

***In Silico* Analysis of Primary Sequence and Tertiary Structure of *Lepidium Draba* Peroxidase**

Y. Fattahian^a, A. Riahi-Madvar^{a,*}, R. Mirzaee^b, M. Torkzadeh-Mahani^a and G. Asadikaram^c

^aDepartment of Biotechnology, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran

^bJiroft University of Medical Sciences, Jiroft, Iran

^cEndocrinology and Metabolism Research Center, Institute of Basic and Clinical Physiology Sciences, and Department of Biochemistry, School of Medicine, Kerman University of Medical Sciences, Kerman, Iran

(Received 2 May 2017, Accepted 16 October 2017)

ABSTRACT

Peroxidase enzymes are vastly applicable in industry and diagnosis. Recently, we introduced a new kind of peroxidase gene from *Lepidium draba* (LDP). According to protein multiple sequence alignment results, LDP had 93% similarity and 88.96% identity with horseradish peroxidase C1A (HRP C1A). In the current study we employed *in silico* tools to determine, to which group of peroxidase enzymes LDP belongs. The tertiary structure of protein was built using homology modeling method and the overall quality of predicted model was evaluated. The physicochemical and spatial properties of protein were measured. The binding site of protein which made a motif, was determined. The predicted tertiary structure possesses acceptable quality, based on several criteria. The predicted motif was also in agreement with previous studies. The isoelectric point (pI) of LDP was calculated as 8.25. Due to neutral pI of LDP and high similarity between structure of LDP and HRP C1A, it is assumed that LDP belongs to the neutral or neutral-basic isoenzymes.

Keywords: *Lepidium draba*, Peroxidase, *In silico*, Bioinformatics, Binding site

INTRODUCTION

Oxidative reactions defined by transferring the electron to peroxide species, the reaction catalyzes by peroxidase enzymes which fall into EC 1.11.1.7 group of enzyme class. This group of enzymes attracts attentions in past few years due to introducing new applications in diagnosis and therapeutics [1]. Peroxidases divided into three superfamilies *viz.* plant peroxidases, animal peroxidases, and catalases. The Plant peroxidase superfamily contains peroxidases of prokaryotic and eukaryotic origin, and based on evolutionary origin and structural similarities the members classified into three classes: class I: peroxidases of prokaryotic origin, class II: secreted-fungal peroxidases, and finally class III: classical secretory plant peroxidases [2]. Horseradish peroxidase (HRP) is the most well-known member of the class III [3]. Peroxidases have vast

applications in industry, environmental, and health sectors as well as pharmaceutical and diagnosis [4]. Ongoing studies until now with new descriptions and somehow novel insights into the enzyme applications (see the review article in reference [1]) make it well worth a subject of studying and further justified the search for finding the new sources and novel applications.

L. draba is an invasive perennial weed, a member of family Brassicaceae [5]. Although this parasite induces economic burden to agriculture industries, it also can be a source of peroxidase enzyme. Previously we have cloned a partial of peroxidase gene from *L. draba* (the sequence has been deposited in the GenBank with accession No. AIJ01351). We named it hereon as LDP (for *L. draba* Peroxidase).

In the present study, we employed *in silico* tools to analyze the properties of this sequence as well as structural properties. Results showed a great similarity between LDP and HRP both in primary sequences and structures. The

*Corresponding author. E-mail: riahi.ali@gmail.com

active sites were also analyzed and compared to other peroxidases. We herein provide theoretical evidences to confirm that the sequence belongs to the neutral or neutral-basic isoenzymes of peroxidase family. The enzyme may be applicable in industry and diagnosis as well as therapeutics.

METHODS

Sequence Retrieval

Sequence of LDP was retrieved from National Center for Biotechnology Information (NCBI) at <https://www.ncbi.nlm.nih.gov> (with accession NO. AIIJ01351).

Sequence Collection and Primary Analysis

Close homologues to LDP were searched by querying AIIJ01351 against non-redundant protein database using NCBI protein BLAST [6]. Another search strategy was based on HMMER algorithm [7] at <http://www.ebi.ac.uk/Tools/hmmer/usingBLOSUM45> similarity matrix. All sequences were imported into CLC-sequence viewer 7.7.1 [8] for further analysis. Also Cello2GO server [9] at www.cello.life.nctu.edu.tw was used to collect GO terms related to the LDP.

Alignments

All alignments were generated at www.tcoffee.org by Espresso-T-Coffee [10] for structural alignments and M-Coffee [11] for conservation analysis. Alignments were analyzed and visualized using Alvis [12] standalone software and Circoletto [13].

Secondary and Tertiary Structures

The secondary structure prediction was based on RaptorX algorithm [14]. The algorithms employed deep convolutional neural fields for secondary structure prediction; and achieved the accuracy of more than 70 percent.

Tertiary Structure Building and Quality Evaluation

A BLAST search against Protein Data Bank (PDB) was performed for finding homologue structures using LDP sequence as a query. BLAST hits with zero e-value were served as templates for building structure in Modeller [15].

The approach was initiated by aligning the templates with each other, then aligning templates with LDP sequence (as a query) and finally models were generated based on homology modeling. All models were refined using GalaxyRefine algorithm [16,17] at www.galaxy.seoklab.org. Quality of initial and refined models was evaluated at Protsav meta server [18] which employed 10 external programs to assess the overall quality of protein 3D models. Ramchandran Plot was assessed at <http://mordred.bioc.cam.ac.uk/~rapper/rampage.php> using Rampage software [19], and more professional structural analyses were done using MolProbity at www.prosbity.biochem.duke.edu.

