BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 10(2)

Genome wide survey and molecular modeling of hypothetical proteins containing 2Fe-2S and FMN binding domains suggests Rieske Dioxygenase Activity highlighting their potential roles in bio-remediation

Nitish Sathyanarayanan[1, 2] & Holenarsipur Gundurao Nagendra[1]*

[1]Department of Biotechnology, Sir M. Visvesvaraya Institute of Technology, Krishnadevarayanagar, Hunasamaranahalli, Bangalore 562 157; [2](Present Address) National Center for Biological Sciences, Tata Institute for Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560065; Holenarsipur Gundurao Nagendra - Email: nagshaila@gmail.com; *Corresponding author

Received January 13, 2014; Accepted January 26, 2014; Published February 19, 2014

Abstract:
'Conserved hypothetical' proteins pose a challenge not just for functional genomics, but also to biology in general. As long as there are hundreds of conserved proteins with unknown function in model organisms such as *Escherichia coli*, *Bacillus subtilis* or *Saccharomyces cerevisiae*, any discussion towards a 'complete' understanding of these biological systems will remain a wishful thinking. Insilico approaches exhibit great promise towards attempts that enable appreciating the plausible roles of these hypothetical proteins. Among the majority of genomic proteins, two-thirds in unicellular organisms and more than 80% in metazoa, are multi-domain proteins, created as a result of gene duplication events. Aromatic ring-hydroxylating dioxygenases, also called Rieske dioxygenases (RDOs), are class of multi-domain proteins that catalyze the initial step in microbial aerobic degradation of many aromatic compounds. Investigations here address the computational characterization of hypothetical proteins containing Ferredoxin and Flavodoxin signatures. Consensus sequence of each class of oxidoreductase was obtained by a phylogenetic analysis, involving clustering methods based on evolutionary relationship. A synthetic sequence was developed by combining the consensus, which was used as the basis to search for their homologs via BLAST. The exercise yielded 129 multi-domain hypothetical proteins containing both 2Fe-2S (Ferredoxin) and FNR (Flavodoxin) domains. In the current study, 17 proteins with N-terminus FNR domain and C-terminus 2Fe-2S domain are characterized, through homology modelling and docking exercises which suggest dioxygenase activity indicate their plausible roles in degradation of aromatic moieties.

Key Words: Hypothetical proteins, multi-domain redox proteins, FMN, FAD, 2F2-2S, FNR, Rieske dioxygenases, aromatic ring cleavage, Xenobiotic, MOSC.

Background:
Over the last decade, more than 150 complete genomes of diverse bacteria, archaea and eukaryotes have been sequenced, and many more are currently in the pipeline **[1].** It is well known that, in any newly sequenced bacterial genome, as many as 30–40% of the genes do not have an assigned function **[2].** This figure is even higher for archaeal and eukaryotic genomes and for the relatively large genomes of bacteria with a complex life style, such as *Anabaena*, *Streptomyces,* etc **[3, 4].**

'Conserved hypothetical' proteins pose a challenge not just to functional genomics, but also to biology in general **[5].** As long

as there are hundreds of conserved proteins of unknown function even in model organisms, such as *Escherichia coli*, *Bacillus subtilis* or *Saccharomyces cerevisiae*, any discussion of a 'complete' understanding of these organisms as biological systems will remain in the realm of wishful thinking. Although it appears likely that the central pathways of information processing and metabolism are already known, crucial elements of these systems could still be lurking among the 'conserved hypotheticals', and important mechanisms of signalling and stress response, in all likelihood, would remain undiscovered **[6].**

