Structure-function relationship of FAD-dependent pyridine nucleotide-disulfide oxidoreductase in *Anaeromyxobacter dehalogenans* 2CP-C reveals perhaps, a unique mechanism for transferring reducing equivalents

Sana Wajid

School of Environmental and Biological Sciences

Rutgers, State University of New Jersey.

December 20, 2013

CONTENTS

ABBREVIATIONS

1LQT = target PDB code

466500 = target

B = β-bridge, S = bend

C = mixed coiled coil.

E = β sheet

G = 3_{10} helix

H = α helix

I = Insight Model

P = Phyre model

R = Rosetta model

T = H-bonded turn

Tem = Template

Thr = Threaded model

Φ = Phi

Ψ = Psi

FIGURES

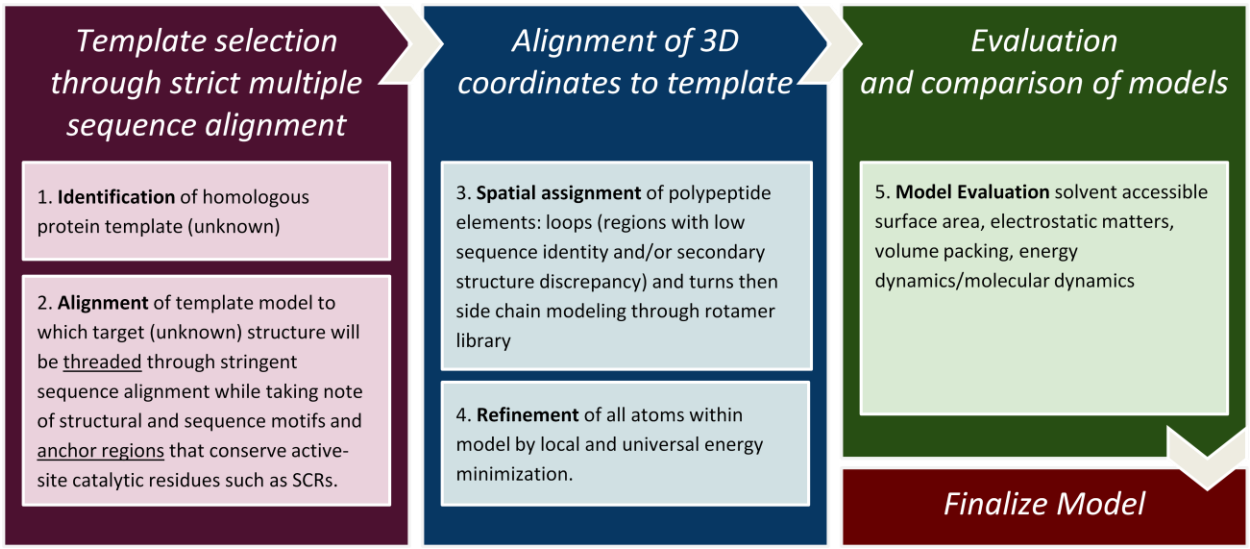


Figure 1. General outline of homology modeling procedure (20-26, 30-33).

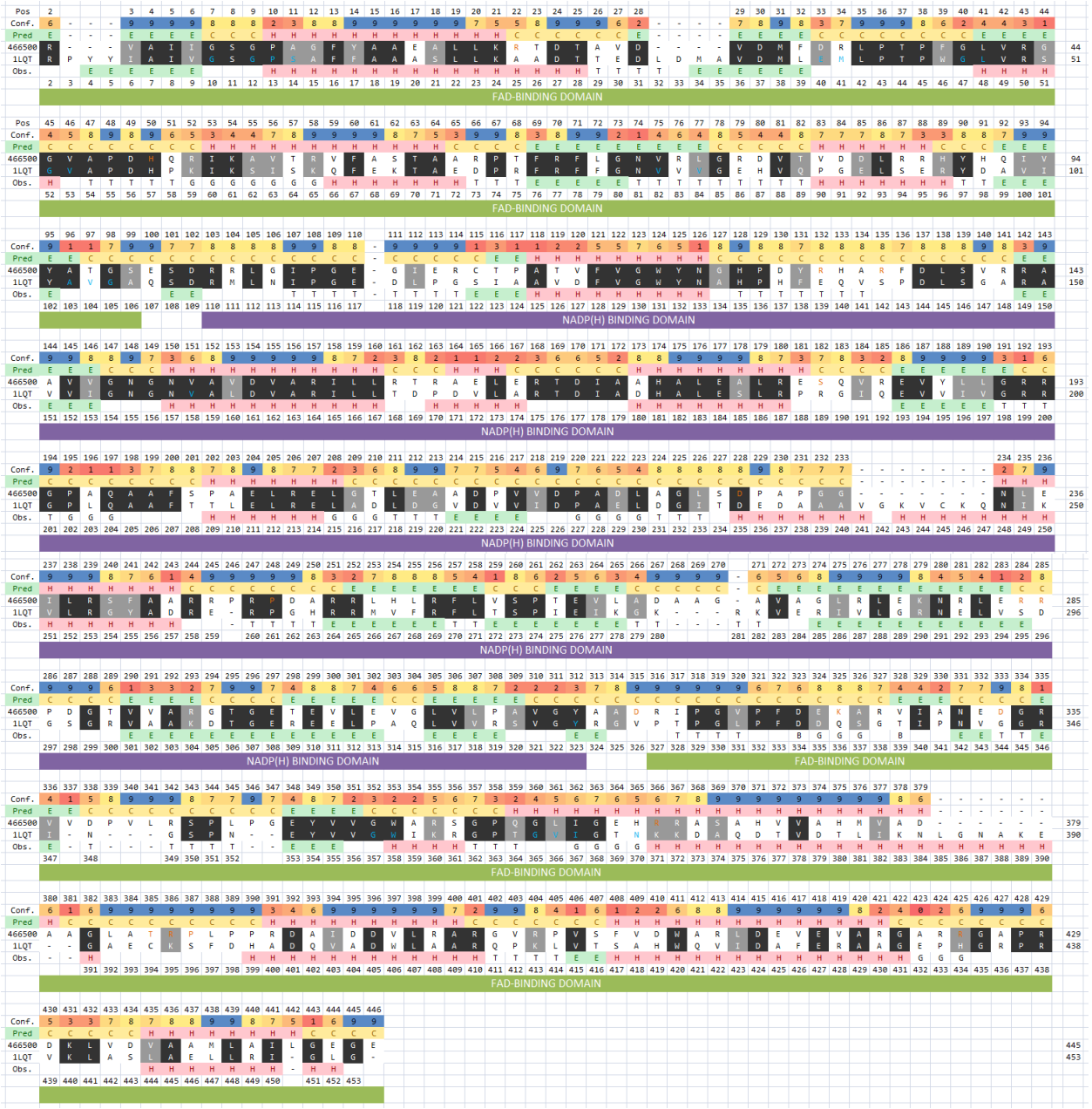


Figure 2. Alignment of 466500 against template 1LQT chain A. ISIS sites are colored orange. PDBe sites are colored blue. Domains: FAD and NADP(H) from chain A are outlined within the figure. Conf = CONCORD prediction Secondary Structure Confidence score, Pred = predicted SS structure where H = α helix, E = β sheet, C = mixed coiled coil. Under observed SS structure: G = 3_{10} helix, T = H-bonded turn, B = β-bridge, S = bend. (45)

