202036	no	7.14e-31
[+]TIGR00289[TIGR00289], TIGR00289 family protein; Homologous proteins related to MJ0570 of Methanococcus jannaschii include both the ...	129390	no	5.61e-28
[+]YjgF_YER057c_UK114_family[cd00448], YjgF, YER057c, and UK114 belong to a large family of proteins present in bacteria, ...	100004	no	2.92e-12
[+]YjgF_YER057c_UK114_family[cd00448], YjgF, YER057c, and UK114 belong to a large family of proteins present in bacteria, ...	100004	no	4.97e-12
[+]TdcF[COG0251], Putative translation initiation inhibitor, yjgF family [Translation, ribosomal structure ...	223329	no	3.08e-10
[+]TIGR00004[TIGR00004], reactive intermediate/imine deaminase; This protein was described initially as an inhibitor of protein synthesis intiatior	129116	no	6.52e-10
[+]Ribonuc_L-PSP[pfam01042], Endoribonuclease L-PSP; Endoribonuclease active on single-stranded mRNA. Inhibits protein synthesis by cleav-	216262	no	9.17e-10

Fig2: CDD-Blast result page showing conserved Domains, family and superfamilies

In Interpro five domains were predicted namely, DUF71 domain, YjgF/Yer057p/UK114 family, Endoribonuclease L-PSP/chorismate mutase-like, Rossmann –like alpha/beta/alpha sandwich fold and unintegrated domain in cross references with PRODOM, PRINTS, PIR, PFAM, SMART, TIGERFAMS, PROFILES, HAMAP, PROSITE, SUPERFAMILY, SIONALP, TMHMM, PANTHER AND GENE3D databases.[Fig 3]



InterProScan (version: 4.8)
Sequence: EHJ68866.1
Length: 746
CRC64: 2932BD59F25F3E5E

Launched Fri, Apr 05, 2013 at 09:54:09
Finished Fri, Apr 05, 2013 at 09:54:20

IPR002761 DUF71 domain
PF01902 ▸ ATP_bind_4
TIGR00290 ▸ MJ0570_dom

IPR006175 YjgF/Yer057p/UK114 family
PF01042 ▸ Ribonuc_L-PSP

IPR013813 Endoribonuclease L-PSP/chorismate mutase-like
G3DSA:3.30.1330.40 ▸ G3DSA:3.30.1330.40
SSF55298 ▸ YjgF/chorismate_mutase-like

IPR014729 Rossmann-like alpha/beta/alpha sandwich fold
G3DSA:3.40.50.620 ▸ G3DSA:3.40.50.620

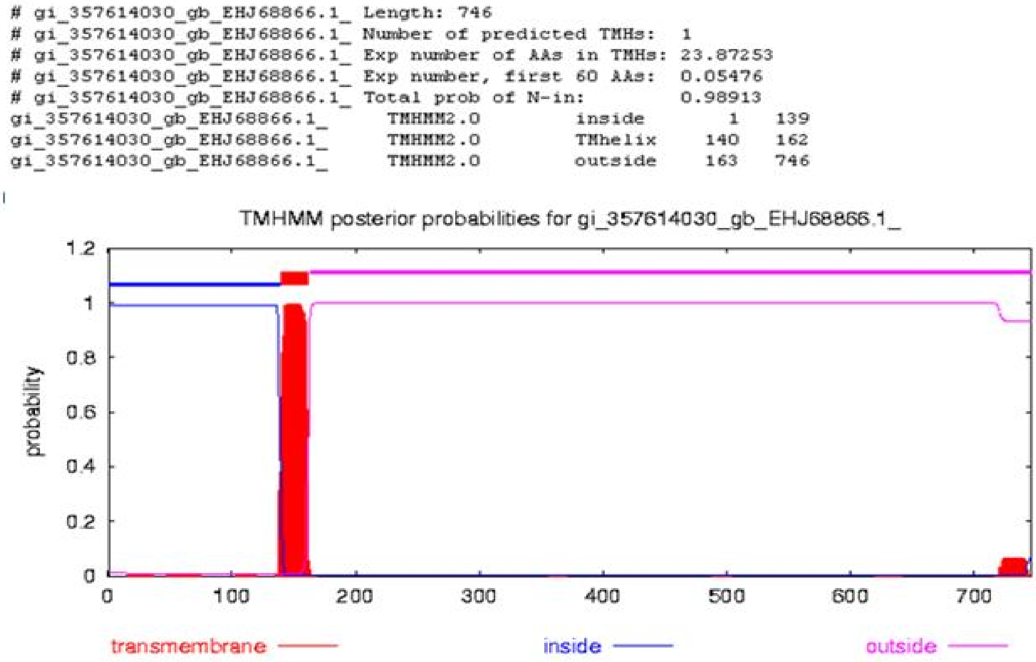
noIPR unintegrated
G3DSA:3.90.1490.10 ▸ G3DSA:3.90.1490.10
PTHR12196 ▸ PTHR12196
SSF52402 ▸ SSF52402