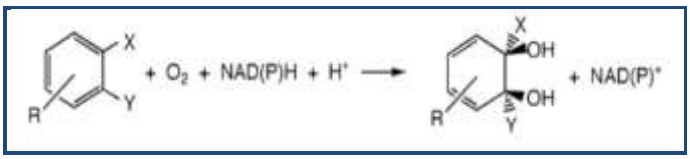


Figure 1: Reaction of ring cleavage mediated by RDO

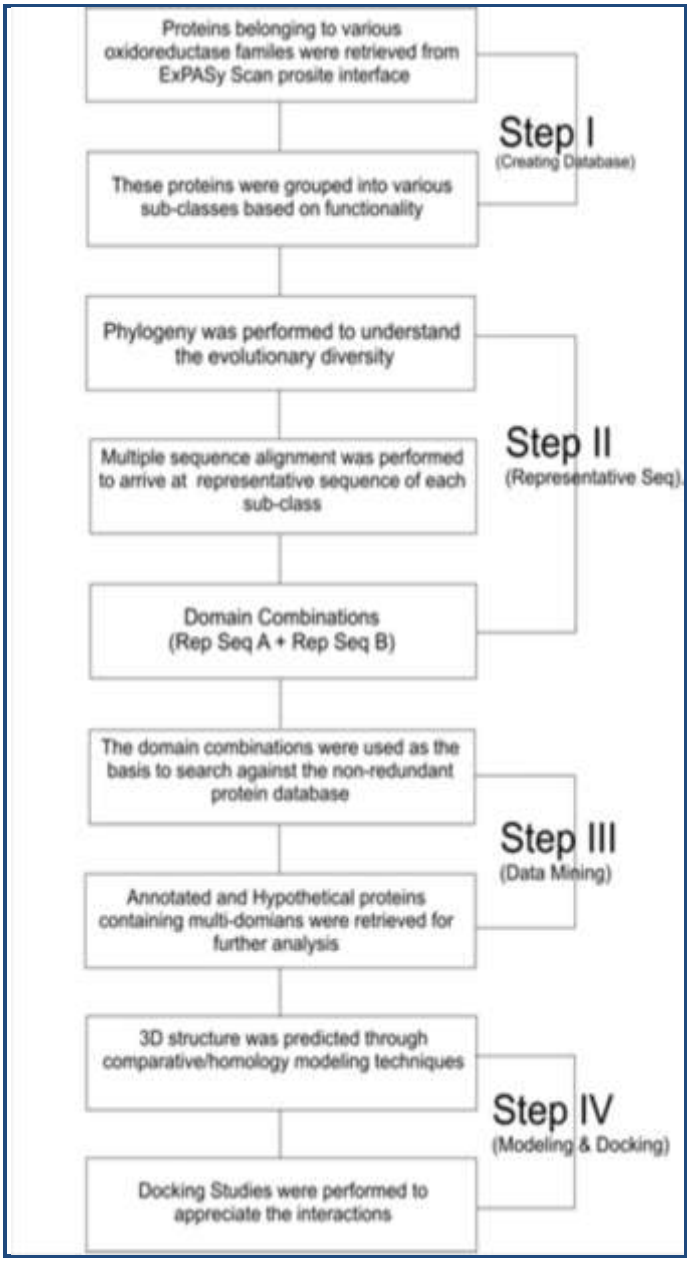


Figure 2: Protocol used in the present study

Aromatic compounds are widely distributed throughout the biosphere predominantly in the form of recycled material **[7]**. Because of the inherent thermodynamic stability of the aromatic ring, natural turnover of these compounds is slow and instead relies on complex microbial degradation pathways. With aromatic compounds comprising >25% of the earth's biomass, these pathways play a crucial role in the biogeochemical carbon cycle. However, despite the abundance of microbial degraders, man-made aromatic pollutants are often recalcitrant to existing bioprocessing pathways. As a result, these xenobiotic compounds, many of which are derived from the processing of crude oil, persist in the environment causing irreversible damage to the biosphere **[7].** Aromatic ring-hydroxylating dioxygenases, also called Rieske dioxygenases (RDOs), are class of multi-domain proteins that catalyze the initial step in microbial aerobic degradation of many aromatic compounds. Two hydroxyl groups are introduced into the aromatic ring yielding cyclic cis-dihydrodiols or cis-diol carboxylic acids **(Figure 1**) [Substituents X and Y can be hydrogen atoms or any of several other groups] **[8, 9].** More than three dozen distinct RDOs have been identified. RDOs consist of a reductase, an oxygenase and in some cases, an additional ferredoxin that mediates electron transfer between the former two components. The oxygenase component catalyzes the insertion of both atoms of molecular oxygen into the aromatic substrate, which is believed to occur at a mononuclear iron site and to be accompanied by electron insertion from a Rieske-type [2Fe–2S] centre. Either the reductase or, where present, the intermediary ferredoxin component, supplies the two electrons from NAD(P)H to the dioxygenase **[10]**. RDOs have been empirically classified according to the various combinations of subunits and electron transfer co-factors involved in reducing the oxygenase component **[10, 11]** as mentioned in **Table 1 (see supplementary material).**

Here we present a protocol to data mine and computationally characterize redox hypothetical proteins possessing multiple domains. Most proteins consist of multiple domains, and domains determine the function and evolutionary relationships of proteins **[12].** Thus, it is important to understand the principles of domain combinations and their associated inter domain interactions especially, in hypothetical proteins.

Primarily, 2Fe-2S (Ferredoxins) and FMN/FAD (Flavodoxins) were considered due to their vital and diverse roles in biological systems, the most important amongst it being their role in Electron Transport Mechanisms. Ferredoxins are small, acidic, electron transfer proteins that are ubiquitous in biological redox systems. Members of the 2Fe-2S ferredoxin family have a general core structure consisting of beta (2)-alpha-beta (2). They are proteins of around one hundred amino acids with four conserved cysteine residues to which the 2Fe-2S cluster is ligated **[13]**. Flavoenzymes have the ability to catalyse a wide range of biochemical reactions. They are involved in the dehydrogenation of a variety of metabolites, in electron transfer from and to redox centres, in light emission, in the activation of oxygen for oxidation and hydroxylation reactions. About 1% of all eukaryotic and prokaryotic proteins are predicted to encode a flavin adenine dinucleotide (FAD) or

Flavin Mono Nucleotide (FMN)-binding domains which are involved in electron transport **[14].**