Structural Alignments

The structural alignment of the predicted sequence with reference structure (PDB ID: 4atj) was based on TM-score [20] as provided by Zhang lab at <http://zhanglab.ccmb.med.umich.edu/TM-align/>. TM-Score scales the structure similarity in which TM-score has the value in (0, 1), where 1 indicates a perfect match between two structures. Based on statistics of structures in the PDB, scores below 0.2 corresponds to randomly chosen unrelated proteins whereas with a score higher than 0.5 assume generally the same fold in SCOP/CATH [20]. (Statistical based information about TM-Score: $0.0 < \text{TM-score} < 0.3$, random structural similarity; $0.5 < \text{TM-score} < 1.00$, in about the same fold).

Active Site Prediction

This stage of research was completely based on predicted tertiary structure. Binding sites and ligands were predicted at <http://www.sbg.bio.imperial.ac.uk/>[21]. MetaPocket [22] and SiteHound-web [23] were employed for predicting the interactive residues; the critical residues, dose predicted as binding sites, were extracted from whole structure using Fit3D server [24]. These residues build a motif which then subjected to a combinatorial (basically combination of geometry and residual composition) search approach against 3d structures of PDB.

RESULTS

Sequences from NCBI Database

A 308 amino acid long sequence of AIIJ1351 was downloaded and served as query for searching approaches.

Table 1. GO Enriched Terms. GO ID and Relative Activities are Extracted from Database and Online Search Tools such as Cello2GO

	<u>GO:0004601</u>	Peroxidase activity
Molecular function	<u>GO:0020037</u>	Heme binding
	<u>GO:0046872</u>	Metal ion binding
Biological process	<u>GO:0042744</u>	Hydrogen peroxide catabolic process
	<u>GO:0005576</u>	Extracellular region
Cellular component	<u>GO:0005618</u>	Cell wall
	<u>GO:0005773</u>	Vacuole
	<u>GO:0016020</u>	Membrane

BLAST hits with zero e-value, at least 95 percent identity, and 95 percent sequence coverage were selected for building the alignments. BLAST results revealed the existence of close homologues for LDP in non-redundant protein database and also close structures. HMMER search tool with a shallow matrix provided access to more distantly related sequences. The sequence belongs to peroxidase family defined by Pfam accession number: peroxidase (PF00141).

By searching through database with a shallow substitution matrix we found homologous sequences in more distant organisms, domains of prokaryotes, the main reason was to find strictly conserved regions and comparing the physicochemical properties of sequences. Closest bacterial peroxidase was an adenylate/guanylate cyclase (h2ckb8) from *Leptonema illini* (species of bacteria spirochetes and the only species of the genus *Leptonema*) [25]. The homologue regions restricted to the C-terminal moiety of sequence. The other distantly related sequence

was A0A0C3RJ46, a short peroxidase from *Nitrosospira* sp. NpAV. GO enriched terms for molecular function, biological process, and cellular components (Table 1) addressed a metalloperoxidase.

Key description of amino acids from HRP C1A (*Armoracia rusticana*) can be found in UniprotKB under accession number P00433, these featured residues are: I36: transition state stabilizer, H40: active site proton acceptor, and finally metal binding sites are F41, C44, V46, G48, D50, Q62, N137, G168, G169, D220, L223, V228. Also Pfam detected active site is H42 similar to Q9LHA7, peroxidase 31 *Arabidopsis thaliana*.

Physicochemical Properties and Amino Acid Composition

The leucine and serine amino acids are the dominant residues in all sequences. The isoelectric point of peroxidases ranges from 4.59 to 9.7 (Fig. 1).

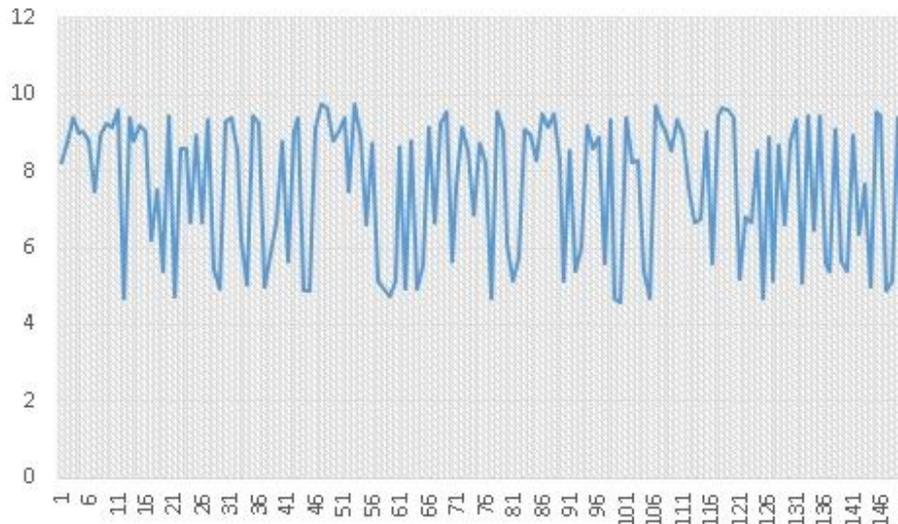


Fig. 1. Isoelectric point of 150 uniref90 peroxidase related sequences. The Y axis is pI and X represents each sequence.