© European Bioinformatics Institute 2006-2012. EBI is an Outstation of the European Molecular Biology Laboratory.

Fig3: Results from InterproScan showing various domains predicted in KGM_05782 protein sequence.

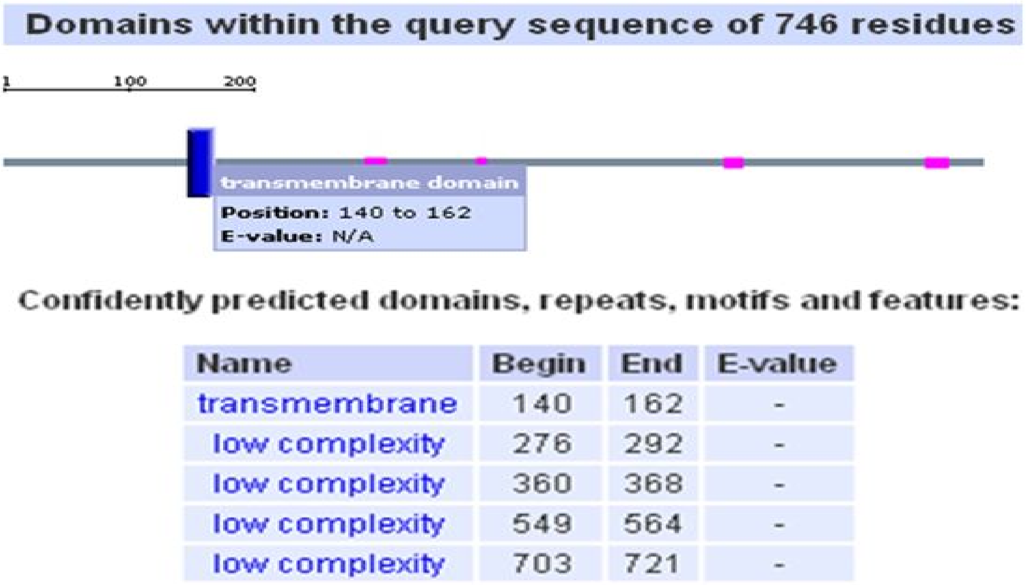
3.3 Prediction of Transmembrane Domain

Transmembrane domain was predicted as a common Domain in hypothetical protein KGM_05782 of Danaus plexippus by using various online tools/software such as SMART, TMHMM, PSORT and Phobius.



```
# gi_357614030_gb_EHJ68866.1_ Length: 746
# gi_357614030_gb_EHJ68866.1_ Number of predicted TMHs:  1
# gi_357614030_gb_EHJ68866.1_ Exp number of AAs in TMHs: 23.87253
# gi_357614030_gb_EHJ68866.1_ Exp number, first 60 AAs:  0.05476
# gi_357614030_gb_EHJ68866.1_ Total prob of N-in:        0.98913
gi_357614030_gb_EHJ68866.1_     TMHMM2.0        inside       1   139
gi_357614030_gb_EHJ68866.1_     TMHMM2.0        TMhelix    140   162
gi_357614030_gb_EHJ68866.1_     TMHMM2.0        outside    163   746
```