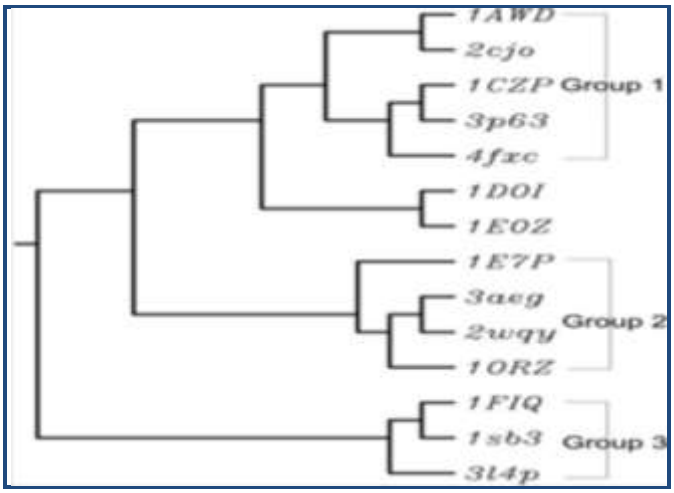


Figure 3: Phylogenetic tree of 2Fe-2S family.

Figure 4: MSA of group 1 of 2Fe-2S family

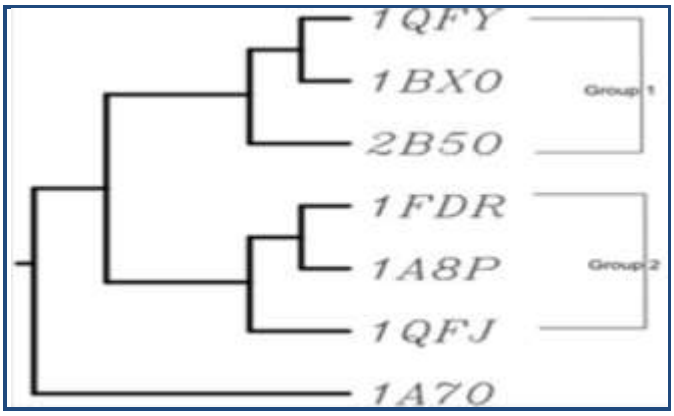


Figure 5: Phylogenetic tree for FNR reductase family

Methodology:

The proteins belonging to oxidoreductase (Ferredoxin, Flavodoxin) families were retrieved from the ExPASy Prosite interface **[15]**. However, engineered and mutated sequences were not considered to avoid redundancy. Additionally, only reviewed sequence from Uniprot containing a structural entry were considered. Binding sites of all the proteins belonging to the same group were analyzed in order to arrive at a consensus pattern through multiple sequence alignment. Extended regions which had no information with the other sequences were clipped to strengthen the alignment. The protocol adopted is shown in **Figure 2**.

The search for Ferredoxin family (PDOC00175) yielded 14 sequences with 2Fe-2S binding signature. As there existed heterogeneity within the group, the sequences were clustered based on phylogenetic analysis. The sequence alignment was performed through ClustalW **[16]** and the tree was obtained using MEGA (NJ method) **[17].** The tree obtained is shown in **Figure 3**. Further to the clustering, multiple sequence alignment was performed using Multalin **[18],** for all the 3 clusters (groups) to obtain a representative sequence containing strong signatures. The multiple sequence alignment of sequences belonging to group 1 yielded better consensus compared to the other clusters, which is as depicted in **Figure 4**.

Similarly, the search for flavodoxin family (FNR reductase - PDOC51384) yielded 7 sequences, whose Phylogenetic tree is shown in **Figure 5.** When multiple sequence alignments of both the clusters were critically analyzed, the MSA of group 1 exhibited strong signatures of the FNR domain when compared to cluster 2, which is depicted in **Figure 6**. Thus, a consensus of the cluster of sequences from group 1, for both the 2Fe-2S and FNR domains respectively, were considered as possible representative patterns, towards generating the probable synthetic sequence, which was used as the basis for BLAST tool **[19]** search against the non-redundant database. Interestingly, this approach yielded 2078 sequences, and clearly contained both 2Fe-2S and FNR domains when analysed through the conserved domain database (CDD) **[20].** Amongst these 2078 sequences, 129 belonged to that special class of hypothetical proteins, which were taken up for further characterization and analysis.

