Analytical Study of Plasmodium Yoelii Yoelii

Akash kumar

Department of Bioinformatics , Uttaranchal College of Science & Technology, Dehradun, India
bioinformaticswelkinakash@gmail.com

Abstract

Protein sequence of Plasmodium yoelii yoelii "prt_seq No definition line found" was extracted from GenPept database (ACCESSION: EAA21357), used to predict its hydrophobic, atomic and amino acid composition, PEST region, solvent accessibility, molecular mass, theoretical pI and finally catalytic site detail was identified with their three dimensional structure. Analysis was made using several bioinformatics tool. PEST region was indentified using pestfind tool, whose abundant availability indicates intracellular half-lives of less than two hour. Hydrophobicity was checked using [1] to know it's buried or exposed state in terms of solvent accessibility, also used to find its different physical and chemical properties, secondary structure composition using Hidden Markov models, formation of disulphide bond, effect of point mutation prediction using heat map representation and finally catalytic site identification were made. Online Active site prediction was made using web server [2], where given protein sequence was matched with server's library of catalytic site, resulting in twenty six identified region with maximum score of 0.015 with Orotidine "5-monophosphate decarboxylase" and minimum score of 0.004 with "Endo-alpha-sialidase". Molecular weight of retrieved protein sequence was found to be 32049.9 and Theoretical Pi value was 6.08, protein close to given Pi and molecular mass was matched using TagIdent web server [3], which reveal 126466 protein. The Instability Index was computed to be 36.19 which classifies the protein as stable and the Aliphatic Index showing relative value occupied by aliphatic side chain (alanine, valine, isoleucine, and leucine) was 80.77. The three dimensional structure of catalytic site was visualized using Jmol software.

Keywords: Catalytic site, Heatmap representation, Propsearch, Solvent accessibility, PEST region.

1. Introduction

Plasmodium yoelii,[P.yoelii], genus *Plasmodium* is a parasite, hosts in insect and vertebrate host for this parasite are mammals [4]. Plasmodium is a genus of *Apicomplexan* parasite cause malaria; more than 200 species has been recognized wherein 11 species infect humans, while others infect for example birds, reptiles, rodents. The chromosomes of these species vary in length from 500 kilo bases to 3.5 mega bases and contain a degenerated chloroplast called apicoplast [5]. P.yoelii yoelii is one of the four malaria species which is transmitted by "Anopheles stephensi" in the laboratory. P.yoelii yoelii have 14 chromosome ranging from size 0.6 MB to 3.8 MB [6]. This species was described in Africa by Landau, Michel and Adam in 1968; mostly used in laboratory to study immune response of malaria infected mice.

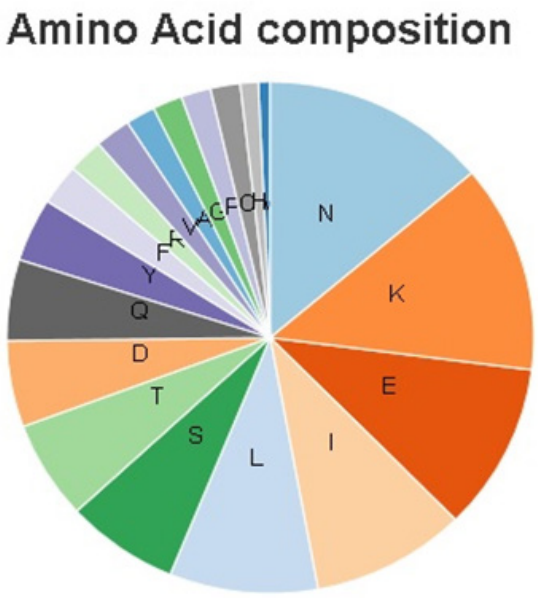
2. Material And Method

The protein sequence was retrieved from GenPept database, whose source accession number was EAA21357.1 available at NCBI website [7]. The retrieved protein sequence was checked for its several physical and chemical properties such as molecular weight, theoretical pI, instability index, aliphatic index and percentage of individual amino acid using Protparam tool from whence chemical structure and formula were deduced. The amino acid sequence was first used to predict 2D structure using Hidden Markov models at [8], which was further analysed and N-terminal amino was checked, its 3D model prediction used 3djigsaw server [9], while three dimensional structure of active site was visualized in Paymol software. Based on data retrieved from 2D structure analysis, protein sequence was checked for its buried or exposed state in terms of hydrophobicity at [1] then protein sequence was used to check its several other properties such as formation of disulphide bond, solvent accessibility, effect of point mutation prediction using heatmap representation and catalytic site identification. Based on hydrophobicity value, theoretical Pi, molecular weight and few other values protein sequence was checked for putative protein family search using Propsearch tool [10]**.** Catlytic site was Identified using [2], where given protein sequence was matched with server's library of catalytic site, which result in twenty six identified region which was then verified with catalytic site atlas at [11].