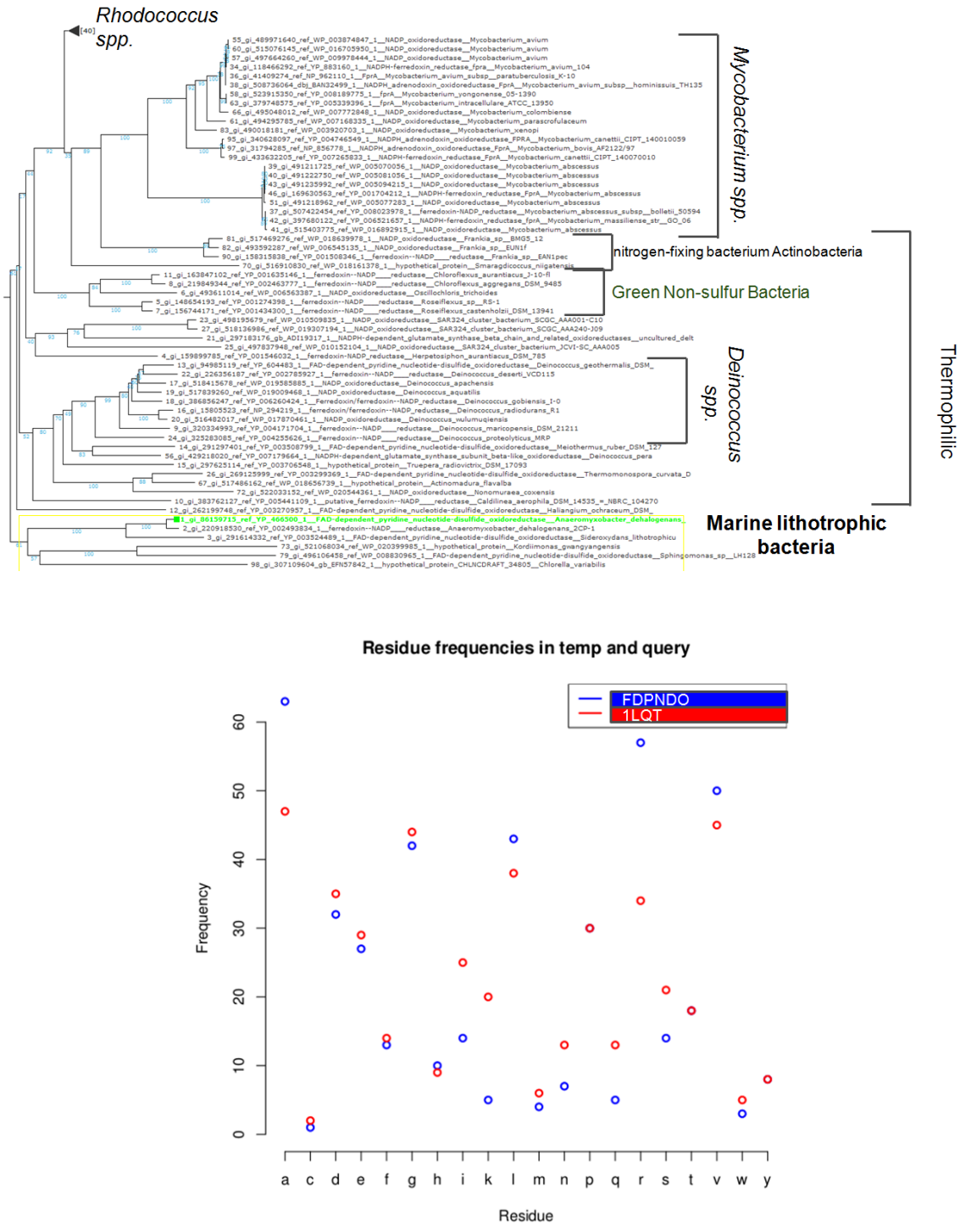


Figure 3. Sequence discovery and analysis of template and target. **A.** Phylogenetic tree of template and target sequences along with top 100 other PSI-BLAST hits found during part 1 of profile-profile alignment (*see Methods*). Tree is created by neighbor-joining method with 1,000 bootstrap samples using MAFFT. **B.** Template and target residue frequencies.

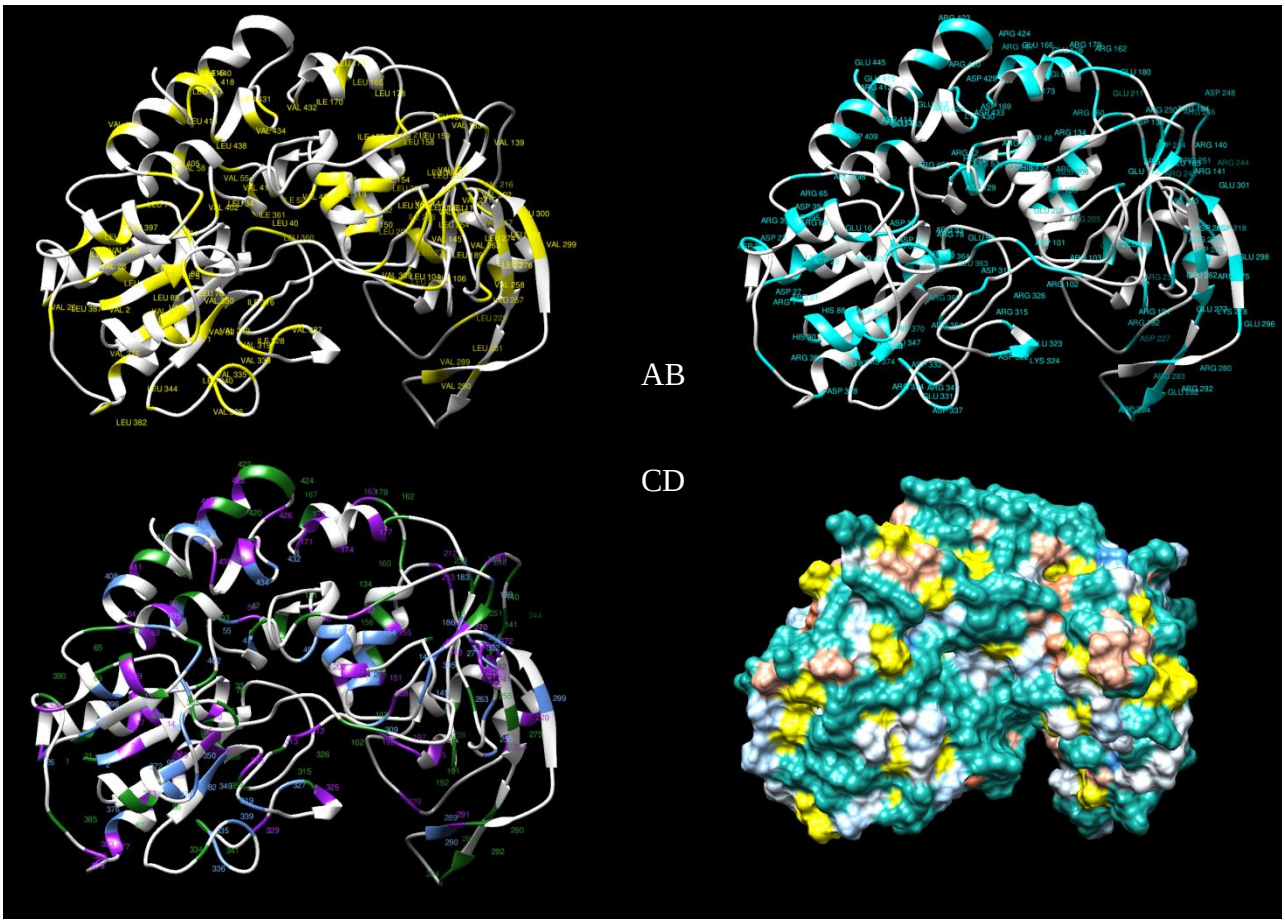


Figure 4. Insight Model (I) is used to depict homology model of protein 466500. For all figures, NADP(H) domain is on the right cleft, while FAD-domain is on the left cleft. Ligand molecules are omitted. **A.** Alipathic side chains are in yellow. **B.** Charged side chains are cyan. **C.** Side chains-color: Ala-purple; Arg-Green; Val-blue. **D.** Solvent accessible surface area of (3C) is modeled with 1.4 probe area.

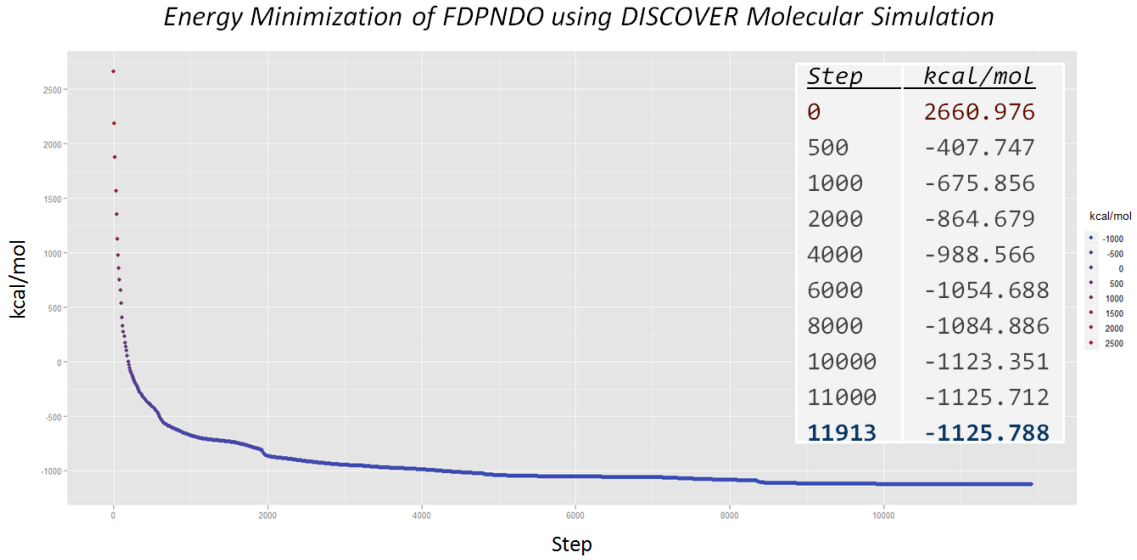


Figure 5. Following model constructing using Accelery's Insight II software, a DISCOVER 3.0 all-atom energy minimization was conducted using 100,000 iterations as plotted.