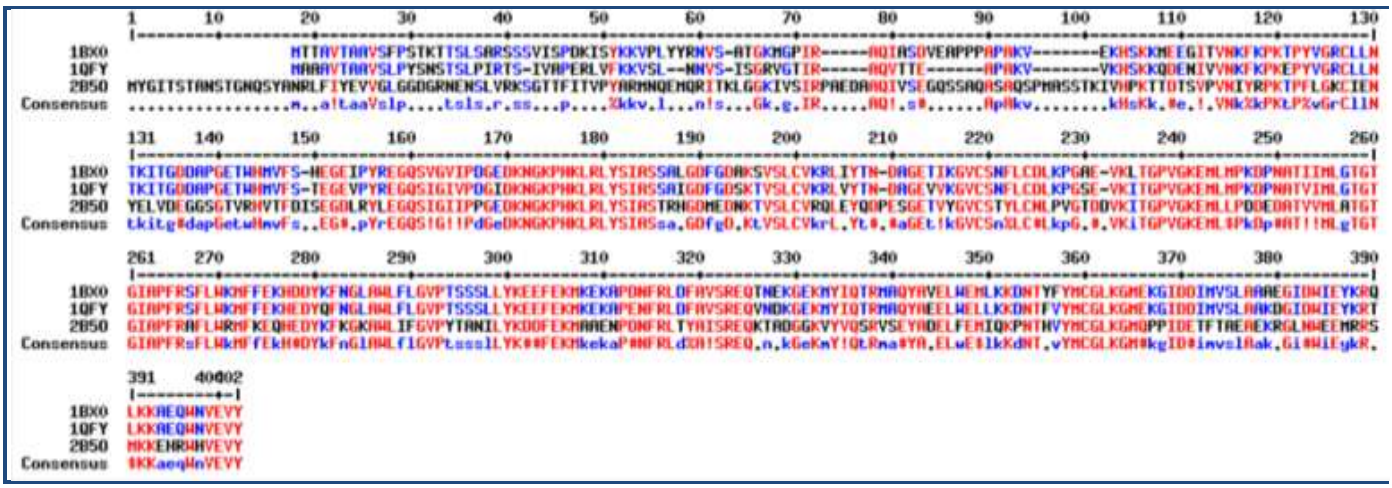


Figure 6: MSA of group 1 of FNR reductase family.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 10(2): 068-075 (2014)

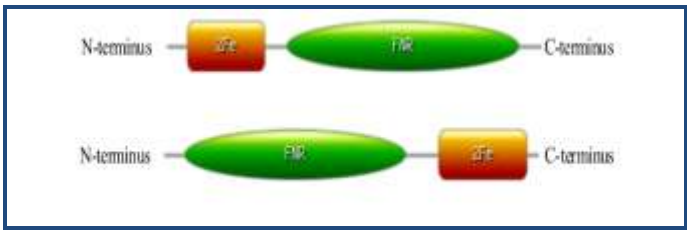


Figure 7: Domain architecture in RDO-reductase class.

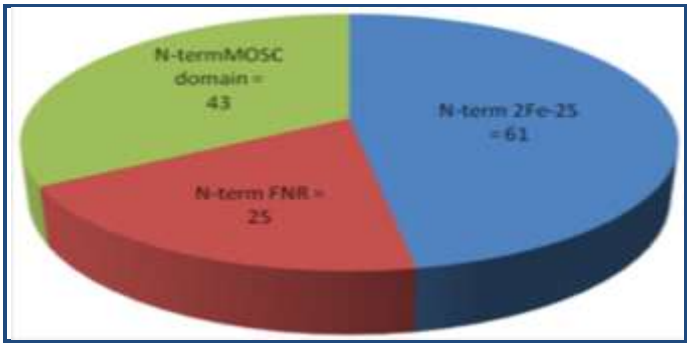


Figure 8: Pie-chart showing the distribution of domains in the 129 hypothetical proteins.

Figure 9: Phylogenetic tree of the hypothetical proteins containing Phylogenetic tree of N-terminus FNR and C-terminus 2Fe-2S

Results & Discussions:

Upon critical evaluation of the 129 multi-domain hypothetical sequences through CDD, significant differences in the location of 2Fe-2S domain, relative to other domains, were found. Of these 129 sequences, 61 contained an N-terminus 2Fe-2S and a C-terminus FNR domain while this order was reversed in 25 sequences as shown in the **Figure 7.** The remaining 43 sequences contained an N-terminus MOSC domain **[21]** (pfam03473 and pfam03476) which is a super family of beta-strand-rich domains identified in the molybdenum cofactor sulfurase and several other proteins from both prokaryotes and eukaryotes. The MOSC domain is predicted to be a sulfur-carrier domain that receives sulfur abstracted by the pyridoxal phosphate-dependent NifS-like enzymes, on its conserved cysteine, and delivers it for the formation of diverse sulfur-metal clusters. The pie chart in **Figure 8** illustrates the distribution of the domains amongst these 129 proteins. In the current study, 25 sequences containing N-terminus FNR and C-terminus 2Fe-2S domain are considered. 61 sequences which contain an N-terminus 2Fe-2S and C-terminus FNR domain has been critically analysed **[22]** while 43 sequences which contain MOSC domain will be considered for modelling in near future. The phylogenetic analysis of the sequences containing a N-terminus FNR and C-terminus 2Fe-2S domains is depicted in **Figure 9.**

The sequences were searched against the PDB database (using the PDB BLAST tool) towards identification of a suitable template. This yielded 2PIA **[11]**, a phthalate dioxygenase reductase from Burkholderia cepacia. Phthalate dioxygenase reductase (PDR) is a prototypical iron-sulfur flavoprotein (36 kilodaltons) that utilizes flavin mononucleotide (FMN) to mediate electron transfer from the two-electron donor, reduced nicotinamide adenine nucleotide (NADH), to the one-electron acceptor, [2Fe-2S]. Of these 25 sequences, 8 sequences had very low (<20%) sequence identity with the template 2PIA, and hence were discarded from further analysis due to lack of clarity. The remaining 17 sequences were considered with confidence for homology modelling exercises, as they exhibited high overall similarity with 2PIA. The overall sequence identity between the query and template was between 20-30% for all the other sequences except 2 sequences which was about 30-50%. The **Figure 10** shows the distribution of the overall sequence identity, identities at the FMN and 2Fe-2S binding regions for each sequence, which clearly illustrates the conservation at critical regions of functional relevance.