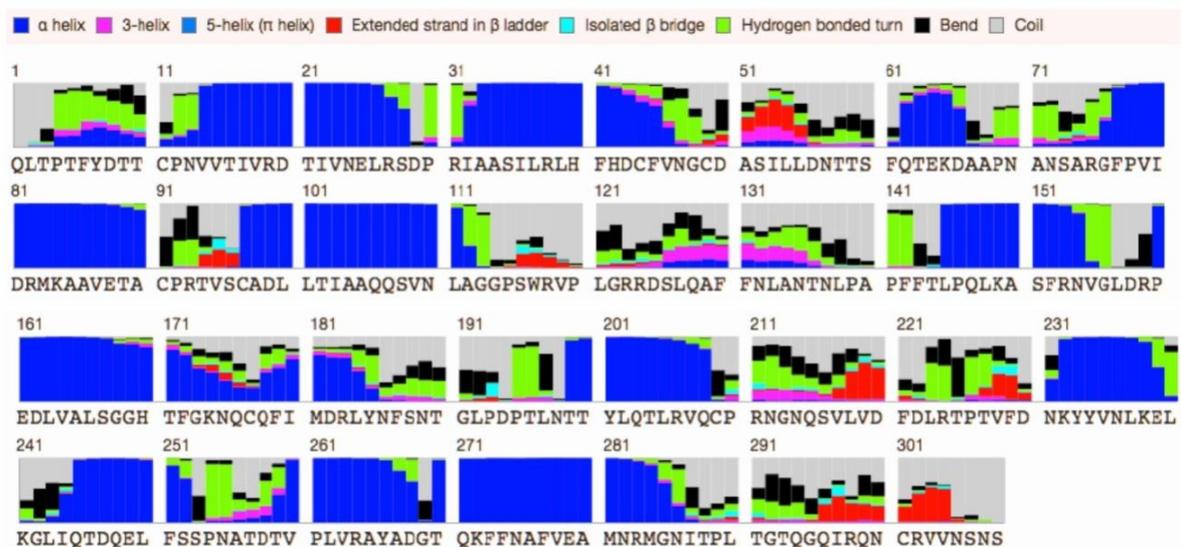


Fig. 2. Graphical representation of secondary structure of LDP. The guide bar on the top of image defined the structure by colors. The height of each column is proportional to the probability of each class related to corresponding residue.

Secondary and Tertiary Structure

Secondary structure components of LDP were mainly coil (52%) followed by helix (46%), and as little as 1.95 percent beta sheet (Fig. 2). These structures can also be

visually observed in predicted structure. BLAST search against PDB revealed the existence of several homologous structures to LDP (Fig. 3).

Modeller outputs were refined and optimized. In each

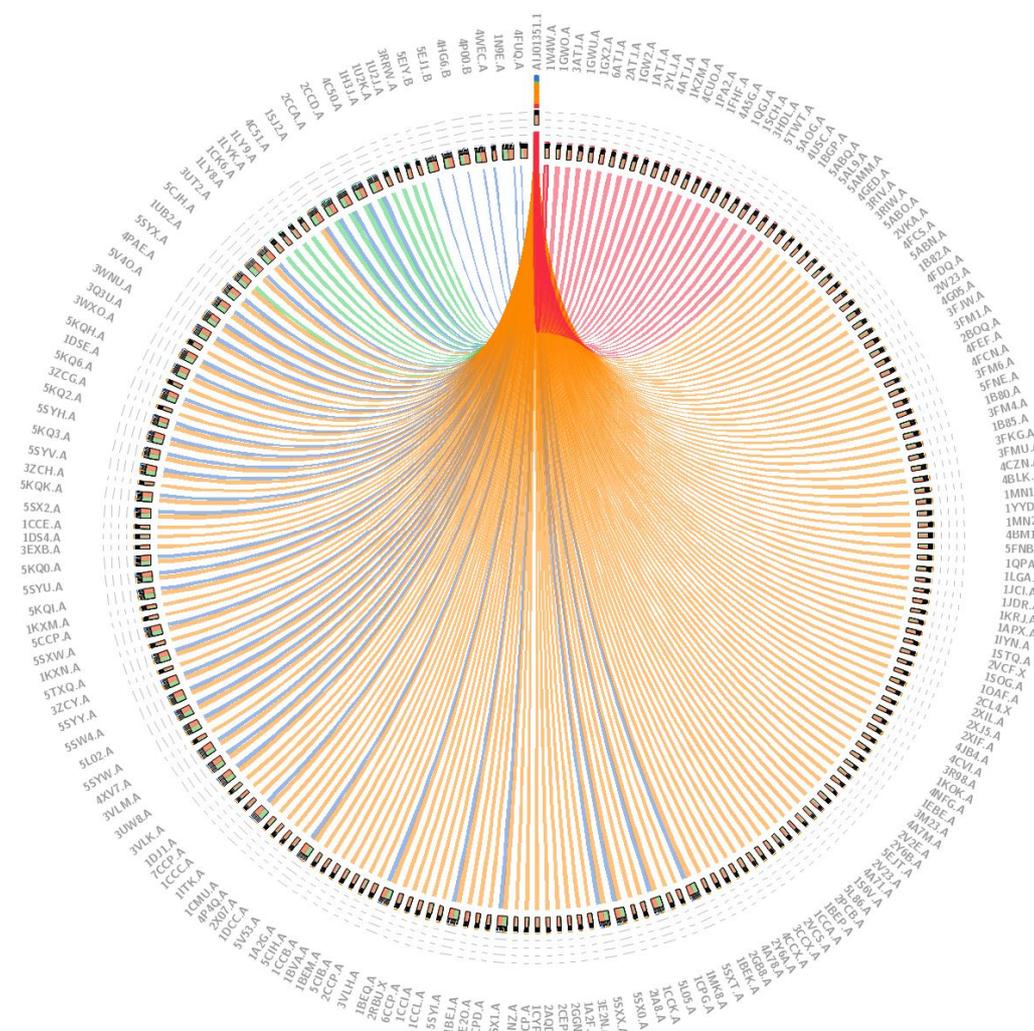


Fig. 3. Graphical visualization of BLAST results. BLAST was done against PDB and queried the sequence of LDP. Ribbons are colored based on identity percentage. Red ribbons represent the fourth quartile up to the maximum score (80% identity); while green ribbons are representative of the second quartile.

step the overall quality of predicted structures was evaluated (Fig. 4). The best constructed model that showed the best overall quality and a good agreement with Molprobit (Table 2) was selected for further analyses. Also Ramachandran plot shows the most proportion of residues in favored region (Fig. 5).