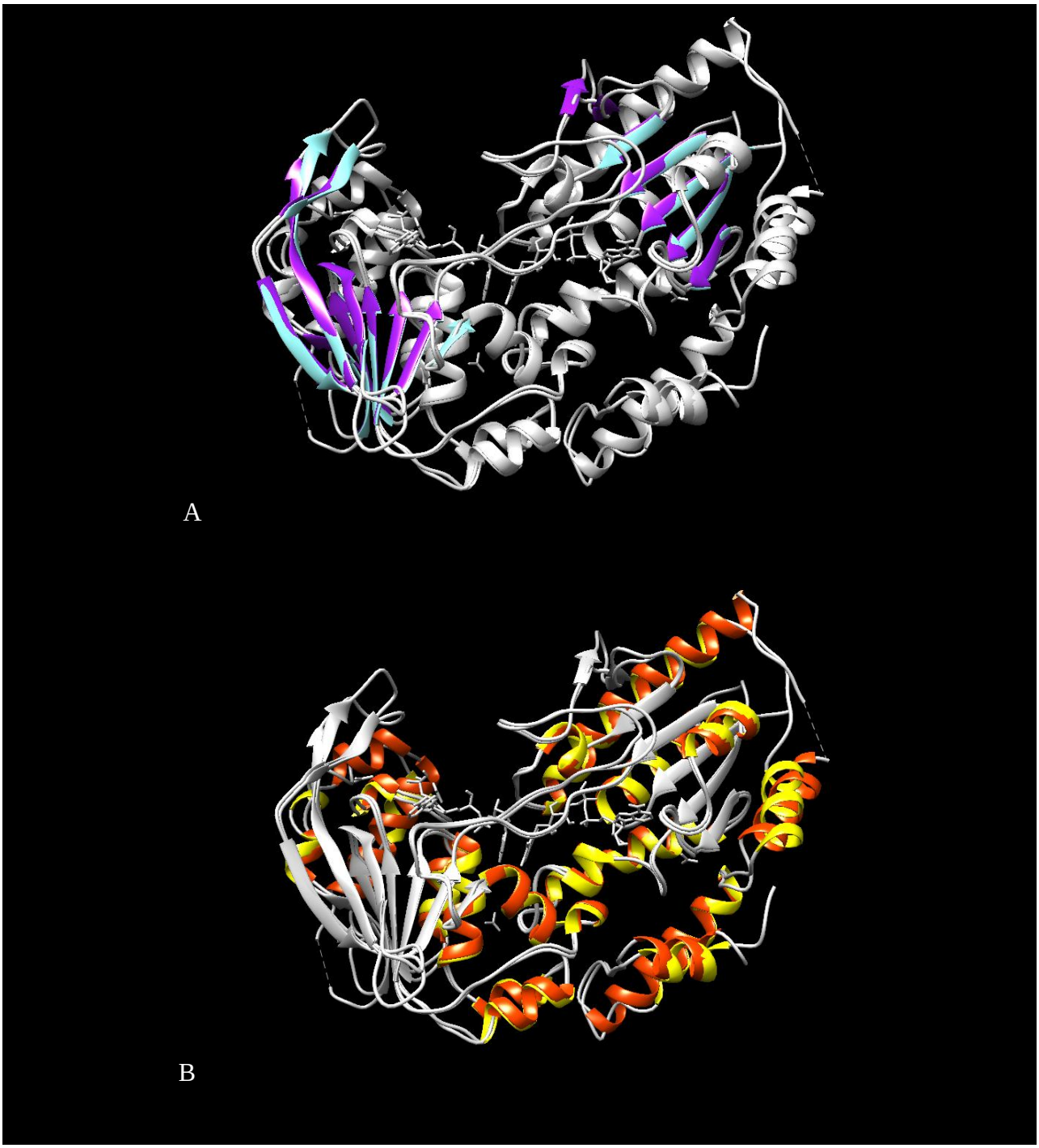


Figure 6. Superposition of Insight II model with template to highlight areas of major secondary structure homology. **A.** Blue - template, purple = target; FAD on right. **B.** Red - template, yellow = target; FAD on right

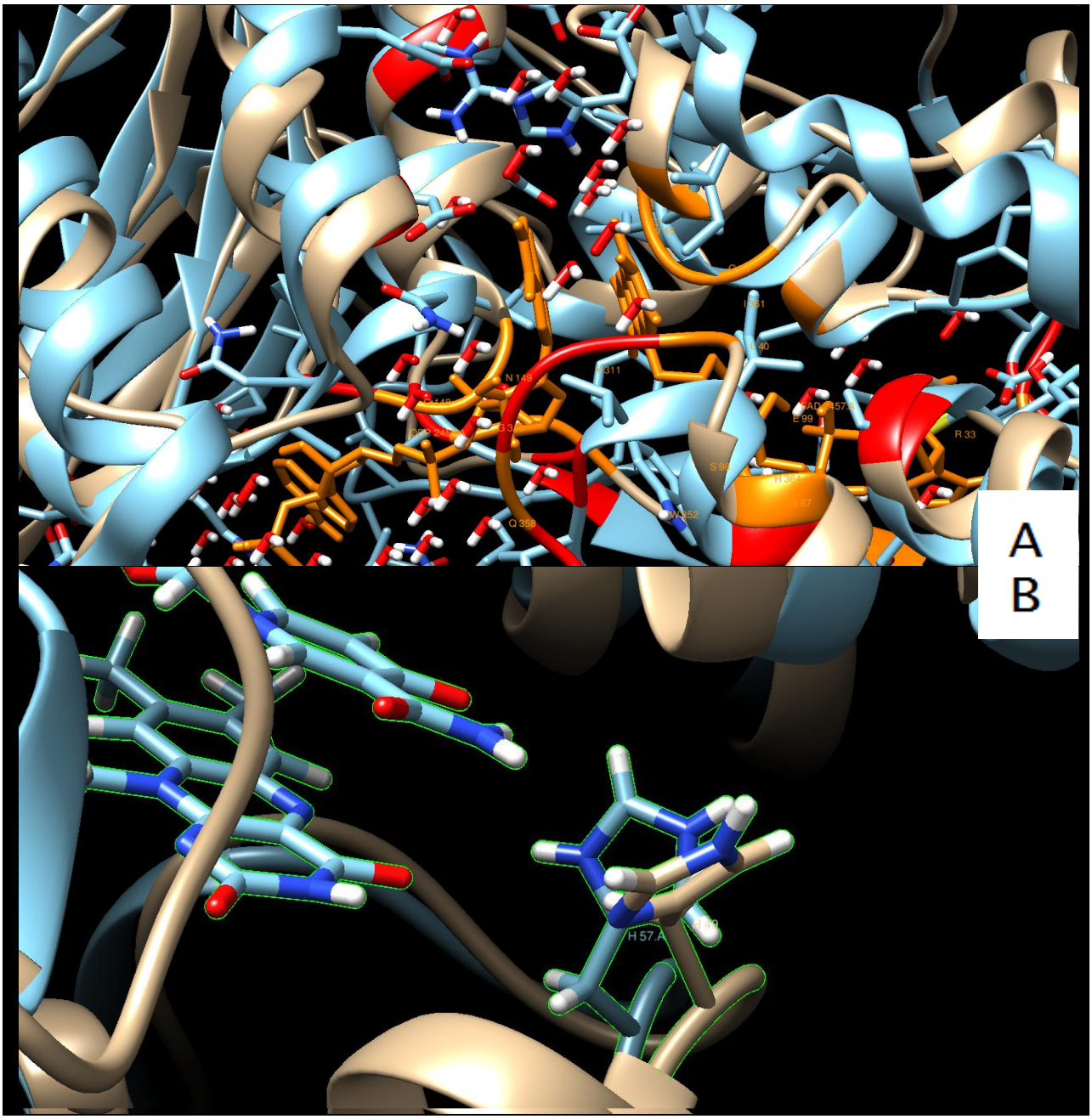
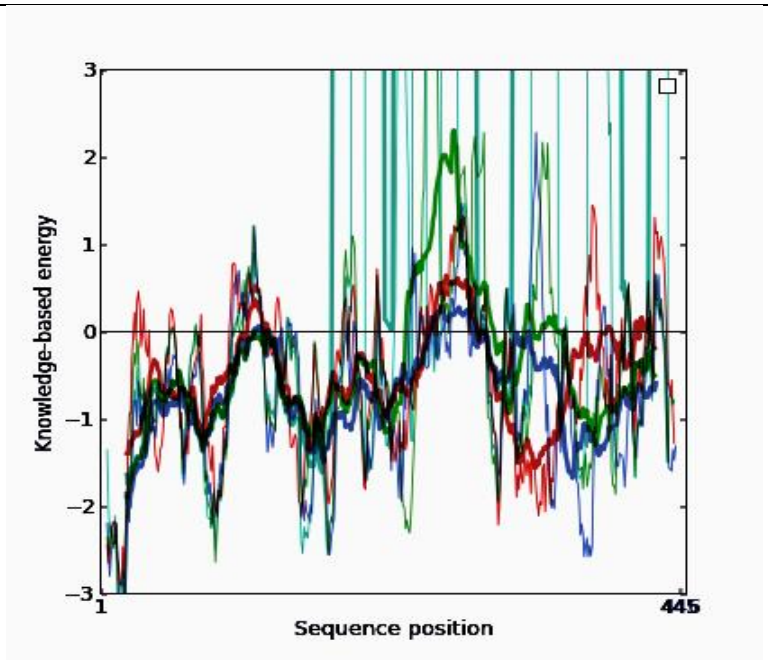
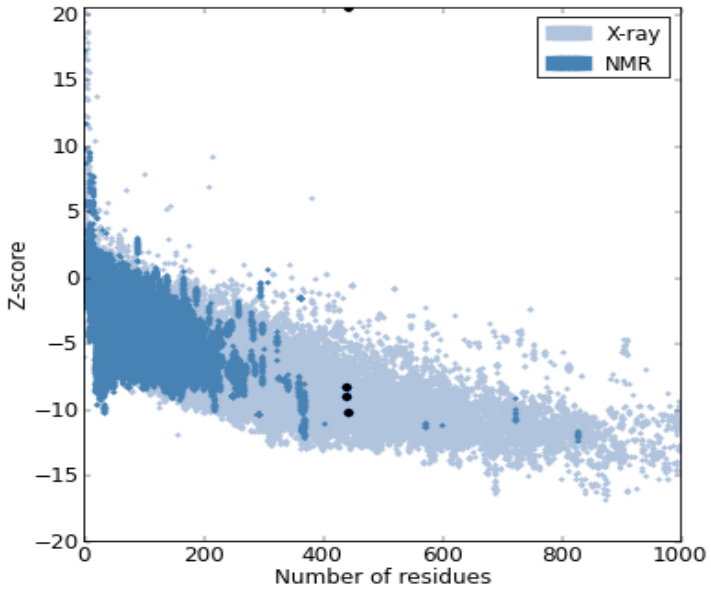


Figure 7. Super position of Insight Model with template 1LQT_A with ligands and water molecules shown. **A.** Water molecules are nearby protein active site which is important for quinine *FprA*:NADPO moiety formation. **B.** Close up of template His-57 residue and aligned target His-50 in close proximity of NADP(H) molecule.