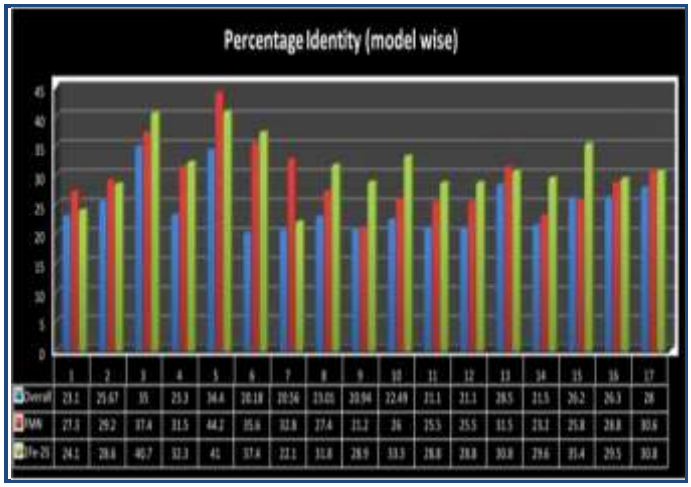


Figure 10: Bar graph showing the overall sequence identity (blue), identity at FMN binding region (red) and 2Fe-2S binding region (green) against the model 2PIA (Please see table 3 for cross-reference).

The FNR family contains two conserved motifs, viz., (R-x-Y-[ST]) where positively charged Arg residue forms hydrogen bonds to the pyrophosphate oxygen atom and (G-x(2)-[ST]-x(2)-L-x(5)-G-x(7)-P-x-G) which is the phosphate-binding motif **[14].** Similarly, 4 conserved Cys residues at positions i, i+5, i+8 and variable i+38 is required for binding of 2Fe-2S ligand **[13].** Both the FMN and 2Fe-2S binding regions are highly

conserved in all the 17 models. In view of the poise in the signatures between the template and the 17 target sequences, model building exercises were carried out with Swissmodel automated mode **[23].** The RMSD between the modelled structure and template for the Ca- atoms confirmed the quality of the models in spite of seemingly low sequence identity **Table 3 (see supplementary material)** and **Figure 10**), in addition to the satisfaction of various criteria calculated using ProCheck **[24]**. Individual models were analysed for the binding of ligands through docking studies which was performed using FlexX algorithm **[25].** To define the structural and functional aspects of the hypothetical protein sequences, modelling of GI ID 289441001 is considered as a case study.

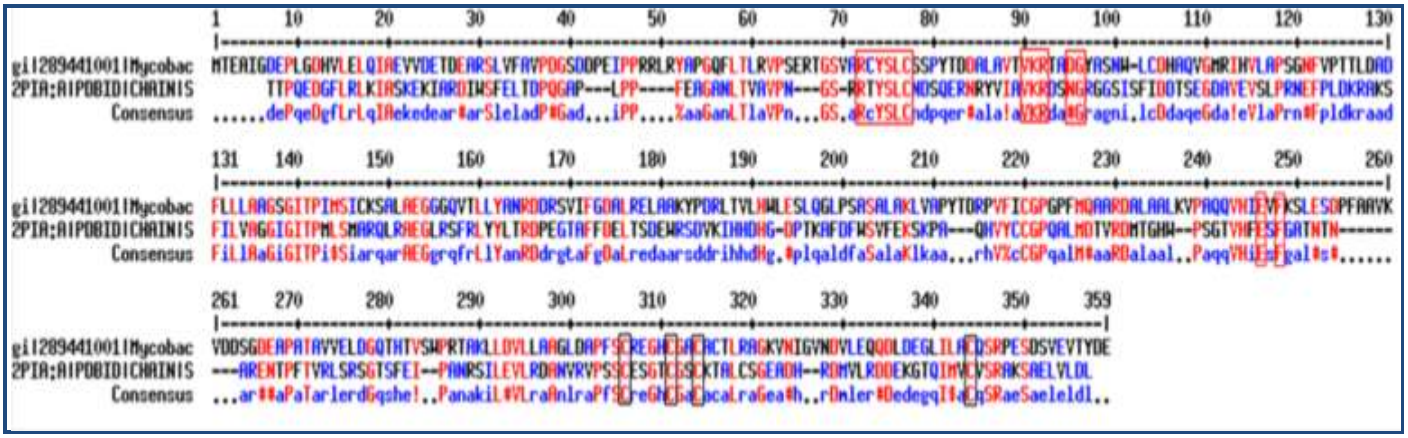


Figure 11: Template to query alignment (FMN motif marked in red and 2Fe-2S binding motif marked in black – C305, C310, C313, C348).