The sequence-order independent protein structure comparison was done with optimized residue to residue

alignment. The TM-score value was 0.99282 that shows a perfect match, Asn175, Phe179, Phe221, Ileu244, Thr246, between two structures (predicted model and 4atj); RMSD evaluated as 0.41 angstrom.

The binding site comprised of Leu37, Arg38, Phe41, His 42, Pro69, Ser73, Pro139, Ala140, Pro141, Phe152, Leu163, Leu166, Ser167, Gly169, His170Phe277 and Met281. These residues are highlighted in yellow in Fig. 6. Residues of this cluster were predicted by several

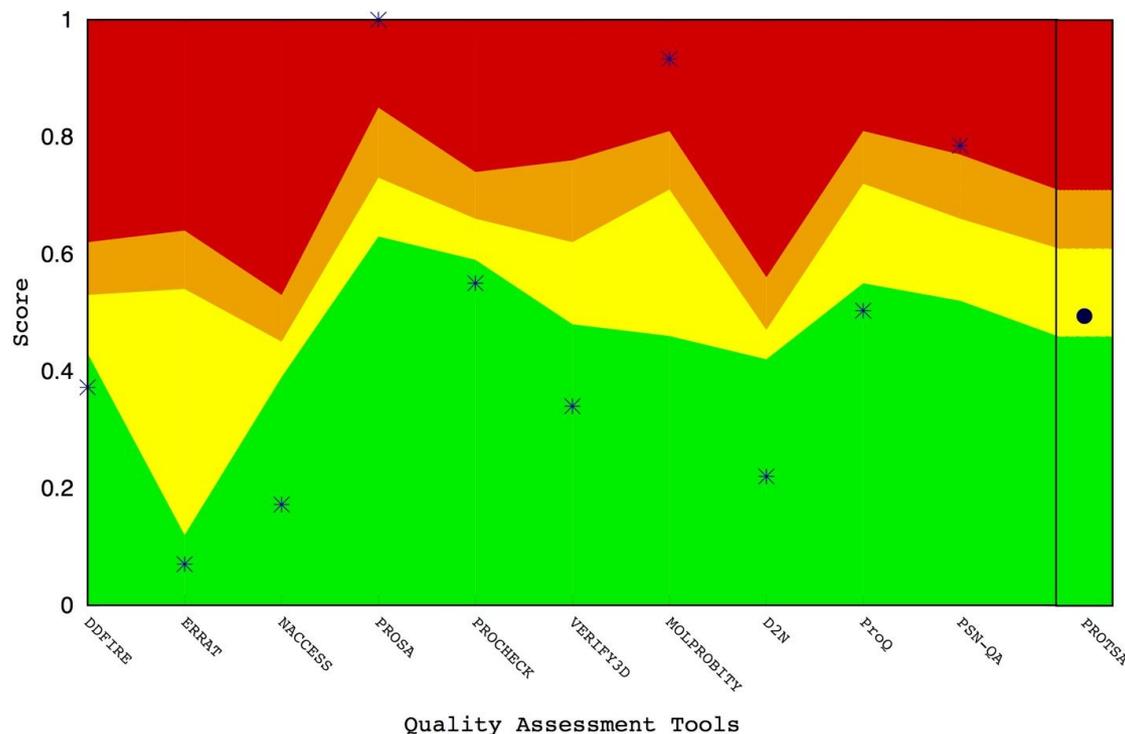


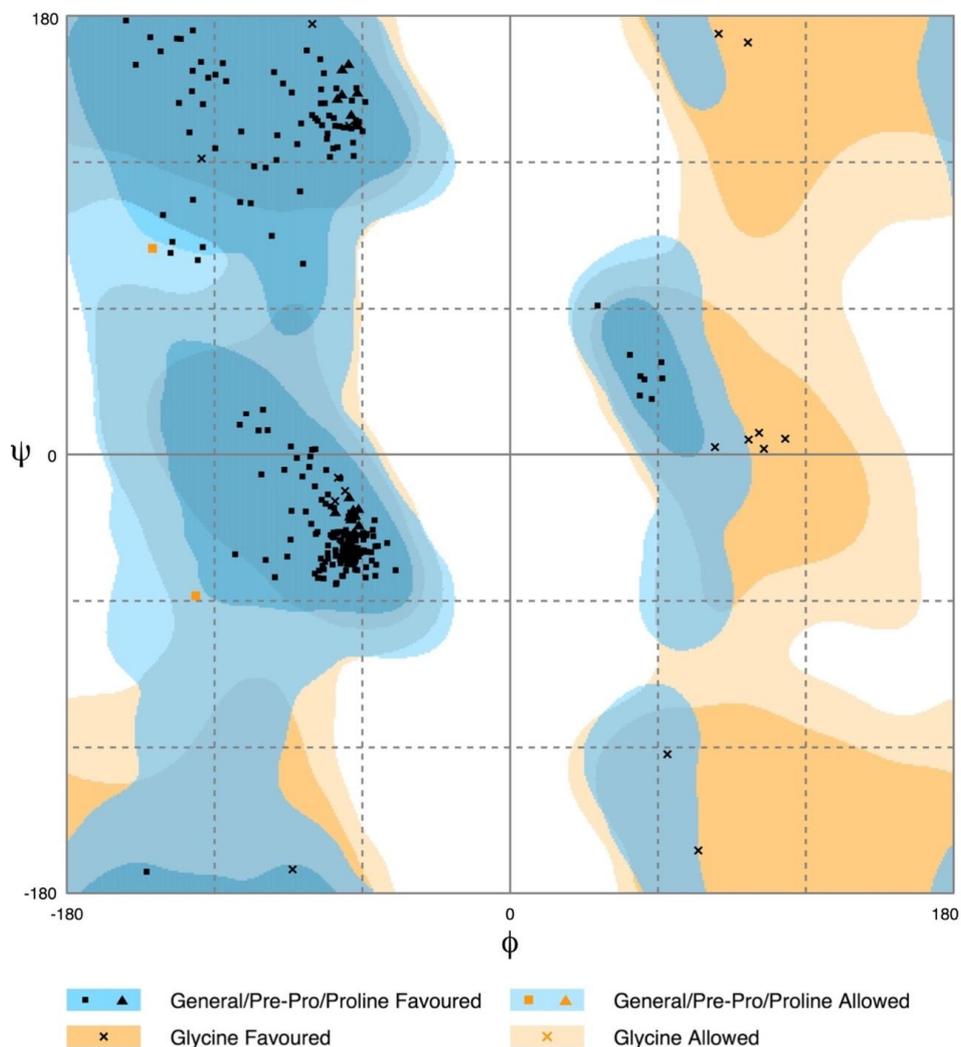
Fig. 4. Color plot of overall quality of the best selected model. Each asterisk is individual Module Quality Score of Input Structure corresponding to the assessment tool mentioned in X axis. The black ball showed Overall Quality Score of Input Structure. Green color indicates the input structure to be in 0-2 Å RMSD, yellow color indicates 2-5 Å, orange color indicates 5-8 Å and red color indicates input structure to be beyond 8 Å.

Table 2. Summary Statistics for Quality Assessment of LDP 3D Structure. Results Came from Geometric Analysis of 3D Structure by Molprobity

Characteristic	Raw count	Percentage	Favorite goal
Poor rotamers	3	1.12%	<0.3%
Favored rotamers	262	97.76%	>98%
Ramachandran outliers	0	0.00%	<0.05%
Ramachandran favored	304	99.35%	>98%
C β deviations > 0.25 Å	7	2.40%	0%