A



B

Figure 8. ProSA result of all 3 models: I,P,R and Thr. **A.** Plot of Residue vs. Knowledge-based energies. Green = R, Cyan = Thr, Red = I, Dark blue = P. **B.** Plot of 4 models against ProSA database of high resolution models.

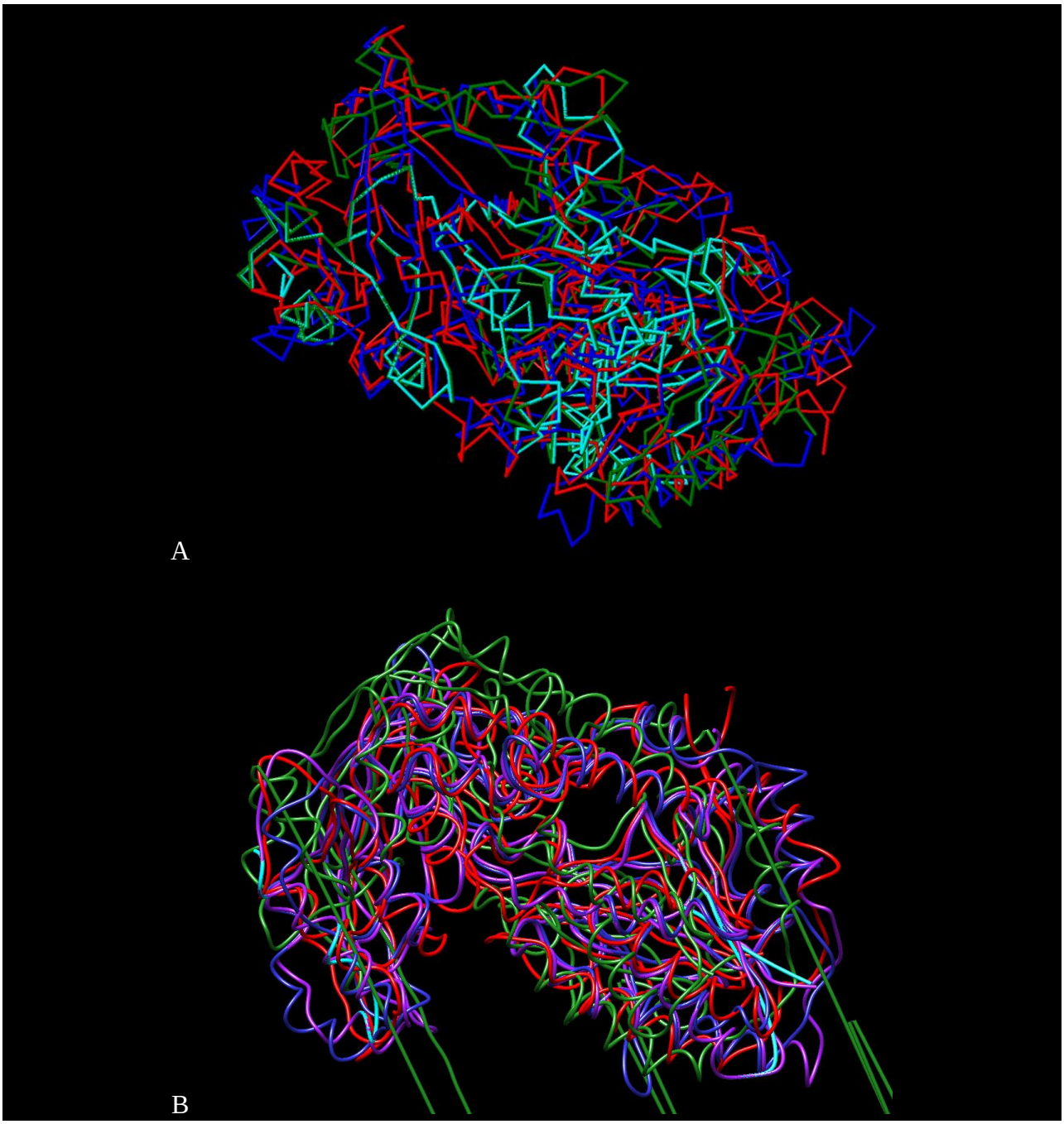


Figure 9. MATRAS Structure Alignment of models and template. Green = R, Cyan = Thr, Red = I, Dark blue = P, purple = Tem. **A.** Structural alignment of models. **B.** Structure alignment of models and target.

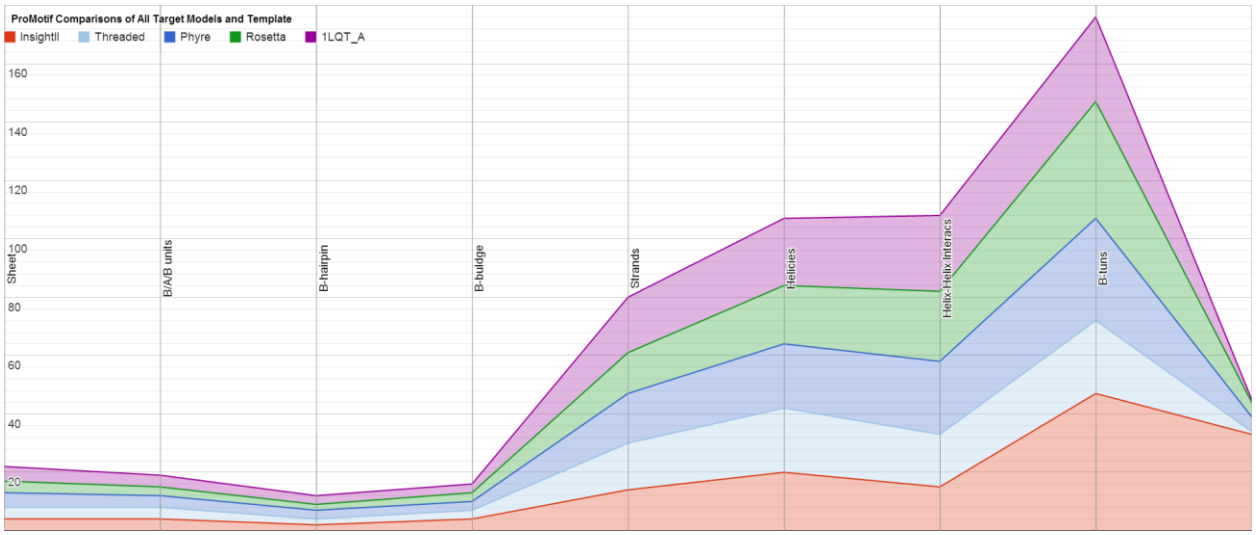


Figure 10. Number of observed Secondary structures within modeled 466500 predictions and template. Green = R, Cyan = Thr, Red = I, Dark blue = P, purple = Tem