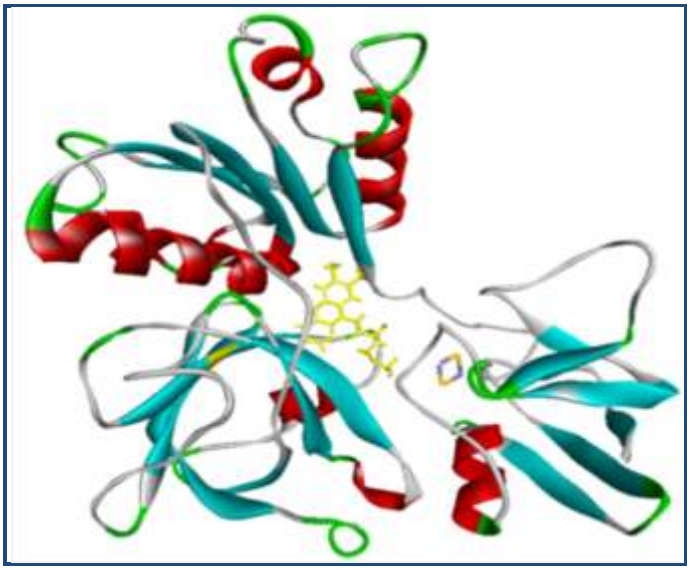


Figure 12: Modelled structure of GI 289441001.

2PIA based model for GI ID 289441001

The query protein 289441001 from *Mycobacterium tuberculosis* was successfully modelled using SWISS model interface, where the overall identity between the query and template is 26.3 %. The alignment between the template and query is shown in **Figure 11**. In spite of the low overall sequence identity, it can be appreciated that the binding regions of 2Fe-2S and FMN exhibit high conservation. The RMSD between the modelled structure and template is found to be 0.22 Å (for 93.2% of the atoms superposed) for Ca atoms. The quality of the model was assessed with PROCHECK (ramachandran map analysis) where 97.7% of the residues were in allowed region and only 2.3% residues were in disallowed region. Interestingly, none of these residues in the outlier regions belong to the functionally important residues. The 2Fe-2S and FMN ligands were docked into the model and all the interactions were found similar to that of the template. The binding of 2Fe-2S Ligand and FMN are shown in **Figures 12, 13 & 14. Table 2 (see supplementary material)** summarizes the residues forming the Pharmacophore (4 Å radius) for FMN ligand in template, FMN ligand redocked to template and model where high residue conservation is observed. The docking of the FMN to the template (using the program FlexX) was done to re-confirm the ligand binding pose, and normalize the artefacts due to the software, if any. The residues highlighted in bold forms H-bonds with the FMN, which further reiterates decent bind of the ligand. The modelled and docked structures were deposited at the Protein Model Data Bank (PMDB) **[26]** where all the models were judged to possess clashes within acceptable limits. **Table 3** summarises the details of all the 17 models generated with 2PIA (which contains 322aa) as the template.