Bad bonds	2/2444	0.08%	0%
Bad angles	27/3333	0.81%	<0.1%

algorithms. The strength of binding site is measured as sum of the interaction energy 1389.37. Lowest and highest interaction energy for all points contributing

to the cluster is between -19.27 and -8.93. Size of cluster in cubic angstrom is 115. These binding sites were extracted as a motif (Figs. 7A and B). All tertiary structures



Number of residues in favoured region (~98.0% expected)	: 304 (99.3%)
Number of residues in allowed region (~2.0% expected)	: 2 (0.7%)
Number of residues in outlier region	: 0 (0.0%)

Fig. 5. Ramachandran Plot of the best constructed model for LDP.

were visualized by Chimera stand-alone software.

The region comprise the 38-69 residues predicted as a portion of binding motif which are highly conserved among all homologous members of our library. The position 52 in the alignment is the most variable part but indeed, occupied by polar residues (Fig. 8).

DISCUSSIONS

With a huge repository of sequence data, structures, and annotated functions, it is possible to inferring functions as well as identification of novel sequences from primary amino acid sequence [26]. After collecting a library of

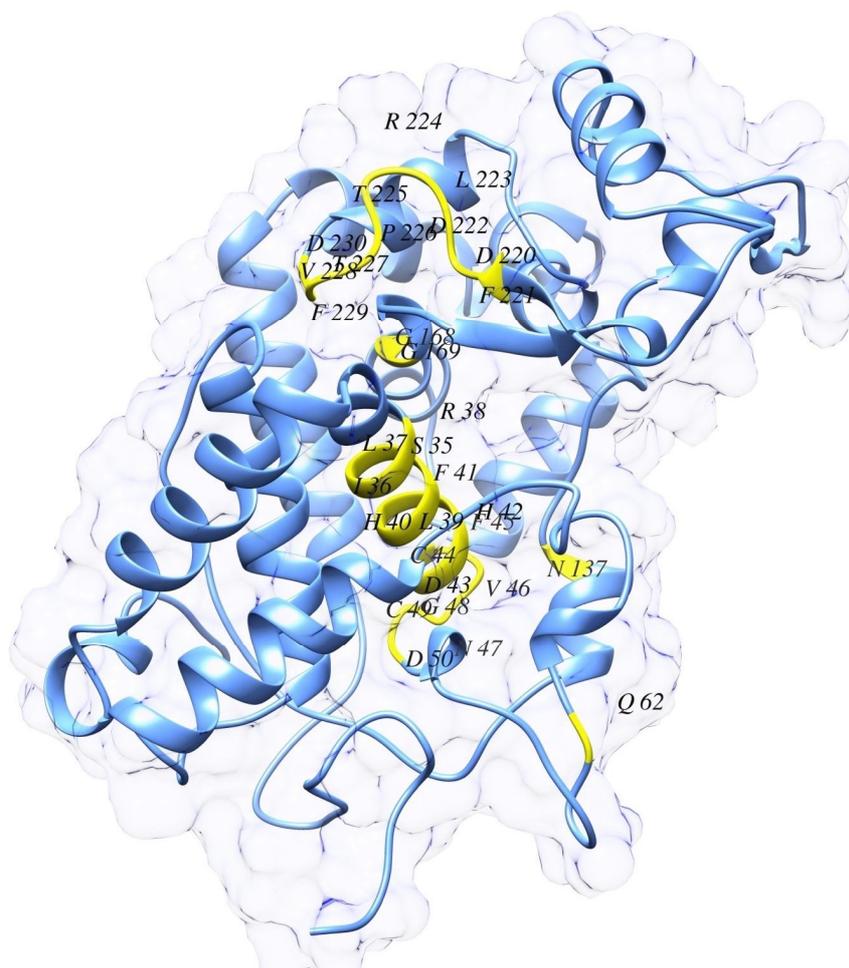


Fig. 6. Tertiary structure of LDP, Coils and helices as the main components of tertiary structure are evident in the image. Yellow colored residues against cyan cartoon backbone are predicted binding sites.