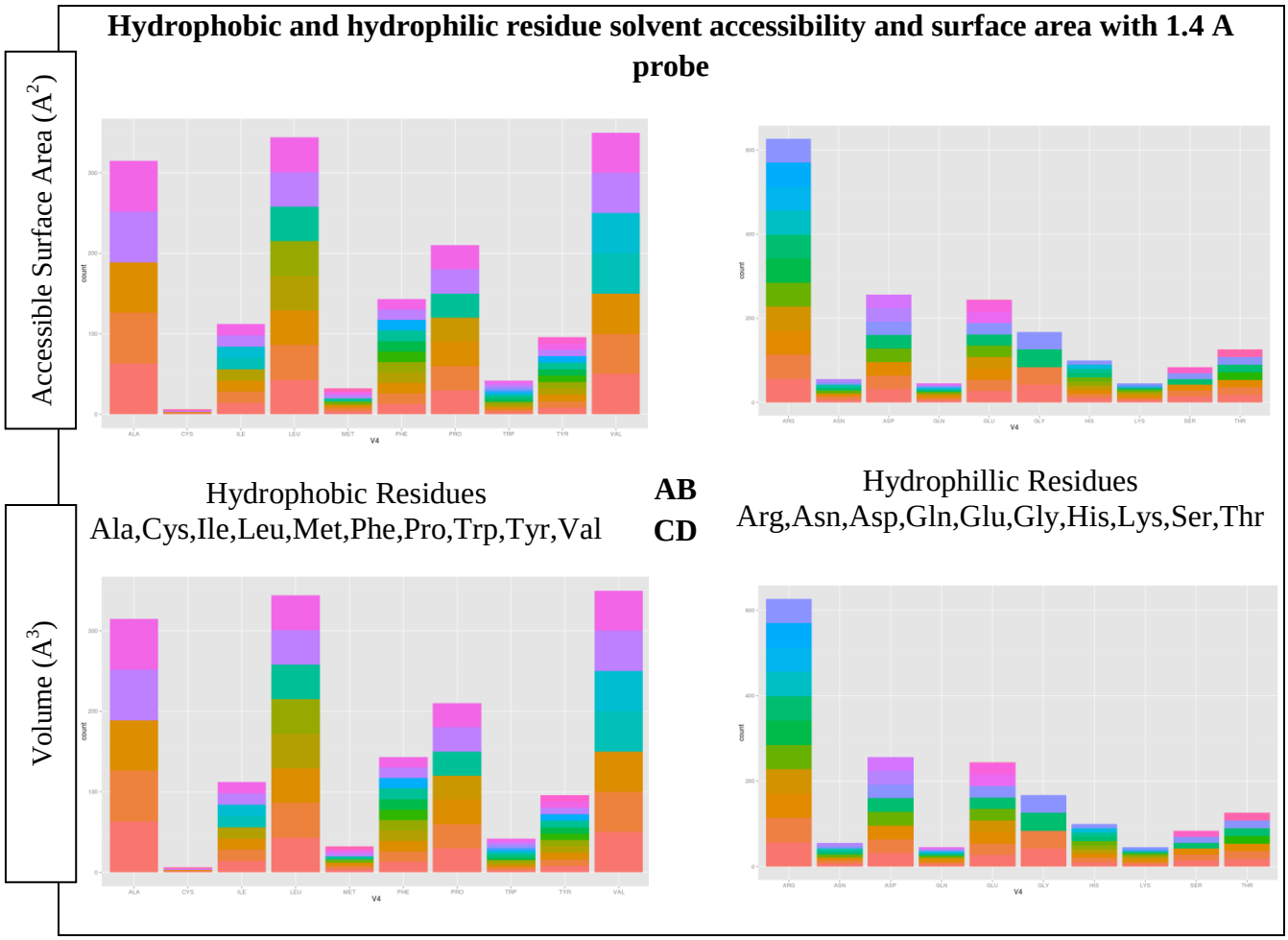


Figure 11. Using Insight Model, solvent accessibility separating hydrophobic and hydrophilic residues was conducted using a 1.4 A surface area probe and volume (A^3). **A.** Surface Area for hydrophobic residues. **B.** Surface area for polar and charged residues. **C.** Volume for Hydrophobic Residues. **D.** Volume for polar and charged residues.

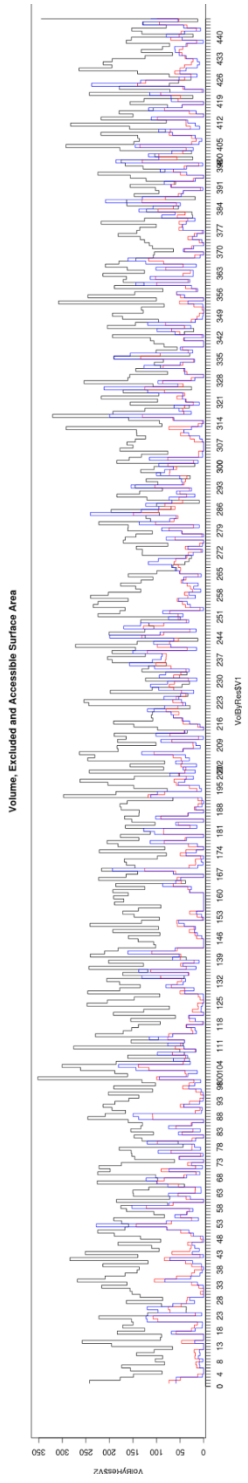


Figure 12. Volume, Excluded and Accessible Surface Area: The Packing Density in Proteins, Standard Radii and Volumes using Insight II model.

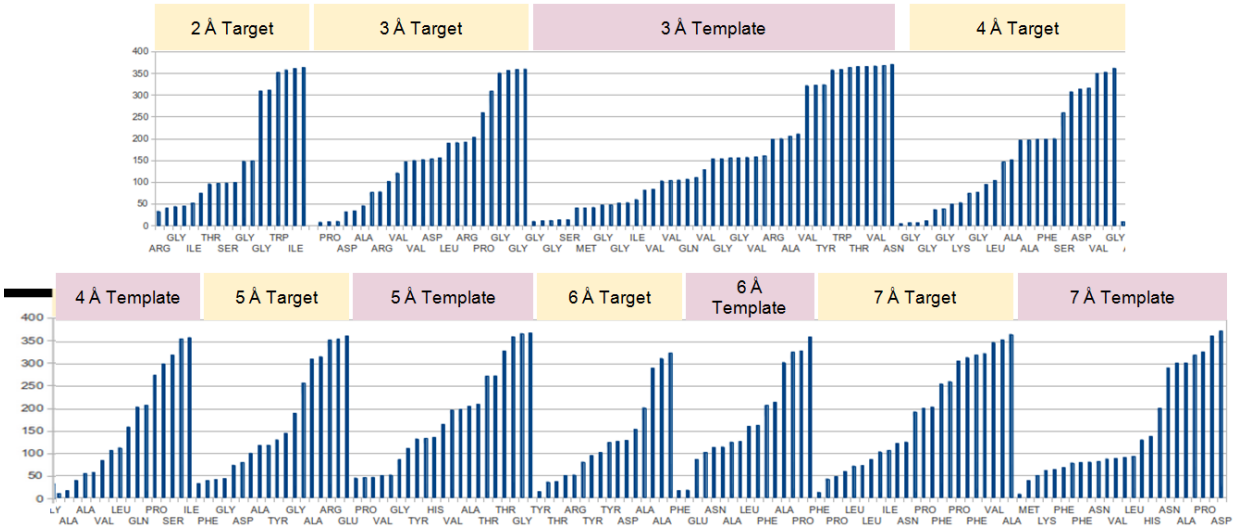


Figure 13. Active site residues found following superposition of Insight model and template 1LQT_A. UCSF Chimera was used to select both ligands, FAD and NADPH then a successive 1 to 7 Angstrom distance probe was used to select all putative active site residues between the target and template.

TABLES

Table 1. Analysis of potential templates discovered using PSI-BLAST vs. PDB database.

Template PDB ID	Source	Chains	Gene Names	Total Score	Query Coverage	E value	%Identity	Resolution (A)
1LQT_A	*Mycobacterium Tuberculosis* H37Rv strain	*A,B*	*fprA, Rv3106, MT3189, MTCY164.16*	384	0.99	1E-118	0.43	1.05, 1.25
2C7G_A	*Mycobacterium Tuberculosis*	*A*	*fprA, Rv3106, MT3189, MTCY164.16*	383	0.99	4E-118	0.43	1.8
1CJC_A	*Bos taurus*	*A*	*FDXR, ADXR*	351	0.98	2E-106	0.44	1.7
2VDC_G	*Azospirillum brasilense*	*A, B, C, D, E, F*	*gltB*	307	0.94	3E-59	0.19	9.5
1H7W_A	*Sus scrofa*	*A, B, C, D*	*DPYD*	187	0.97	6E-31	0.16	1.9

Template PDB ID	Identities (residues, %)	Positives (residues, %)	Gaps (residues, %)	Sequence Length	E. C. Number
1LQT_A	197	42.9	252	459	1.18.1.2
2C7G_A	196	42.7	252	459	1.18.1.2
1CJC_A	200	43.6	243	459	1.18.1.6
2VDC_G	66	19.1	97	346	1.4.1.13
1H7W_A	58	15.9	97	364	1.3.1.2

Table 2. Comparison of genomes within target clade from Figure 3A.

Organism	***Anaeromyxobacter dehalogenans 2CP-2***	***Anaeromyxobacter dehalogenans 2CP-1***	***Sideroxydans lithotrophicus ES-1***
PHYLUM	PROTEOBACTERIA-DELTA	PROTEOBACTERIA-DELTA	PROTEOBACTERIA-BETA
Goldstamp	Gc00340	Gc00932	Gc01345
Size	5013	5029	3004
G+C	75	75	58
OXYGEN REQUIREMENT	Anaerobe	Anaerobe	Aerobe
CELL SHAPE	Rod	Rod	
MOTILITY	Motile	Motile	
SPORULATION	Yes	Yes	
TEMPERATURE RANGE	Mesophile	Mesophile	
HABITAT	Soil, Terrestrial	Soil, Terrestrial	Fresh water
ENERGY	Heterotroph	Heterotroph	Iron oxidizer
	Non-Pathogen	Non-Pathogen	Chemolithoautotroph

Mycobacterium tuberculosis H37Rv	*Sphingomonas sp. LH128*	*Kordiimonas gwangyangensis DSM 19435*
ACTINOBACTERIA	PROTEOBACTERIA-ALPHA	PROTEOBACTERIA-ALPHA
Gi19905	Gi20709	Gi11583
4352	6458	4082
65	65	58
		Obligate aerobe
Rod		Rod
Non-Motile		Motile
No		
Mesophile		
Host		Aquatic, Sediment, Marine

Table 3. References for all validating methods used for the analysis of all 4 models.