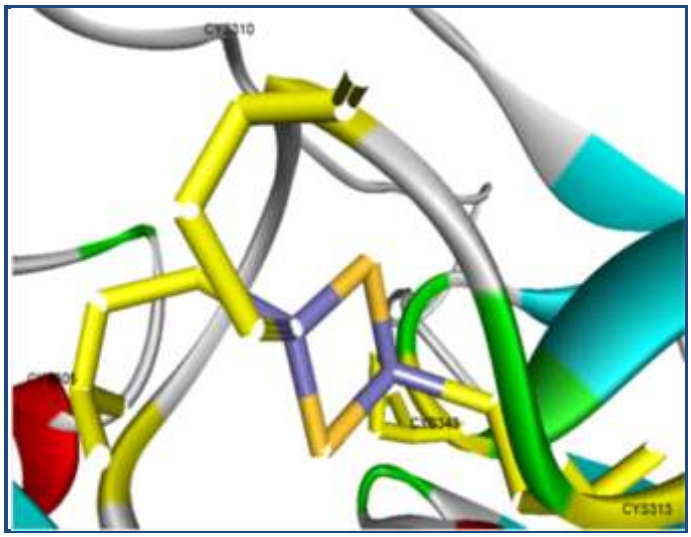
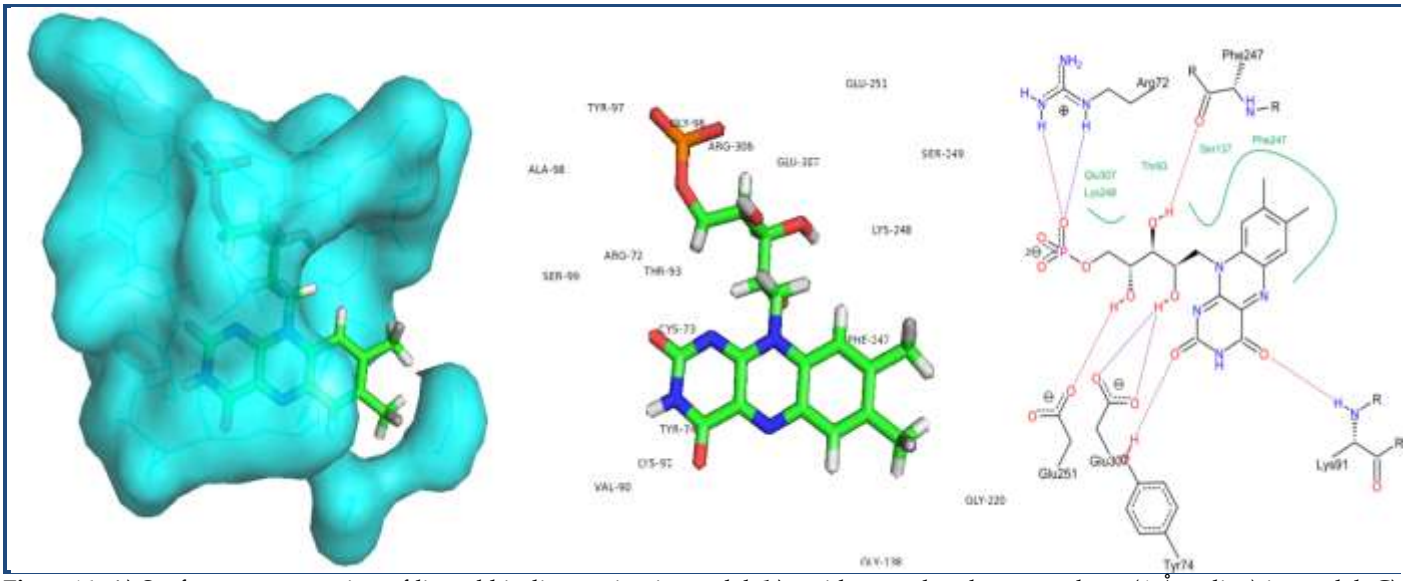


Figure 13: 2Fe-2S interaction in the model with the conserved cysteins (C305,C310,C313,C348).



Figue 14: A) Surface representation of ligand binding region in model; **b)** residues at the pharmacophore (4 Å radius) in model; **C)** 2D representation of the ligand-residue interaction in model

Conclusion:

129 hypothetical proteins from across the genomes have been data mined, and the 3D description of 17 sequences has been derived with confidence. The statistics related to comparative modelling and docking studies (with acceptable energy values) have revealed a strong interaction of 2 redox ligands, viz., 2Fe-2S and FMN with the binding residues, which further strengthens the argument of these proteins being involved in cleavage of aromatic compounds. Though degradation of aromatic compounds by microorganisms is a well established Fact **[27, 28],** characterization of hypothetical sequences in the Present study could aid in better understanding of these microbial systems. A large number of microbial systems containing these dioxygenases have also been mined and characterized in the present investigation, which could provide insights into their degradation properties. Thus, this study on multi-domain hypothetical proteins could prove critical in two ways viz., in understanding the mechanism of uptake of nutrients which contain aromatic ring structures and hence enabling engineering of these proteins towards effective degradation of harmful xenobiotics.

References:

[1] Galperin MY & Koonin E, *Nucleic Acids Res*. 2004 **32**: 5452 [PMID: 15479782]
[2] Bork P, *Genome Res.* 2000 **10:** 398 [PMID: 10779480]
[3] Bentley SD et al. *Nature* 2002 **417**: 141 [PMID: 12000953]
[4] Kaneko T *et al. {DNA} Res*. 2001 **8:** 205 [PMID: 11759840]
[5] Galperin MY, *Comp Funct Genomics*. 2001 **2**: 14 [PMID: 18628897]
[6] Galperin MY & Koonin EV, *Nucleic Acids Res.* 2004 **32:** 5452 [PMID: 15479782]
[7] Gibson J & Harwood CS, *Annu Rev Microbiol.* 2002 **56**: 345 [PMID: 12142480]
[8] Bertini MA *et al. Coord Chem Rev*. 1996 **151:** pp145
[9] Butler CS & Mason JR, *Adv Microb Physiol*. 1997 **38:** 47 [PMID: 8922118]
[10] Karlsson A *et al. J Mol Biol*. 2002 **318**: 261 [PMID: 12051836]
[11] Correll CC *et al. Science* 1992 **258**: 1604 [PMID: 1280857]
[12] Vogel C, *Curr Opin Struct Biol.* 2004 **14:** 208 [PMID: 15093836]
[13] Valentine RC, *Bacteriol Rev*. 1964 **28**: pp497
[14] Joosten V & Van Berkel WJ, *Curr Opin Chem Biol*. 2007 **11**: 195 [PMID: 17275397]
[15] Bairoch A, *Nucleic Acids Res.* 1993 **21**: 3097 [PMID: 8332530]
[16] Thompson JD, *Nucleic Acids Res*. 1994 **22:** 4673 [PMID: 7984417]
[17] Kumar S *et al. Brief Bioinform*. 2008 **9**: 299 [PMID: 18417537]
[18] Corpet F *Nucleic Acids Res*. 1988 **16**: 10881 [PMID: 2849754]
[19] Altschul SF *et al. Nucleic Acids Res.* 1997 **25**: 3389 [PMID: 9254694]
[20] Marchler-Bauer A, *Nucleic Acids Res*. 2011 **39:** D225 [PMID: 21109532]
[21] Anantharaman V & Aravind L, *FEMS Microbiol Lett*. 2002 **207**: 55 [PMID: 11886751]
[22] Sathyanarayanan N & Nagendra HG, *Bioinformation*. 2012 **8:** 1154 [PMID: 23275712]
[23] Schwede T *et al. Nucleic Acids Res*. 2003 **31:** 3381 [PMID: 12824332]
[24] Laskowski RA, *J Appl Crystallogr*. 1993 **26**: pp283
[25] Kramer B *et al. Proteins* 1999 **37**: 228 [PMID: 10584068]
[26] Castrignano T, *Nucleic Acids Res.* 2006 **34**: D306 [PMID: 16381873]
[27] Gibson DT *et al. Biochemistry* 1970 **9:** pp1631
[28] Cowles CE *et al. J Bacteriol*. 2000 **182**: 6339 [PMID: 11053377]