Fig4 Transmembrane region predicted by TMHMM



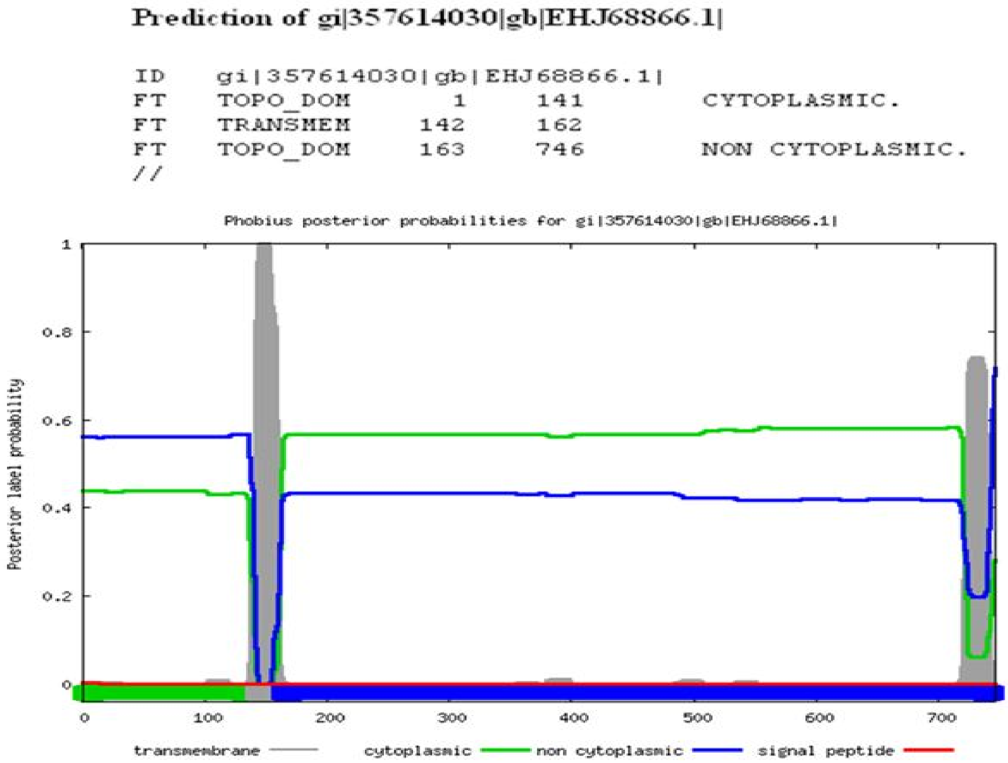
Domains within the query sequence of 746 residues

transmembrane domain
Position: 140 to 162
E-value: N/A

Confidently predicted domains, repeats, motifs and features:

Name	Begin	End	E-value
transmembrane	140	162	-
low complexity	276	292	-
low complexity	360	368	-
low complexity	549	564	-
low complexity	703	721	-