homologous sequences it was observed that not only was critical annotated residues identical but also placed in the same position in the primary sequence, suggesting the high conservancy level of the enzymes and also the critical structure, necessary for peroxidase activity.

In the spite of high conservancy in critical residues, length, and even structure of peroxidases, there is a vast difference between isoelectric points. Isoelectric point of LDP was calculated as 8.25 which is like most collected peroxidases. In our recent work, pI of LDP was calculated as 7.7, using computing pI/Mw tool at ExPASy Server (unpublished data). However, it seems that pI: 8.25 is more reasonable, because of some characteristics of LDP which

are similar to neutral-basic isoenzymes. The lowest pI referred to peroxidase A2-like Acc. No. UPI000510AEF7 from *Pyrus bretschneideri* (Chinese white pear) (4.59 with 58 percent identity to LDP) and the highest pI was theoretically calculated for a peroxidase from *Cynara cardunculus*, var. *scolymus* (Globe artichoke) (*Cynara scolymus*) with Acc. No. A0A118K1I0 (9.75 with 52 percent identity to LDP). There is not any established correlation between enzymatic function and pI [27] similar fact is likely true for LDP, it means that the enzymatic activity may not affected by pI. Residues of higher frequencies were nearly identical in our library members for example the most abundant residues in most sequences

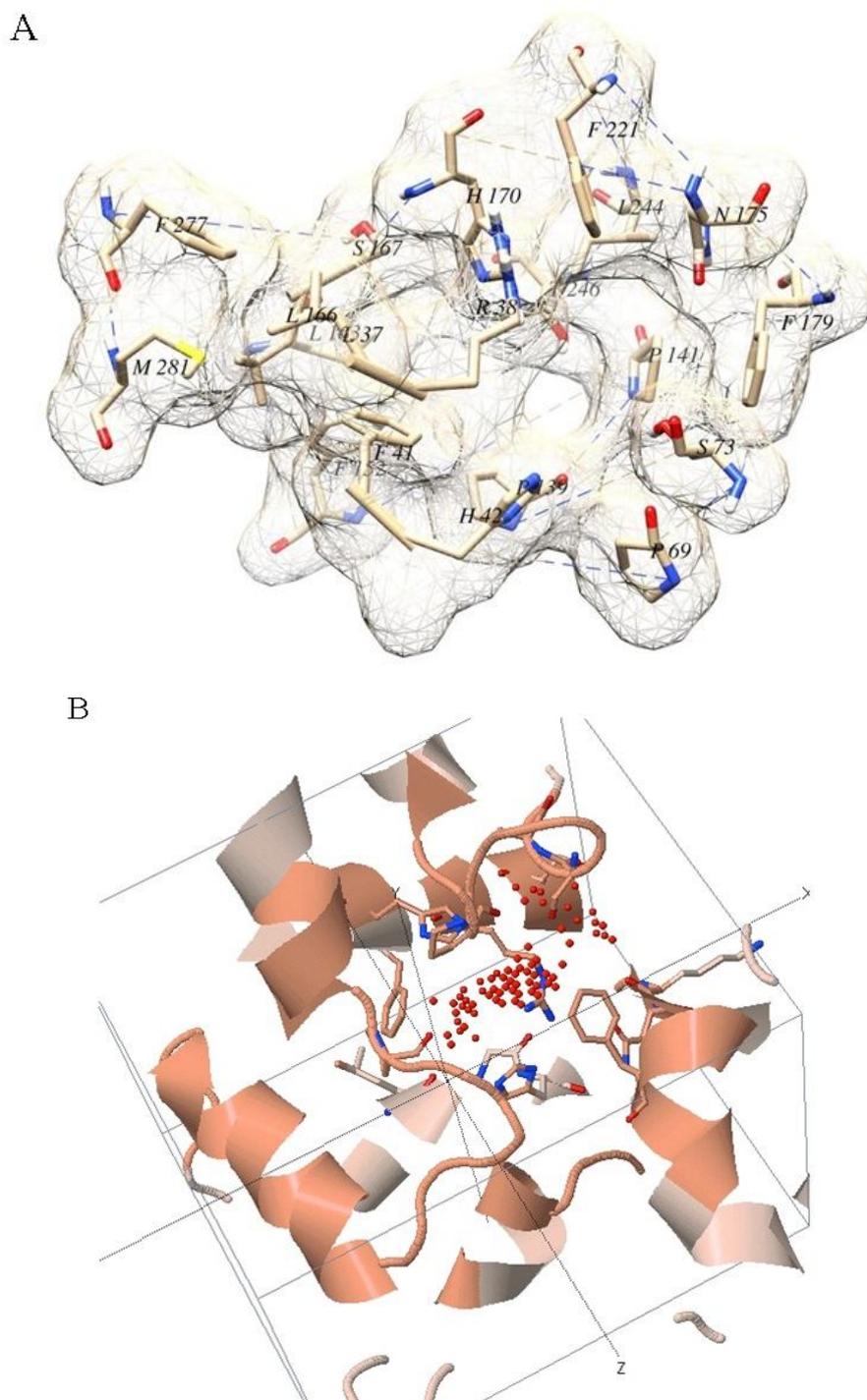


Fig. 7. A: Spatial motif with mesh surface corresponding to predicted binding site. The motif extracted from whole structure by introducing the binding site residues. B: Binding pocket comprised of binding residues and heme group in the center of pocket.