Edited by P Kangueane
Citation: **Sathyanarayanan & Nagendra,** Bioinformation 10(2): 068-075 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: Components of Rieske dioxygenases

Class	Reductase	Intermediate electron transfer component	Oxygenase	Examples
IA	FMN Cys4[2Fe–2S]	None	Cys2His2 [2Fe–2S] Fe^{2+}	Phthalate dioxygenase
IB	FAD Cys4[2Fe–2S]	None	Cys2His2 [2Fe–2S] Fe^{2+}	Benzoate dioxygenase
IIA	FAD	Cys_4[2Fe–2S]	Cys2His2 [2Fe–2S] Fe^{2+}	Dibenzofuran 4,4a-dioxygenase
IIB	FAD	Cys_2His_2[2Fe–2S]	Cys2His2 [2Fe–2S] Fe^{2+}	Biphenyl dioxygenase
III	FAD Cys4[2Fe–2S]	Cys_2His_2[2Fe–2S]	Cys2His2 [2Fe–2S] Fe^{2+}	Naphthalene dioxygenase

Table 2: Residues in the Pharmacophore in the template and model for FMN

No.	Residues with in 4 Å in template	Residues with in 4 Å in template (FMN redocked)	Residues with in 4 Å in model
1	R55	R55	R72
2	T56	T56	C73
3	Y57	Y57	Y74
4	S58	S58	S75
5	V73	V73	V90
6	K74	K74	K91
7	R75	R75	R92
8	G79	G79	G96
9	R80	R80	Y97
10	G81	G81	A98
11	E223	E223	E245
12	F225	F225	F247
13	S274	S274	A320

Table 3: Summary of 17 models

No	Multi domain hypo protein	Species	AA	RMSD (Ca Atoms)	PMID Id-
1.	107104513	Pseudomonas aeruginosa PACS2	391	0.41 [76%]	PM0077744
2.	11610564	Pseudomonas sp. Y-2	335	0.28 [90%]	PM0078733
3.	107103607	Pseudomonas aeruginosa PACS2	321	0.34 [95%]	PM0078734
4.	254822729	Mycobacterium intracellulare ATCC 13950	364	0.26 [85]	PM0078735
5.	41409379	Mycobacterium avium subsp. paratuberculosis K-10	362	0.33 [83]	PM0078736
6.	41409643	Mycobacterium avium subsp. paratuberculosis K-10	364	0.43 [81%]	PM0078737
7.	163754039	Kordia algicida OT-1	357	0.40 [83%]	PM0078738
8.	90407852	Psychromonas sp. CNPT3	336	0.42 [85%]	PM0078739
9.	154497184	Bacteroides capillosus ATCC 29799	386	0.39 [77%]	PM0078740
10.	149912010	Moritella sp. PE36	350	0.47 [81%]	PM0078741
11.	212712724	Providencia alcalifaciens DSM 30120	323	0.41 [90%]	PM0078742 *
12.	188026245	Providencia stuartii ATCC 25827	323	0.41 [90%]	PM0078742 *

13.	41406594	Mycobacterium avium subsp. paratuberculosis K-10	364	0.33 [86%]	PM0078743
14.	226328372	Proteus penneri ATCC 35198	323	0.44 [89%]	PM0078744
15.	288549425	Enterobacter cancerogenus ATCC 35316	311	0.46 [91%]	PM0078745
16.	289441001	Mycobacterium tuberculosis T46	358	0.22 [86%]	PM0078746
17.	342862403	Mycobacterium colombiense CECT 3035	364	0.21 [86%]	PM0078747

[* 100% Identical sequences from different species]