3. Result and Discussion

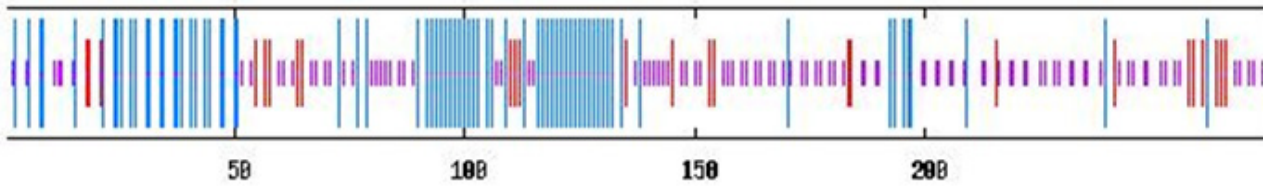
The amino acid sequence of 274 residues of Plasmodium yoelii yoelii was retrieved from genpepet database whose accession number was (ACCESSION: EAA21357) available at NCBI website [13]. The retrieved protein sequence was submitted to know its amino acid composition at [14] figure-1 and secondary structure composition figure-2, figure-3, Figure-4 and figure-5 shows individual and overall hydrophobicity of given protein using Hphob./Kyte&Doolittle method. Table-1 showing propsearch result, where we found that protein was 94% similar with Spindle pole body component SPC42 from organism Saccharomyces cerevisiae (strain ATCC 204508 / S288c) also called Baker's yeast. Figure-7 shows solvent accessibility which demonstrates the exposed and buried part of submitted protein. Figure-8 shows the PEST region detail which check for Proline, glutamic acid, serine and threonine series. Protein rich in PEST region have half-lives of less than two hour and are generally flanked by cluster of positively charged amino acid. Figure-9 shows Heatmap representation of data where individual data is graphically represented. It uses colour-coding to represent the values taken by a variable in a hierarchy. Figure-10 showing first five catalytic site found in protein. First four protein sequence was from same catalytic site EC number (4.01.01.0023) and catalytic site EC label as "Orotidine 5'-monophosphate decarboxylas" with highest score of 0.015 while fifth catalytic site (1get-0cx) was from catalytic site EC number (1.08.01.0007) and catalytic site EC label "Glutathione reductase (e.c.1.6.4.2) wild-type complexed with nadp and fa" with score of 0.011. Orotidine 5'-monophosphate decarboxylas also known as OMP Decarboxylase enzyme which is extraordinarily efficient catalyst and best known for accelerating the unanalyzed reaction rate by a factor of 10^{17} and involved in pyrimidine biosynthesis. Fig-11 and fig-12 shows positional and structural view of catalytic site 1eix-4, while fig-13, fig-14 and fig-15 shows catalytic site 1get-0cx in detail. Atomic composition of given sequence revealed that the total number of atoms present was 4504 where 1395 carbon, 2254 hydrogen, 386 nitrogen, 462 oxygen and 7 sulphur were present, constructing chemical formula as $C_{1395}H_{2254}N_{386}O_{462}S_{7}$.

Fig.1 *Amino Acid composition of given protein sequence*



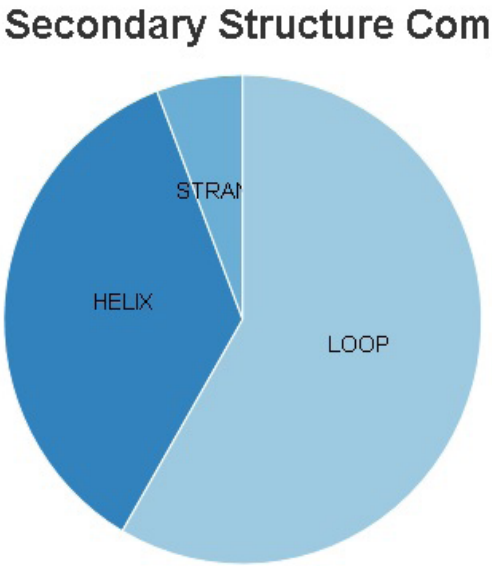
The above diagram shows that protein, contains high concentration of Asparagine (N, 13.9%) and lysine (K, 13.1%). Asparagine might be responsible for alpha-helices structure, nitrogen transportation and capping hydrogen bond interaction, it also provide key site for N-linked glycosylation which play crucial role in cell-cell interaction, cell-extracellular attachment and protein folding; while Lysine may play important role in salt-bridge to make protein more stable. Both amino acids play major role in affectation of central nervous system of vertebrates infected by Plasmodium species.