Fig5 Transmembrane region as predicted in SMART



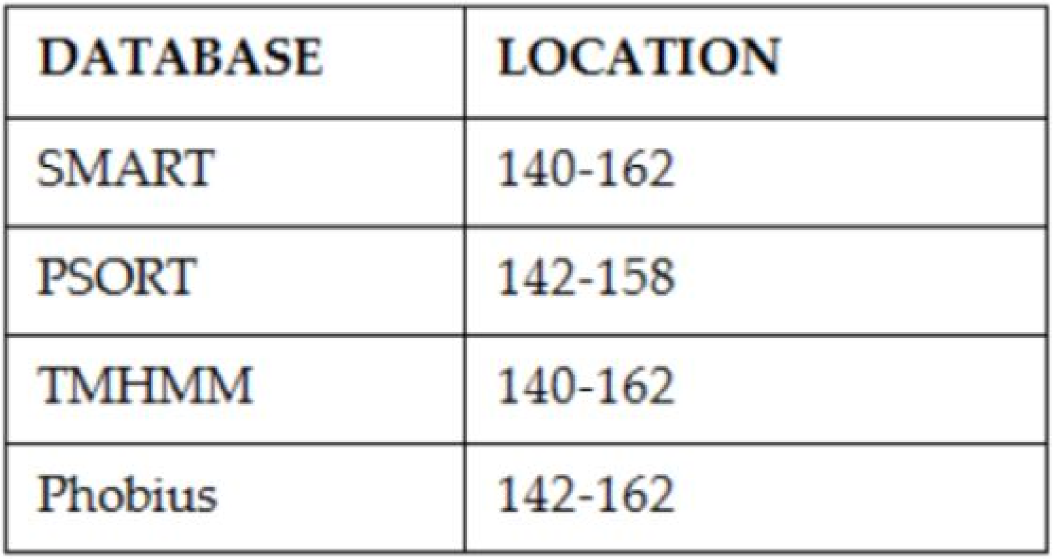
Prediction of gi|357614030|gb|EHJ68866.1|

```
ID    gi|357614030|gb|EHJ68866.1|
FT    TOPO_DOM       1     141        CYTOPLASMIC.
FT    TRANSMEM     142     162
FT    TOPO_DOM     163     746        NON CYTOPLASMIC.
//
```

Fig6 Transmembrane region as predicted in Phobius

Transmembrane region was predicted from 140-162 [Table 1] accounting for a stretch of aproximatly 20 amino acids.

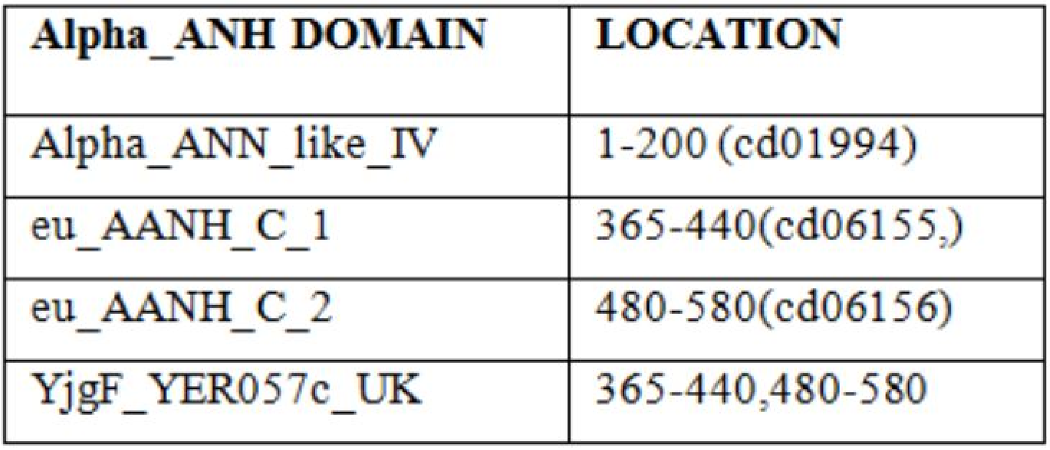
Table 1: Showing location of transmembrane domain.