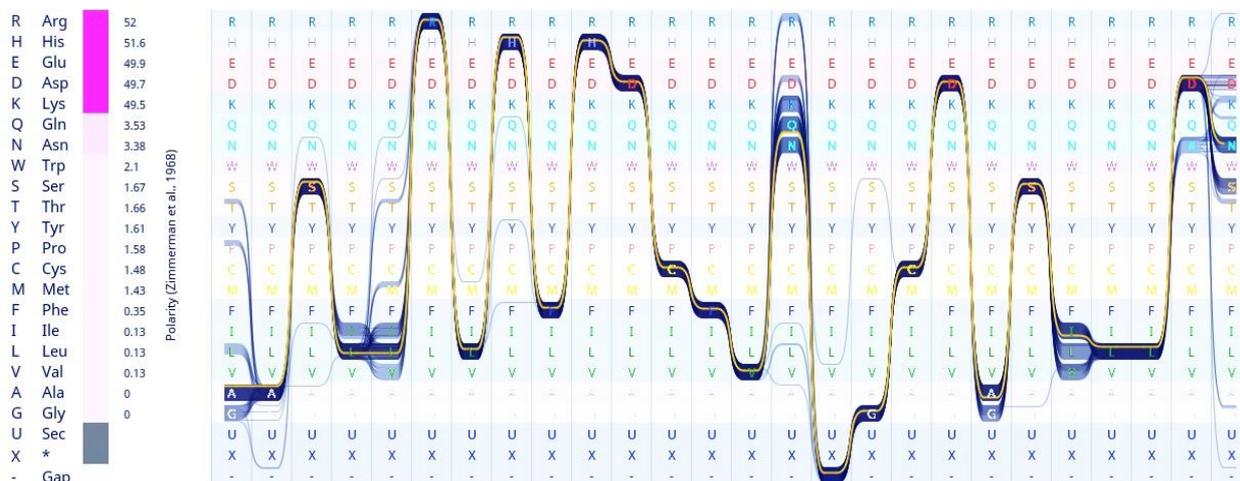


Fig. 8. Residues 38-69 of LDP, extracted from alignment file of 150 uniref90 homologous sequences. Each amino acid is indicated by one letter code, each blue line corresponds to each sequence in alignment file; The polarity bar to the left, scales the standard amino acids into polarity. The ochre line links the residues of LDP. This region is annotated in database as binding site and active site also is a portion of our prediction. Conservancy of residues is clearly evident in the picture.

(including LDP) were threonine (10.1 %), leucine (9.4), and serine (5.5). Nevertheless, the huge difference of pI can be explained by differences in frequencies of other amino acids.

We believe that getting access to a reliable structure, lead to deduce and gather numerous information. This assumption makes sense, while several authors concluded the functions and properties of proteins from crystal structures (see the reference [26] for more information). Building the tertiary structure of a protein from its primary sequence is a major scientific problem [28]. Bioinformatics algorithms are introduced in recent decade to solve the problem. Considering the high cost of experimental approaches to this mean [29], using bioinformatics tools is rationale despite its questionable outputs. Fortunately, there are number of algorithms for predicting the tertiary structure of protein from its amino acid sequence including: homology modeling, *ab initio*, and threading [30], and also various software are designed for quality assessment of predicted structure. Existence of several homologue models in protein data base for LDP, justified our approach [15] to choose homology modeling. Our final model underwent refinement and careful selection based on comparative

quality evaluation. The final model achieved a good agreement with Molprobit (Table 2) and of course encompass a high overall quality Fig. 4, similarly ramchandran plot confirms the reliability of structure, with high number of favored residues (more than expected) Fig. 5. Perfectly matching the predicted structure with reference structure is another support for the accuracy of our model. Moreover, this similarity revealed the same activity as horseradish peroxidase. In the previous study, the tertiary structure of LDP was modeled employing homology modeling technique at SWISS-MODEL server. The structure of LDP was solved using recombinant HRP C1A (PDB code 1gx2.1A) as a template. Our previous results demonstrated that similar to HRP, LDP contains heme-binding region and two calcium binding sites (unpublished data).

The LDP sequence is fallen into Clan: Peroxidase (CL0617) which contains 2323 species, this huge pool of species is supportive information for abundance of peroxidase activity among all domains of life. Choosing a shallow substitution matrix (BLOSUM 45) [31] in searching tactic, allowed us to find more distinct sequences even in prokaryotic species which indicating a

common evolutionary origin [32]

The identification of critical residues is a pivotal step for understanding the enzyme activity [33], in this regard, we attempted to find binding sites as critical residues. The residues involved in binding, all together made a motif which is in line with previous studies. Arg38, Phe41, His42, Asn70, His174 and Asp247 are important residues in binding site of HRP [34] which are interestingly conserved in LDP (unpublished data).

Peroxidases catalyze various oxidative reactions in which electrons are transferred to peroxide species (often H₂O₂) and substrate molecules are oxidized. Although peroxidase activity and related enzymatic machines are abundant in various organisms, finding the related motif similar to active site of our enzyme was failed likely due to software limitations and complexity of extracted motif.