Fig.2 *Protein secondary structure by HNN*



```
HNN :
    Alpha helix       (Hh) :    90 is  32.85%
    3(10) helix       (Gg) :     0 is   0.00%
    Pi helix          (Ii) :     0 is   0.00%
    Beta bridge       (Bb) :     0 is   0.00%
    Extended strand   (Ee) :    36 is  13.14%
    Beta turn         (Tt) :     0 is   0.00%
    Bend region       (Ss) :     0 is   0.00%
    Random coil       (Cc) :   148 is  54.01%
    Ambigous states   (?)  :     0 is   0.00%
    Other states           :     0 is   0.00%
```

Fig.3 *Diagrammatic representation of Secondary Structure Composition*



Predicting secondary structure by HNN have showed that, more than half of this protein is random coil which in turn mean that monomer subunits are randomly arranged while still being bonded to adjacent unit.

Fig.4 *The individual values of hydrophobicity for the 20 amino acids using the scale Hphob./Kyte&Doolittle.*

```
Ala:  1.800  Arg: -4.500  Asn: -3.500  Asp: -3.500  Cys:  2.500  Gln: -3.500
Glu: -3.500  Gly: -0.400  His: -3.200  Ile:  4.500  Leu:  3.800  Lys: -3.900
Met:  1.900  Phe:  2.800  Pro: -1.600  Ser: -0.800  Thr: -0.700  Trp: -0.900
Tyr: -1.300  Val:  4.200  : -3.500  : -3.500  : -0.490
```

Fig.5 *Prot scale output for hydrophobicity of given protein*

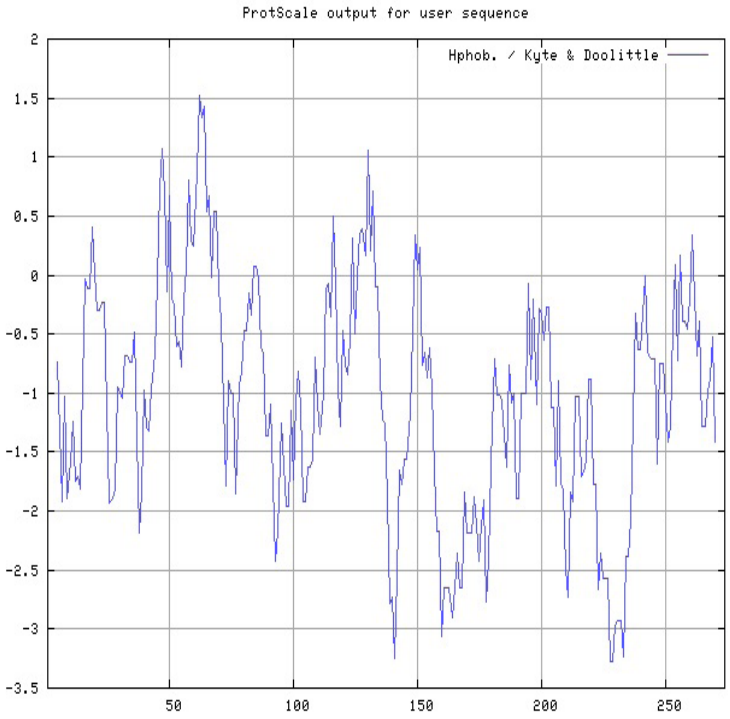


Table 1. *Propsearch result for submitted sequence*

Rank	ID	DIST	LEN2	POS1	POS2	pI	DE
1	sp42_yeast	7.98	363	1	363	8.12	Spindle pole body component SPC42.
2	ste4_schpo	8.56	264	1	264	5.10	Sexual differentiation protein ste4.
3	ybm6_schpo	8.65	327	1	327	8.85	Hypothetical 38.1 kDa protein C582.06c in chromosome II.
4	yp9a_caeel	9.27	261	1	261	7.63	Hypothetical 30.4 kDa protein C28H8.10 in chromosome III.
5	ybb0_yeast	9.29	280	1	280	4.85	Hypothetical 32.6 kDa protein in SCT1-HIR1 intergenic region