DATABASE	LOCATION
SMART	140-162
PSORT	142-158
TMHMM	140-162
Phobius	142-162

Alpha ANH domain [38] [39] is also predicted as a functional domain. This is a subfamily of Adenine nucleotide alpha hydrolases superfamily [Table 2]. Adeninosine nucleotide alpha hydrolases superfamily includes N type ATPase and ATP sulphurylases. It forms an alpha/beta/alpha fold which binds to Adenosine group. This subfamily of proteins is predicted to bind ATP. This domain has a strongly conserved motif SGGKD at the N terminus.[40]

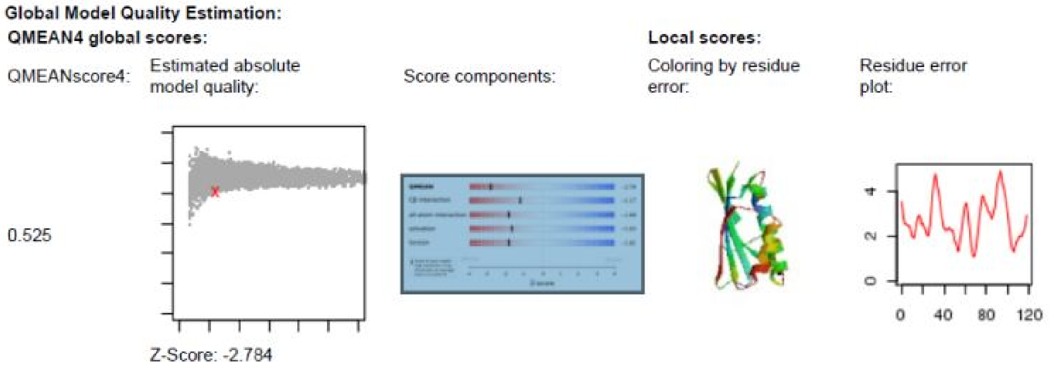
Table 2: Showing location of ATP binding domain



Alpha_ANH DOMAIN	LOCATION
Alpha_ANN_like_IV	1-200 (cd01994)
eu_AANH_C_1	365-440(cd06155,)
eu_AANH_C_2	480-580(cd06156)
YjgF_YER057c_UK	365-440,480-580

3.2 Protein 3D Structure Prediction

We used swiss model and modeller for constructing 3D structure of hypothetical protein KGM_05782 that utilize 3rk1 protein file as templates to model structure. The resultant structure was evaluated using QMEAN Z-score.



Global Model Quality Estimation:

QMEAN4 global scores:

QMEANscore4: 0.525

Estimated absolute model quality:

Z-Score: -2.784

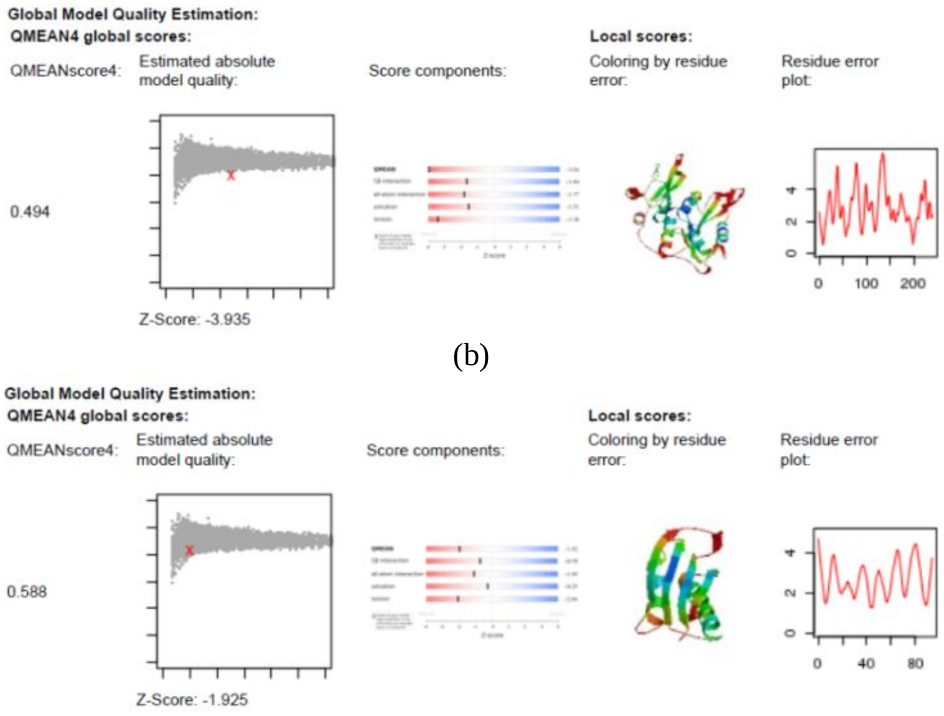
Score components:

Local scores:

Coloring by residue error:

Residue error plot:

(a)



(b)

(c)

Fig7: Results of various parameters for Model-1 (a), Model-2 (b) and Model-3 (c) used for secondary structure predication.

Three models were predicted with QMEAN Z-score as -2.784, -3.935 and -1.925 respectively for Model-1, Model-2 and Model-3. The protein’s secondary structure was visualized using pymol molecular viewer and found to have 2 major helices and five beta sheets.

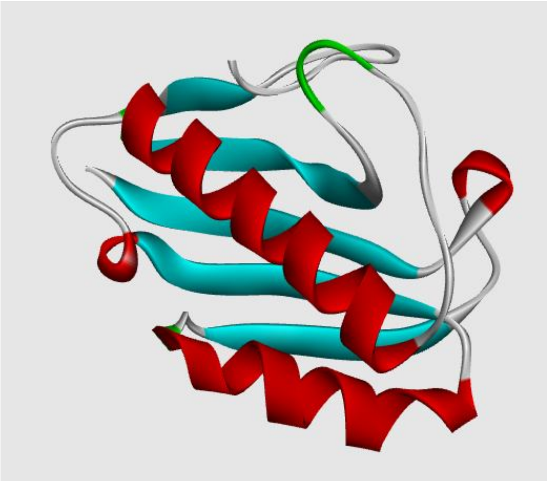


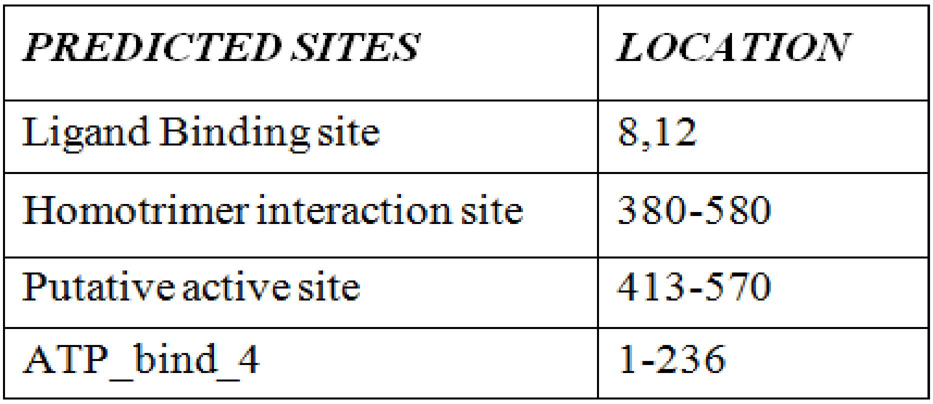
Fig8: Structure visualization of the predicted 3D structure of the hypothetical protein KGM_05782 of Danaus plexippus in Discovery Studio

4 Discussion and Conclusion

Computational sequence analysis and prediction of unknown or uncharacterized proteins is a key for genome annotation. In this study we identified 9000+ hypothetical proteins. Out of this one protein named KGM_05782 was subject to detailed studies and explored using various online/offline softwares and tools freely used for academics. Our considered Protein sequence contains 3.8% proline residues and studies have suggests that proline can be used as an energy substrate for flight muscle during distant migrations. [42][43][44]