Method	Validating Method/Technique	Source	References
PROCHECK	Stereochemical Quality	http://www.doe-mbi.ucla.edu/Software/PROCHECK.html	70
Verify3D	Biophysical structural class, packing	http://nihserver.mbi.ucla.edu/Verify_3D/	71
Errat	Quality of non-bonded atomic interactions	http://nihserver.mbi.ucla.edu/ERRAT/	72
PROVE	Validation using atomic Voronoi volumes	http://www.doe-mbi.ucla.edu/Software/PROVE.html	73
ProSA	Conformation and sequence energies	https://prosa.services.came.sbg.ac.at/prosa.php	74

Table 4. Domain prediction results for template using primary sequence. (50-61)

Domain Existence	Start	Stop	Length	Analysis	Signature	E-Value	Description
FAD	1	209	208	SUPERFAMILY	SSF51971	0	Nucleotide-binding domain superfamily
FAD/NADPH	1	445	444	PANTHER	PTHR11938:SF4	0	Ferredoxin/Ferredoxin--NADP reductase-related
FAD/NADPH	1	445	444	PANTHER	PTHR11938	0	FAD NADPH dehydrogenase/oxidoreductase
NADP(H)	250	312	62	SUPERFAMILY	SSF51971	0	Nucleotide-binding domain superfamily
FAD/NADPH	1	445	444	PIRSF	PIRSF000362	4.80E-183	Adrenodoxin-NADP+ reductase
FAD/NADPH	1	444	443	CDD	PLN02852	2.25E-121	ferredoxin-NADP+ reductase
FAD	2	92	90	Gene3D	G3DSA:3.40.50.720	1.70E-71	NAD(P)-binding Rossmann-like Domain
NADP(H)	312	441	129	Gene3D	G3DSA:3.40.50.720	1.70E-71	NAD(P)-binding Rossmann-like Domain
FAD	93	311	218	Gene3D	G3DSA:3.50.50.60	2.40E-66	Homologous Superfamily to FAD/NAD(P)-binding domain
FAD	2	24	22	PRINTS	PR00419	1.10E-21	Adrenodoxin reductase family signature
FAD	28	41	13	PRINTS	PR00419	1.10E-21	Adrenodoxin reductase family signature
FAD	71	81	10	PRINTS	PR00419	1.10E-21	Adrenodoxin reductase family signature

FAD	143	157	14	PRINTS	PR00419	1.10E-21	Adrenodoxin reductase family signature
FAD	2	387	385	PFAM	Pyr_redox_2	1.90E-06	Pyridine nucleotide-disulphide oxidoreductase
FAD	2	310	308	Pfam	PF07992	1.90E-06	Pyridine nucleotide-disulphide oxidoreductase
FAD	4	54	50	CDD	cl17500	1.59E-04	NAD(P)-binding Rossmann-like domain

Table 5. Predicted ISIS: interaction sites identified from sequence (64)

Domain	Target	Template
FAD	R-22	A-25
	H-50	.-57
NADPH	R-132	E-139
	R-135	S-142
	S-182	R-189
	D-228	.-235
	P-248	.-261
	R-284, R-285	S-295, D-296
	D-315	V-326
FAD	K-325	Q-336
	D-333	G-344
	R-366	K-371
	T-385, R-386, P-387	C-394, K-395, S-396
	R-425	H-434

Table 6. Hydrogen bond formation between residue and FAD molecule and its aligned consensus with target model serves as anchor points for protein secondary structure prediction. (46-47)

Target	Template	Consensus
.; ., ., A-11	G-10; G-12, P-13, S-14	A
D-33, R-34	E-40, M-41	., .
G-40, L-41	G-47, L-48	., .
G-45, V-46	G-52, V-53	., .
V-75, L-77	V-82; V-84	.; L
A-96, T-97, G-98	A-103, V-104, G-105	., T, .
V-151	V-158	.
Y-313	Y-323	.
G-352, W353	G-357, W358	., .
G-360, L-361, I-362	G-365, V-366, I-367	., [LV], .
H-365	N-370	[HN]

Table 7. Assigned Loops for Rosetta Model

Start	Stop	Weak alignment identity	Gaps	Predicted region of mixed/ random coils	Secondary Structure prediction discrepancy	Comment (SS numbering with respect to threaded model)
2	6		x			3 residue gap, start of Threaded_B1 unclear
27	34		x			4 residue gap, start of Threaded_B2 unclear
64	80	x		x		Low SS prediction confidence
113	117	x		x	x	Template shows beta sheet (6) before A5
132	136	x		x		Site of 2 ISIS (132, 135)
161	165	x		x		Low SS prediction although a7 from template is observed in threaded target
180	183	x		x		Site of ISIS S-182
201	203	x		x		Unclear where a9 starts
229	234	x	x	x		unclear where threaded_B8 starts
245	250	x	x		x	Site of ISIS P-248; template
267	271	x	x			Start of threaded B11 unclear
284	293	x		x		ISIS R-284,R-285, unclear where B11 ends
315	318	x				Site of ISIS D-315
332	342	x	x	x	x	D-333, target region shows B16,B17, unclear in prediction
383	391	x				Site of 3 ISIS residues: T-383, R-384, P-385
423	425	x		x	x	Template shows 310 helix following a15, this is not observed within target; ISIS site R-425

Table 8. Promotif, PDBSUM, Verify3d, Procheck, Errat, Prove, ProSA analysis of all secondary structures found within all model (70-73).

ProMotif									
Model	**Sheet**	**β/α/β units**	**β-hairpin**	**β-buldge**	**Strands**	**Helicies**	**Helix-Helix Interacs**	**β-tuns**	**Gamma-turns**
InsightII	4	4	2	4	14	20	15	47	33
Threaded	4	4	2	3	16	22	18	25	1
Phyre	5	4	3	3	17	22	25	35	5
Rosetta	4	3	2	3	14	20	24	40	5
1LQT_A	5	4	3	3	19	23	26	29	1

Model	**PDBSUM ID**	**Template**	**Verify_3D**
			of the residues had an averaged 3D-1D score > 0.2
InsightII	f321	1LQT_A	0.9462
Threaded	f656	1LQT_A	0.6368
Phyre	f654	Multiple	0.9396
Rosetta	f655	Multiple	0.9238
1LQT_A	1lqt	1LQT_A	0.9956

Model	**Ramachandran plot**				
	Core	**Allow**	**Gener**	**Disallow**	**Residue properties: Max.deviation:**
InsightII	0.592	0.34	0.056	0.014	18.9
Threaded	0.926	0.06	0.011	0.006	14.3
Phyre	0.898	0.08	0.003	0.016	19.1
Rosetta	0.841	0.12	0.025	0.014	10.5
1LQT_A	0.939	0.06	0.005	0	4.3

Model	**Errat**		**Prove**					**ProSA**
	Pass /Fail	**Score**	**Z-score mean**	**Z-Score-RMS**	**% Outliers**	**Bond len/angle:**	**Bad Contacts**	**Z-Score**
InsightII	Pass	64.18	1.358	2.037	13.6	5.3	0	-8.14
Threaded			N/A / Fail			5.3	123	20.44
Phyre	Pass	77.85	1.359	36.088	6.5	13.8	10	-9.05
Rosetta	Pass	Failed	0.878	26.958	7.5	4055.1	36	-7.6
1LQT_A	Pass	92.05	0.799	25.162	2.8	5.3	1	-10.57

Table 9. TM-Score and Chimera Match-Maker Superposition of all 4 Models. (74,75)

Model	TM-SCORE					Chimera MatchMaker (using alignment)	
	Score	RMSD of common residues	Common Residues	RMSD	Superposition in the TM-score: Length(d<5.0)=	RMSD	Atom Pairs
InsightII	0.2792	15.81	445	3.7	33	1.281	180
Threaded	0.2758	16.024	445	3.58	20	0	431
Phyre	0.3261	14.488	446	2.89	34	0.642	361
Rosetta	0.2838	156.453	445	3.5	23	0.197	394

1. INTRODUCTION

Organohalide-respiring bacteria (OHRB) living in oxygen-depleted and sometimes toxic environments must adapt and respond to rapid changing redox with substrate-availability conditions due to molecular turnover and do so through a diverse metabolic and ecogenomic toolkit (2,3). *Anaeromyxobacter dehalogenans* spp. are the first anaerobic bacteria that group with the strictly aerobic order Myxococcales according to 16S rRNA gene phylogeny and are usually found in aquatic submerged-soils that are oxygen depleted as microaerophiles whereas all known *Myxobacteria* are strictly aerobic organotrophs (1-3). Unlike the order *Myxococcales*, *A. dehalogenans* spp. do not exhibit fruiting body development or secondary metabolite production but utilize gliding motility and sporulation (1, 2). Metabolic versatility through a wide spectrum of electron acceptors is another recognizable factor for the OHRB *A. dehalogenans* spp. (1,2).