In above table first column shows rank , the second column gives the SWISSPROT or PIR id, then the distance score followed by the length of the overlap between query and the subject, the positions of overlap, the calculated pI and the definition line for the found sequence. A distance score of below 8.7 indicates a 94% chance of similarity between the two proteins.

Fig.7 *Diagramatic representation of Solvent Accessibility*

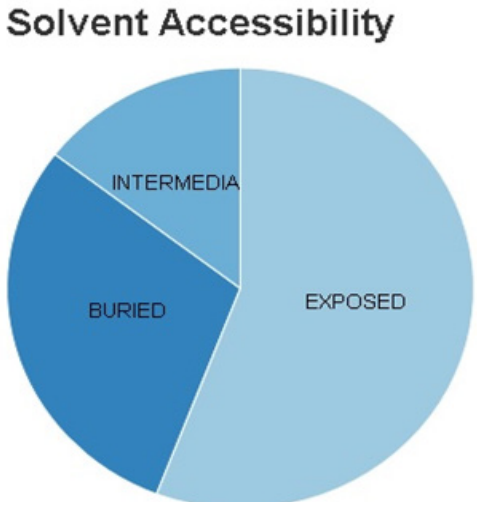


Fig.8 *PEST region detail*

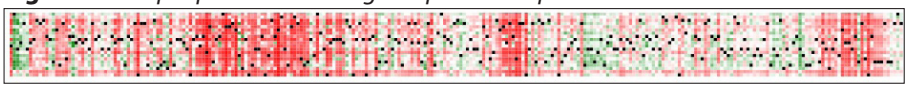
```
PEST-find: Finds PEST motifs as potential proteolytic cleavage sites.

    1 PEST motif was identified in  from 1 to 274.

Poor PEST motif with 15 amino acids between position 232 and 248.
  232 KETNQDVNTLINQTTPK 248
      PEST score: -2.60
```

Sequence was checked for PEST region, where one PEST motif was identified between 232 and 248 position which was of 15 amino acid and show very poor score of -2.60.

Fig.9 *Heat map representation of given protein sequence*



In heat map representation larger values were represented by small dark gray or black squares (pixels) and smaller values by lighter squares.

Fig.10 *First five Catalytic site in given protein*

Catalytic site	Protein	PDB EC (if known)	Score	Catalytic site EC number	Catalytic site EC label
1eix-4	no_def		0.015	4.01.01.0023	Orotidine 5'-monophosphate decarboxylase
1eix-7	no_def		0.015	4.01.01.0023	Orotidine 5'-monophosphate decarboxylase
1eix-6	no_def		0.014	4.01.01.0023	Orotidine 5'-monophosphate decarboxylase
1eix-5	no_def		0.013	4.01.01.0023	Orotidine 5'-monophosphate decarboxylase
1get-0cx	no_def		0.011	1.08.01.0007	Glutathione reductase (e.c.1.6.4.2) wild-type complexed with nadp and fa
1dub-0	no_def		0.006	4.02.01.0017	2-enoyl-coa hydratase

Above figure showing first five highest scoring catalytic site detail, which can be used for further analysis.