Peroxidase activity has been detected in a number of enzymes, predominantly classified to EC 1.11.1.7. Here we introduce a new peroxidase sequence which is deposited in database under accession number AIJ1351. We built a reliable tertiary structure with well agreement to reference structure 4atj. The binding residues, hence structural motif, are in line with other studies that support the notion that our sequence is a neutral peroxidase which can be applicable in diagnostics and therapeutic purposes including cancer therapy while several *in vitro* studies have shown the effectiveness of peroxidase and indole acetic acid (IAA) couple to destroy tumors [35-37].

ACKNOWLEDGMENTS

We are grateful to Mr. Mohammad Reza Rahbar (Shiraz University of Medical Sciences) for his helpful discussion and manuscript editing.

REFERENCES

- [1] F.W. Krainer, A. Glieder, Appl. Microbial. Biotechnol. 99 (2015) 1611.
- [2] K.G. Welinder, Curr. Opin. Struct. Biol. 2 (1992) 388.
- [3] N.C. Veitch, Phytochemistry 65 (2004) 249.
- [4] A.M. Azevedo, V.C. Martins, D.M. Prazeres, V. Vojinović, J.M. Cabral, L.P. Fonseca, Biotechnol. Annual Rev. 9 (2003) 199.
- [5] G.A. Mulligan, C. Frankton, Can. J. Bot. 40 (1962) 1411.
- [6] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezuk, S. McGinnis, T.L. Madden, Nucl. Acids Res. 36 (2008) W5.
- [7] R.D. Finn, J. Clements, W. Arndt, B.L. Miller, T.J. Wheeler, F. Schreiber, A. Bateman, S.R. Eddy, Nucl. Acids Res. 43 (2015) W30.
- [8] C.G. Workbench, v3. 6.(2010), Now new version can be available at <http://www.clebio.com/products/cleclgenomicsworkbench>
- [9] C.-S. Yu, C.-W. Cheng, W.-C. Su, K.-C. Chang, S.-W. Huang, J.-K. Hwang, C.-H. Lu, PloS One. 9 (2014) e99368.
- [10] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, C. Notredame, Nucl. Acids Res. 34 (2006) W604.
- [11] S. Moretti, F. Armougom, I.M. Wallace, D.G. Higgins, C.V. Jongeneel, C. Notredame, Nucl. Acids Res. 35 (2007) W645.
- [12] R.F. Schwarz, A.U. Tamuri, M. Kultys, J. King, J. Godwin, A.M. Florescu, J. Schultz, N. Goldman, Nucl. Acids Res. 44 (2016) e77.
- [13] N. Darzentas, Bioinformatics 26 (2010) 2620.
- [14] S. Wang, W. Li, S. Liu, J. Xu, Nucl. Acids Res. 44 (2016) W430.
- [15] B. Webb, A. Sali, Protein Struct. Prediction 1137 (2014) 1.
- [16] L. Heo, H. Park, C. Seok, Nucl. Acids Res. 41 (2013) W384.
- [17] G.R. Lee, L. Heo, C. Seok, Proteins: Struct., Function and Bioinformatics. 84 (2015) 293.
- [18] A. Singh, R. Kaushik, A. Mishra, A. Shanker, B. Jayaram, Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1864 (2016) 11.
- [19] S. Lovell, I. Davis, W. Arendall, P. de Bakker, J. Word, M. Prisant, J. Richardson, D. Richardson, Proteins 50 (2003) 437.
- [20] Y. Zhang, J. Skolnick, Nucl. Acids Res. 33 (2005) 2302.
- [21] M.N. Wass, L.A. Kelley, M.J. Sternberg, Nucl. Acids Res. 38 (2010) W469.
- [22] Z. Zhang, Y. Li, B. Lin, M. Schroeder, B. Huang,

- Bioinformatics 27 (2011) 2083.
- [23] M. Hernandez, D. Ghersi, R. Sanchez, Nucl. Acids Res. 37 (2009) W413.
- [24] F. Kaiser, A. Eisold, D. Labudde, J. Comput. Biol. 22 (2015) 698.
- [25] A.C. Parte, Nucl. Acids Res. 42 (2013) D613.
- [26] D. Lee, O. Redfern, C. Orengo, Nat. Rev. Mol. Cell Biol. 8 (2007) 995.
- [27] K.G. Welinder, Plant Peroxidases 1980-1990, Topics and Detailed Literature on Molecular, Biochemical, and Physiological Aspects, 1992.
- [28] R. Cao, D. Bhattacharya, B. Adhikari, J. Li, J. Cheng, Bioinformatics 31 (2015) i116.
- [29] N. Guex, M.C. Peitsch, Electrophoresis 18 (1997) 2714.
- [30] D. Baker, A. Sali, Science 294 (2001) 93.
- [31] W.R. Pearson, Curr. Protocol. Bioinform. 43 (2013) 3.
- [32] K.G. ng Welinder, Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology. 1080 (1991) 215.
- [33] S. Sankararaman, F. Sha, J.F. Kirsch, M.I. Jordan, K. Sjölander, Bioinformatics 26 (2010) 617.
- [34] N.C. Veitch, A.T. Smith, Advances in Inorganic Chemistry, Academic Press, 2000.
- [35] C. Huang, L.-Y. Liu, T.-S. Song, L. Ni, L. Yang, X.-Y. Hu, J.-S. Hu, L.-P. Song, Y. Luo, L.-S. Si, World J. Gastroenterol. 11 (2005) 4519.
- [36] Y.-M. Jeong, M.H. Oh, S.Y. Kim, H. Li, H.-Y. Yun, K.J. Baek, N.S. Kwon, W.Y. Kim, D.-S. Kim, Die Pharmazie-An International Journal of Pharmaceutical Sciences 65 (2010) 122.
- [37] D.-S. Kim, S.-E. Jeon, K.-C. Park, Cellular Signalling 16 (2004) 81.