The recently sequenced genome of A. dehalogenans strain 2CP-C by Thomas et al. (2008), reveals a genome architecture with high G+C content of 74.9% and asymmetrical lagging and leading strand in the circular chromosome that indicate a large gene deletion event of 1.5 Mb which may explain the mosaic features between strain 2CP-C and the order Myxococcales (1, 2). High G+C content is an indicator of recent gene acquisitions and little phylogenetic relationship between aerobic organisms in the form of monophyletic clades signify multiple and independent acquisition of aerobic metabolism (1,2,34). Comparative genomics of *A. dehalogenans* strain 2CP-C against delta-Proteobacteria and *Myxobacteria* reveal potential multiple ancient horizontal gene transfer events from a most probable facultative aerobic ancient ancestor that led to strain 2CP-C aerobic and anaerobic metabolic versatility and in the end allowed an advantage

over changing nutrient availability consistent with top and marine sediment soil ecological niches (1,2,34,35).

Flow of electrons within a metabolic network depends on variety of electron donors, acceptors and their catalysts; consequently an overall metabolic flexibility is found within the prokaryotes (1,6). One such indicator of metabolic flexibility is the abundance of *c*-type cytochromes as their function allows storage of electrons inside specific transition metal bounded heme centers that function as biological capacitors during times of nutrient deficiency; accordingly, strain 2CP-C genome contains 68 putative *c*-type cytochrome genes with 83.8% containing multiple heme binding motifs along with a 40-heme cytochrome rare anomaly and suggests a strong necessity for metabolic enzymes (2,6). Moreover, breadth of Ni-Fe and Fe-S clusters, indicators of versatile metabolism, are found in *A. dehalogenans* strain 2CP-C (2).

Overall, *A. dehalogenans* spp. are facultative aerobes that are able to derive energy through organohalide respiration: halogenated organic or aliphatic compounds such as acetate or hydrogen are used as electron donors, aromatic pollutants such as chlorinated phenols as electron acceptors (chlororespiration) and in general hydrocarbons as terminal electron acceptors (2, 3, 5). Bioinformatics studies on the *A. dehalogenans* strain 2CP-C proteome reveal seminal OHRB putative genes for two reductive dehalogenase genes for degradation of halogenated compounds , haloalkane dehydrogenase and haloacid dehalogenase that subgroup *A. dehalogenans* strain 2CP-C as non-obligate organohalide respiring bacteria (2,5). Furthermore, according to Thomas et al. 2008, the derivation of the versatile ecogenomic toolkit of anaerobic *A. dehalogenans* strain 2CP-C may have likely been through convergent evolution of with an aerobic ancestor (2).

2. MATERIALS & METHODS

2.1. SEQUENCE AND DOMAIN ANALYSIS

In this study the target FAD-dependent pyridine nucleotide-disulfide oxidoreductase (FDPNDO) protein sequence from *Anaeromyxobacter dehalogenans* 2CP-C has been retrieved from GenBank database of NCBI (GenBank accession no: YP_466500) and is henceforth also known as the 'target'. Putative conserved domain of FPNDO was computed using **SSDB** (http://www.kegg.jp/kegg/ssdb/) (50), **SCOP** (http://scop.mrc-lmb.cam.ac.uk/scop/) (51), **CATH** (www.cathdb.info) (52), **SMART** (http://smart.embl-heidelberg.de) (53), and **CDD** (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) (54). **InterPro** (http://www.ebi.ac.uk/Tools/pfa/iprscan) (55) and **Pfam** (http://pfam.sanger.ac.uk) (56) were used to predict protein domain architecture and correlated superfamilies. **ProKnow 2.0** (http://services.mbi.ucla.edu/proknow) (57) and **EzyPred** (http://www.csbio.sjtu.edu.cn/cgi-bin/EzyPred.cgi) (58) were used to make inferences on protein function from primary sequence.

2.2. PRIMARY STRUCTURE ANALYSIS

A primary biophysical and biochemical study on FDPNDO molecular weight, amino acid composition, grand average of hydropathicity (GRAVY) was performed using **ProtParam** by Expasy tool (http://web.expasy.org/cgi-bin/protparam) (59). Dipeptide compostion contingency table was constructed using R statistical package **seqinr** (http://cran.r-project.org/web/packages/seqinr) (60) while various plotting was done using R package **ggplot2** (62).

2.3. SECONDARY STRUCTURE ANALYSIS

Secondary structure of FDPNDO was predicted using **CONCORD** server (http://helios.princeton.edu/CONCORD/) that aggregates several prediction methods such as **PSIPRED**, **DSC**, **GOR IV**, **Predator**, **Prof**, **PROFphd**, and **SSpro** and generates a prediction output using a consensus method (61). Percentage of secondary structure conformation is calculated by counting type of secondary structure element character as defined by **STRIDE** and **DSSP** (63). Binding site prediction was predicted using Interaction Sites Identified from Sequence (**ISIS**); furthermore, these results were also used to identify residues of importance when determining Structurally Conserved Regions (SRCs) and loop regions (64).

2.4. COMPARATIVE MODELING OF PROTEIN

Construction of FAD-dependent pyridine nucleotide-disulfide oxidoreductase model involved five standard steps that are adapted from protocols and are outlined in Figure 1 (20-26,30-33).

Template selection through strict multiple sequence alignment

After selecting a template, alignment between target and template was identified using a two-step profile-profile alignment method. First, a stringent DELTA-BLAST using BLOSUM 90 matrix was performed against non-redundant and top 500 matches were aligned using MAAFT and another DELTA-BLAST was performed against PDB and the best scoring match was chosen as the template (1LQT chain A). Second, two alignments were aligned together and template, target and top 5 alignments based on identity were picked out. Target protein was also scanned on multiple meta servers for best template with the highest identity score like NCBI related structures (http://structure.ncbi.nlm.nih.gov), **Geno3D** (http://geno3d-pbil.ibcp.fr). The two

alignments were re-aligned using MUSCLE and the alignment between the target and query was manually adjusted to accommodate higher C-terminal identity (SF1).

Software used for modeling

Generating a 3D homologous protein structure from an alignment can be done through various software. For this paper, four final models were generated using: proprietary Accelerys Insight II, open source Rosetta and Phyre web server.

Accelrys Insight II (http://accelrys.com/), is a molecular modeling graphical interface which contains Discover: an environment for simulation of molecular mechanics including energy minimization and dynamics. Insight II protocol involves identification of template structure and generating an alignment, assigning coordinates to structurally conserved regions, constructing loop conformations, refining structure through Discover energy minimization and resolving the final structure (23). **Rosetta** (https://www.rosettacommons.org/), developed by the Meiler Lab, is used for comparative modeling which aims to resolve tertiary structure fold architecture from primary structure of another protein that may or may not be homologous (20). Rosetta protocol includes a primary sequence alignment between template/target then subsequent threading of target onto template backbone 3D coordinates and then loop building using a multiple template fragment library (20). Similar loop building approach is used by Protein Homology/analogY Recognition Engine V 2.0 or **Phyre2** Webserver (www.sbg.bio.ic.ac.uk/phyre2/) which offers a final resolved model with only the inclusion of template target sequence, and an email address (21). Phyre2 uses the top ten highest PDB matches to the target to construct a ‘fold’ library from multiple structures (21).

Alignment of 3D coordinates to template and loop assignment

Within Insight II, coordinates were assigned to Structurally Conserved Regions (SCRs) between the 3D template and aligned target and with this approach, identical residues that occur in the alignment are assigned identical coordinates while non-identical regions are assigned identical backbone coordinates and side chains are calculated within the program. Regions for *de novo* loop assignments with Rosetta were manually inspected regions that contain the following characteristics as outlined by Combs et al. 2013: regions of mixed/random coils and also weak alignment identity and/or secondary structure discrepancies (20). Furthermore, areas of structural importance such as conserved active site residues were especially kept out of loop regions. Loop fragment libraries of 3-mer and 9-mer that contain sequence, Cartesian coordinate and secondary structure information were generated by web server **Robetta** (http://robetta.bakerlab.org) and along with the loop definitions are used by the loop modeling cyclic coordinate descent (CCD) algorithm (20, 48).

2.5. ENERGY MINIMIZATION

After identifying Structurally Conserved Regions (SCRs) within FDPNDO and constructing a rough model using Insight II, this model was then subjected to energy minimization. Hydrogen molecules were added to lone pairs to all molecules, with the capping mode on and pH set to 7 (23-25). Potentials from `Forcefield` are set to the potential function for a molecule. Discover 3.0 Molecular Simulation: Energy Minimization using `Simple_Minimize` with 100,000 iterations and derivative set to 0.01 (23-25). An all-atom refinement of Rosetta generated model following CCD is then refined through the '`relax`' script which aims to energetically minimize the model by iterative side-chain repacking through selection from rotamer libraries for an

overall side chain optimization then a gradient-based minimization of side chain and whole model (20,26).