Transmembrane domain predicted as common domain in hypothetical protein of Danaus plexippus by using various tools/software's such as SMART, THMM, PSORT and Phobius. Transmembrane region was predicted from 140 -162 that explained it to be a membrane bound protein. It is predicted that it may be a part of some important cascade and work as channelizing protein. The YjgF/YER057c/UK family of proteins is found highly conserved and currently lacks a consensus biochemical function. In a earlier work on Salmonella enteric, strains lacking yjgF has led to a working model in which YjgF functions to remove potentially toxic secondary products of cellular enzymes that can be correlated with its feeding ability on milkyweeds having cardenolide content.[45] This also indicates the highly conserved YjgF/YER057c/UK114 family of proteins responsible for the survival of this weed feeding group. We also predict some other sites and domains which are weakly characterized as Alpha_ANH domain and ligand binding sites. [Table 3]

Table 3 Concluded sites in hypothetical protein



PREDICTED SITES	*LOCATION*
Ligand Binding site	8,12
Homotrimer interaction site	380-580
Putative active site	413-570
ATP_bind_4	1-236

Nevertheless, these predicted data provide a powerful framework for furthur underderstanding of genomes through iterative function assignments and annotations.

References

[1] http://danaus.genomeprojectsolutions-databases.com/

[2] S. Zhan et al., “The monarch butterfly genome yields insights into long-distance migration,” *Cell.* vol.147, no.5, pp. 1171–1185, Nov.2011.

[3] A. Hsiao, and M.D. Kuo, “High-throughput Biology in the Postgenomic Era,” *J Vasc Interv Radiol*, vol.17, pp.1077–1085, 2006.

[4] http://genomics.energy.gov/

[5] http://eol.org/pages/2682739/hierarchy_entries/39081589/overview

[6] A.F Yakunin et al.,”Structural proteomics: a tool for genome annotation”, *Curr Opin Chem Bio.*, vol.8, no.1, pp. 42-8, Feb.2004.

[7] H.A. Watkins and E.N. Baker “Structural and Functional Analysis of Rv3214 from Mycobacterium tuberculosis, a Protein with Conflicting Functional Annotations, Leads to Its Characterization as a Phosphatase,” *J Bacteriol.*, vol.188,no.10, pp. 3589-3599,May 2006.

[8] J. M. Johnston et al.,” Crystal Structure of a Putative Methyltransferase from Mycobacterium tuberculosis: Misannotation of a Genome Clarified by Protein Structural Analysis,” *J Bacteriol.*, vol.185, no.14, pp.4057–4065. July 2003.

[9] http://www.ncbi.nlm.nih.gov/

[10] A. Ikai,” Thermostability and aliphatic index of globular proteins,” *J Biochem.*, vol.88, no.6, pp.1895-1898.Dec.1980.

[11] SC.Gill and PH H.Von,” Calculation of protein extinction coefficients from amino acid sequence data,” *Anal Biochem.*,vol.182,no. 2,pp. 319-26,Nov.1989

[12] K.Guruprasad, BV. Reddy, and MW. Pandit,” 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 Eng.*, vol.4, no.2, pp. 155-61, Dec. 1990.

[13] J.Kyte and R.F.Doolottle,” A simple method for displaying the hydropathic character of a protein,” *J Mol Biol.*, vol. 157,no. 1,pp. 105-132,May 1982.

[14] http://us.expasy.org/tools/protparam.html.

[15] A. Bateman et al.,” The Pfam Protein Families Database,” *Nucleic Acids Res.*, vol.28,no.1,pp. 263–266, Jan. 2000.

[16] http://pfam.sanger.ac.uk/.

[17] http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi/.

[18] A.M.bauer et al.,” CDD: a Conserved Domain Database for the functional annotation of proteins,” *Nucleic Acids Res.*, vol.39, pp.D225-D229, Jan.2011.

[19] E.M. Zdobnov and R. Apweiler,” InterProScan--an integration platform for the signature-recognition methods in InterPro,” *Bioinformatics*, vol. 17, no.9, pp. 847-8, Sept. 2001.