Fig.11 *Position of catalytic site 1eix-4*

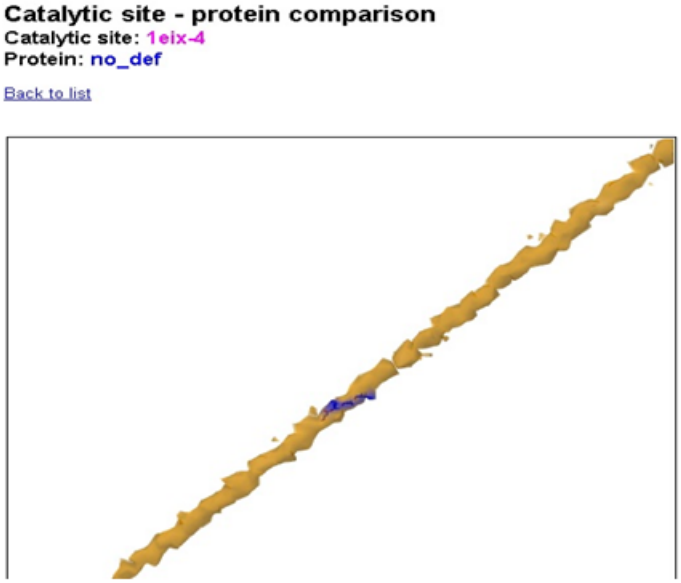


Fig.12 *Showing catalytic site 1eix-4*

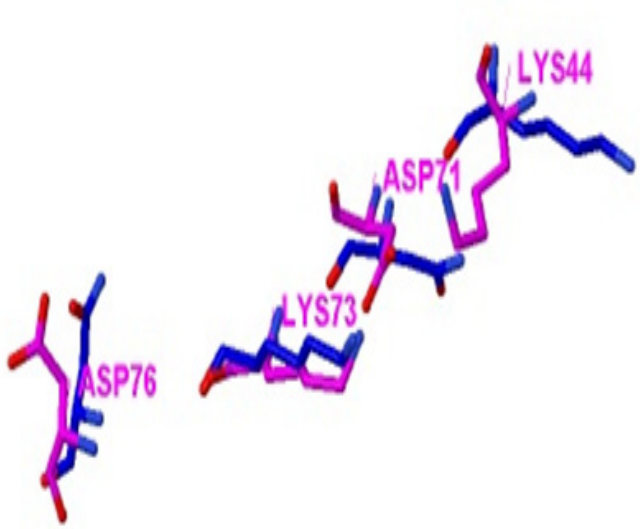


Fig.13 *Position of catalytic site 1get-0cx*

Fig.14 *Showing catalytic site 1get-0cx*

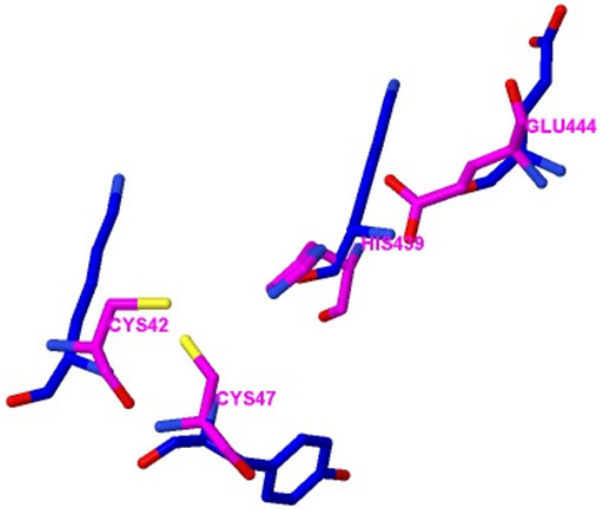
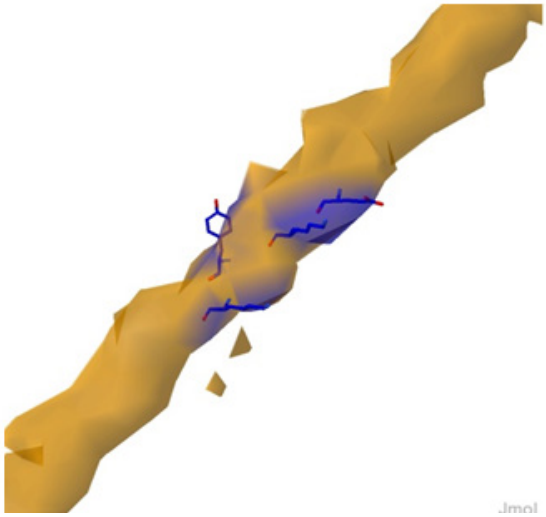


Fig.15 *Close view of catalytic site 1get-0cx*



4. Conclusion

Searching of protein sequence from large database, analysing its various physical and chemical properties and identifying active/catalytic site detail with three dimensional structure on the basis of their amino acid sequence is very interesting and challenging task for bioinformatists. Hydrophobicity was checked after confirming atomic and amino acid composition details: revealed that it contain 1395 atom of carbon, 2254 atom of hydrogen, 462 atom of oxygen, 386 atom of nitrogen and sulphur in very small amount with 7 atom only- whereas amino acid composition detail showed us that it contain Asparagine [highest 13.9%]. The N-terminal of sequence considered was Methionine. Total number of negatively charged residues was 44 and positive residues were 42; while total number of atoms calculated was 4504. Solvent accessibility showed that exposed part is more than buried part in given protein sequence. Molecular mass of retrieved protein sequence was 32049.9 kdal and theoretical pI value at which this protein sequence carries no net electrical charge was 6.08. A search was made based on its molecular mass and theoretical pI value using Propsearch tool wherein was found sp42_yeast protein with 94% similarity. Finally protein sequence was submitted to find its catalytic site based on above collected data and found 1eix_4 site with highest score 0.014 which further detail were visualized using Jmol software. Protein sequence was also submitted to be checked for relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine) and found its Aliphatic Index value as 80.77. The Instability Index (II) is computed to be 36.19 which classify the protein as stable. Hope that the work done above will be useful to know more about this Malaria spreading species of genus Plasmodium and will be helpful for new research worldwide on different aspects.

5. References

1. SIB Swiss Institute of Bioinformatics: http://web.expasy.org/cgi-bin/protscale/protscale.pl?1.
2. Integrated into UniProtKB/TrEMBL (Dec 2003) http://www.ebi.ac.uk/ena/data/view/EAA21357.
3. SIB Swiss Institute of Bioinformatics, TagIdent: http://web.expasy.org/cgi-bin/tagident/tagident0.pl.
4. Combet C., Blanchet C., Geourjon C. and Deleage G (2000) NPS@:Network Protein Sequence Analysis TIBS Vol. 25, No 3 [291]:147-150.
5. Bates, PA. Kelley, LA., MacCallum, RM. and Sternberg, MJE. (2001) Proteins: Structure, Function and Genetics, Suppl, 5, 39-46.
6. Gasteiger, E. Hoogland, C. Gattiker, A. Duvaud, S. Wilkins, MR. Appel, RD. Bairoch, A. (2005) The Proteomics Protocols Handbook, Humana Press, 571-607.
7. National Center for Biotechnology Information (NCBI), Science and Health, Biomedical and Genomic Information: http://www.ncbi.nlm.nih.gov/.
8. Kumar, Akash, Mohd Zakir Khawaja, and Vivek Dhar Dwivedi (2013) "Evolutinary Analysis and Motif Discovery in Pinopsin from Vertebrates", *Indian Journal of Bioinformatics and Biotechnology,* 2.3: 65-68.
9. Offman MN, Fitzjohn PW, Bates PA (2006) Developing a move-set for protein model refinement, Bioinformatics; *22(15):1838-1845.*
10. Prof.Dr. Heinz-Uwe Hobohm, University of Applied Sciences, KMUB-Bioinformatics: http://abcis.cbs.cnrs.fr/htbin-post/PROPSEARCH/propsearch_cgi.pl.
11. G J Bartlett, C T Porter, N Borkakoti & J M Thornton (2002) "Analysis of catalytic residues in enzyme active sites", Journal of Molecular Biology, Vol 324 p 105-121: http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/CSA/CSA_Browse.pl.
12. DB Source: accession AABL01000533.1(Oct-2002) http://www.ncbi.nlm.nih.gov/protein.
13. https://www.predictprotein.org/.
14. Kumar, Akash, Mohd Zakir Khawaja, Ahsan ul haq Qureshi (*2013)* "Predicted structure of an unknown protein sequence from US patent 8211428" *Online Journal of Bioinformatics Volume 14 (2): 118-125.*
15. Nilmeier JP, Kirshner DA, Wong SE, Lightstone FC (2013) Rapid catalytic template searching as an enzyme function prediction procedure, *PLoS ONE* **8**(5): e62535, doi:10.1371/journal.pone.0062535.
16. Kirshner DA, Nilmeier JP, Lightstone FC (2013) Catalytic site identification – a web server to identify catalytic site structural matches throughout PDB, *Nucleic Acids Research*, doi:10.1093/nar/gkt403.
17. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2005) *Protein Identification and Analysis Tools on the ExPASy Server;* (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press, pp. 571-607.
18. Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, Nucl. Acids. Res. 32: D129-D133.
19. James W. Torrance, Gail J. Bartlett, Craig T. Porter, Janet M. Thornton (2005) Using a Library of Structural Templates to Recognise Catalytic Sites and Explore their Evolution in Homologous Families, J.Mol.Biol. 347:565-8.