[20] E. Quevillon et al.,” InterProScan: protein domains identifier,” *Nucleic Acids Res.*, vol.33, no. 2, pp. W116-W120, Mar. 2005

[21] http://smart.embl-heidelberg.de/smart/set_mode.cgi?

[22] I.Letunic et al.,”SMART 5: domains in the context of genomes and networks,” *Nucleic Acids Res.* vol.34, pp.D257-D260, Jan.2006.

[23] http://www.cbs.dtu.dk/services/TMHMM/

[24] A. Krogh et al.,” Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes,” *J Mol Biol.*, vol. 305, no.3, pp.567-80, Jan. 2001.

[25] http://psort.hgc.jp/form.html

[26] K. Nakai and P. Horton," PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization," *Trends Biochem Sci.,* vol. 24, no. 1, pp.34-36, Jan.1999.

[27] http://phobius.sbc.su.se/

[28] http://www.ebi.ac.uk/Tools/pfa/phobius/

[29] L. Käll, A. Krogh, and E.L.Sonnhammer," A combined transmembrane topology and signal peptide prediction method*,"* *J Mol Biol.,* vol.338, no. 5, pp. 1027-1036, May2004.

[30] K. Arnold et al.," The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling," *Bioinformatics,* vol. 22,pp. 195-201,2006

[31] N. Guex, and MC. Peitsch," SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling," *Electrophoresis,* vol.18, no. 15, pp.2714-23. Dec. 1997.

[32] T. Schwede et al.," SWISS-MODEL: an automated protein homology-modeling server," *Nucleic Acids Res.,* vol.31, no. 13, pp. 3381-3385,2003.

[33] http://www.ebi.ac.uk/pdbsum/

[34] RA. Laskowski," PDBsum: summaries and analyses of PDB structures," *Nucleic Acids Res.,* vol.29, no. 1, pp. 221-222, Jan. 2001.

[35] http://salilab.org/modeller/

[36] A. Šali and T. L. Blundell," Comparative protein modelling by satisfaction of spatial restraints," *J. Mol. Biol. vol.* 234, pp. 779-815, 1993.

[37] http://spdbv.vital-it.ch/

[38] http://www.pymol.org/

[39] P. Bork and E.V. Koonin," A P-loop-like motif in a widespread ATP pyrophosphatase domain: implications for the evolution of sequence motifs and enzyme activity*,"* *Proteins,*vol. 20, no. 4, pp. 347-355, Dec. 1994.

[40] L.Aravind , V. Anantharaman ,and E.V. Koonin," Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA," *Proteins,* vol. 48, no. 1,pp. 1-14, Jul. 2002

[41] E. Mistiniene et al.," Oligomeric assembly and ligand binding of the members of protein family YER057c/YIL051c/YJGF," *Bioconjug. Chem.,* vol. 14, no. 6, pp. 1243-1252, Nov.2003.

[42] Scaraffia PY, Wells MA., "Proline can be utilized as an energy substrate during flight of Aedes aegypti females. "J Insect Physiol. 2003 Jun;49(6):591-601 PMID: 12804719

[43] Kay BK, Williamson MP, Sudol M., "The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. "FASEB J. 2000 Feb;14(2):231-41. PMID : 10657980

[44] Micheu S, Crailsheim K, Leonhard B. "Importance of proline and other amino acids during honeybee flight--Apis mellifera carnica POLLMANN). "Amino Acids. 2000;18(2):157-75. PMID: 10817408

[45] Jennifer A. Lambrecht,1 Beth Ann Browne, and Diana M. Downs, "Members of the YjgF/YER057c/UK114 Family of Proteins Inhibit Phosphoribosylamine Synthesis in Vitro "J Biol Chem. 2010 November 5; 285(45): 34401–34407. doi: 10.1074/jbc.M110.160515 PMC: 2966054