2.6. VALIDATION OF THE MODEL

Model evaluation is an integral concluding step to homology modeling. Each residue within the protein model can be associated to two conformational angles: angle of rotation around N-Cα is φ (phi) while angle of rotation around Cα-C' is ψ (psi) and both of these values determine the conformation of the main-chain atoms as are the only degrees of freedom around the rigid peptide bond and for that reason, an overall model which is accurate avoids steric collisions between the side chain and the main chains (36). A Ramachandran plot plots φ and ψ angles and shows allowed and disallowed conformational regions where favorable regions correspond to secondary structures: right/left hand helices and strands cluster in specific regions of the plot as tighter clustering is representative of an energetically favorable model (36). Overall model quality can be derived from the plot by figuring out the number of modeled residues that fall within the most favorable or core regions. The Structure Analysis and Verification Server or **SAVES** (http://services.mbi.ucla.edu/SAVES/) comprises of is a metaserver that provides the analysis of protein structure through the following programs: **Promotif**, **PDBSUM**, **Verify3D**, **Procheck**, **Errat**, **Prove** (49, 70-73)

2.7. MODEL SUPERPOSITION FOR LIGAND-RESIDUE INTERACTIONS ANALYSIS

Evaluation of the model properties, similarities and atomic deviations against the template protein structure were important to this research. For ligand active site residue interaction

studies, **MatchMaker** tool within **UCSF Chimera** (https://www.cgl.ucsf.edu/chimera/), a GUI of an extensible molecular modeling and visualization software, was used to superimpose target model against the template using the original pair wise alignment (Figure 2) and iterated by pruning long atom pairs until no pair exceeds 2.0 angstroms, provided alignment was used (43). **Insight II** was used for rechecking the model for defects by the `Struct_Check` function under `ProStat` which allowed regions of the model to be colored that exceeded 3 standard deviation of Root Mean Square Deviation or RMSD units, which compares two vectors of atoms results in a final distance between conformations (23-25). Similar but perhaps a better alternative was also used. The template-modeling score (TM-score) measures overall topology difference between a target and template and higher scores signify better structural matches (75).

2.8. COMPARISON OF GENERATED MODELS

A multiple 3-D alignment of the four models generated was constructed using **MA**rkovian **TRA**nsition of Structure evolution or **MATRAS** (http://strcomp.protein.osaka-u.ac.jp/matras/) (44). Feng et al. 2010 compiled **Potentials 'R' Us**, a collection of four-body potentials, short-range potentials along with 23 two-body potentials calculated energies inputted to the **Knowledge-based potential server for proteins** (http://gor.bb.iastate.edu/potential/), their compiled comparisons reveal overall energies to the generated target and template models (27).

3. RESULTS

3.1. SEQUENCE ANALYSIS OF FDPNDO

The primary sequence of FDPNDO (446 residues) as retrieved from GenBank (Accession no: YP_466500), is encoded by gene *Adeh_3296* (Gene ID: 3889389) with a gene description of FAD-dependent pyridine nucleotide-disulfide oxidoreductase (FDPNDO), showed this protein is basic (theoretical pI: 9.08) with a molecular mass of 48,180 Dalton (48.18 kDa) (59). Aliphatic index is an indicator of thermostability of a protein as it calculates the relative volume occupied by four aliphatic amino acids (alanine, leucine, isoleucine and valine) (60). The aliphatic index of FDPNDO is 96.48 and is a strong indicator of protein stability.

Although *A. dehalogenans* strain 2CP-C isn't thermophillic, it contains a high G+C content that is found within thermophilic organism. It was discovered that AI of thermophilic bacteria is significantly higher than of ordinary proteins (60). Moreover, Gromiha et al. (1999) found that preferred specific exchanges among amino acid residues exist in thermophilic bacteria: Gly→Ala, Ser→Thr, Lys→Arg, Asp→Glu, Met→Ala | Leu, Cys→Ile, Ala, Val, and Trp→Tyr and these exchanges increase thermostability of proteins (12). Assessment of FDPNDO amino composition shows the top 5 amino acids: alanine (14.1%), arginine (12.8%), valine (11.2%), glycine (9.4%) and leucine (9.6%) and bottom 5 amino acids: cysteine (.22%), tryptophan (.67%), methionine (.89%), lysine (1.12%), glutamine (1.12%) suggests a likely indication of preferred amino acid substitution to confer higher protein stability even though *A. dehalogenans* 2CP-C is mesophilic (12,60).

For single amino acid composition, influence of arginine is positive in archaeal and negative in bacterial proteins (13). In FDPNDO, 4-mer with minimum frequency of 2 are: ADAA, AVDV, DAAG, TEVL, VARG, VDVA and VVDP which suggest a very short sequence motif of placing charged arginine or aspartic acid between small, nonpolar residues and within the model this is highlighted in Figure 4 (60). The Grand average of hydropathicity (GRAVY) is calculated by taking the sum of hydropathy values of all amino acids and divided by total amino acids where a more positive final value correlates with hydrophobicity (12). The GRAVY index of FDPNDO is negative (-0.139), indicating a hydrophilic interaction with water. CONCORD secondary structure prediction shows that FDPNDO is half random coils (50.45%) with the rest helical (29.82%) and strand (19.73%).

3.2. TEMPLATE SELECTION THROUGH MSA

Precise sequence alignment between target and template that contain structurally conserved regions (SCRs) are a strong basis for finding a template. Potential templates were identified using DELTA-BLAST against PDB and reveal conservative function across bacteria and eukaryotic organisms between three structures with 43%-44% sequence identity (Table 1). Against these three structures, further analysis of number of gaps and E-value resulted in choosing *Mycobacterium Tuberculosis* oxidoreductase chain A (PDB ID: 1LQT_A) as a template to build the comparative model.

3.3. TEMPLATE AND FPRA GENE DISCOVERY

A known *FprA* protein that joins the structural family of glutathione reductase, template 1LQT chain A is a 50 kDa dimeric flavoenzyme that uses its NADPH cofactor to store and transfer two

reducing equivalents from NADPH to metabolic partners and belongs to E.C 1.18.1.2 (9). Small, monomeric, hydrophilic and ubiquitous, Ferredoxin-NADP(H) reductase (FNR) protein families are classified under E.C. number of 1.18.1.2 and contain a FAD prosthetic group that is used to catalyze electron transfer between NADP(H) and a proton acceptor binding partner either the smaller ferredoxin (Fd), more versatile flavodoxin (Fld) (14-16). Flavodoxins (Fld) use a flavin mononucleotide (FMN) as a redox cofactor and their known features include a highly acidic nature composed of around 160 to 180 residues and are found mostly in bacteria with the exception of some red and green algae while ferredoxins (Fd) are multifunctional electron carrier proteins that use an iron sulfur compound as a cofactor and generally react with FNR (14,15,17). It is important to note that under low iron conditions flavodoxins can be substituted ferredoxins, an important flexibility mechanism for organisms such as *A. dehalogenans* 2CP-C which must adapt to rapidly changing redox conditions (14).

Accordingly, FNR protein architecture consists of two distinct domains: N-terminus binding cofactor FAD and C-terminus cavity to bind NADP(H) (14). A close structural homologs to template $FprA_{1LQT}$ is bovine adrenodoxin (AR), which share the following conserved catalytic residues: Lys24, Arg213, Lys246, Arg362, and Asp374 (9). Furthermore, template $FprA_{1LQT}$ has a theoretical pI of 5.53, about 3 units more acidic than the target, an aliphatic index of 92.81 and GRAVY score of -0.216 (59). Classic Rossman fold topology is exhibited by both the FAD and NADPH binding domains of 1LQT; however, notable deviations from the overall fold similarities between other *FprA* flavoenzymes include replacement of strand 3,4 with an alpha helix within the FAD-binding domain (9). When bound to the NADP-binding domain solvent accessible template cleft where substrate ADP resides, NADP nicotinamide ring becomes

parallel (offset of 9°) to FAD isoalloxazine ring and C4 of NADP+ is 3.27 Å from N5 of FAD in chain A and 3.31 Å in chain B (9).

Most notable feature of 1LQT is the formation of a quinone analog construction complex between $FprA_{1LQT}$:NADPO which is perhaps formed through a catalytic triad mechanism of serine proteases with interacting residues Glu214, His57 and a water molecule (7, 9).

Proposed mechanism NADPO formation in 1LQT by Bossi et al. 2002 involves the following steps. First Glu-214 increases the basicity of His57 through a short-lived tetrahedral transition state the enzyme which can then accept a proton from water forming a positively charged His57 stabilized by Glu214 and hydroxyl ion (7,9). Nucleophilic attack of carbonyl C4 atom of NADP+ by hydroxyl ion which forms the second tetrahedral intermediate that aims to regenerate the carbonyl carbon (7,9). Terminal steps from deacylation to free enzyme include a couple reaction in the C4 of hydroxylated NADPH (C4-OH) moiety: hydride transfer from C4 and H+ release from hydroxyl (7,9). Occurrence of this mechanism had not yet been verified past Bossi et al. 2002 (7,9).

3.4. PHYLOGENOMICS, DOMAIN AND MOTIF ANALYSIS

Phylogenomic analys of FDPNDO primary sequence against top 100 PSI-Blast hits is shown in Figure 3 with *Rhodococcus* spp. as the out group. Overall, tree shows FDPNDO is mostly related among other marine chemolithotrophic bacteria such as *Anaeromyxobacter dehalogenans* 2CP-1, *Sideroxydans lithotrophicus* ES-1, *Kordiimonas gwangyangensis* DSM 19435, *Sphingomonas spp.* LH128 (*see* Table 2). Here it is important to note the variety of reparatory abilities of clade members along with G+C content outside of related strain 2CP-1.

Domain prediction of FDPNDO by Pfam revealed one significant match Pyr_redox_2 (1 to 386) and conserved domain (CD) search revealed 2 overlapping matches, first to super family cl17500 (E-value 1.59e-04) that comprises of Rossman fold-like domain and NAD(P)-binding proteins not found in as multi-domain and second to PLN02852 (E-value 2.25e-121) consisting of ferredoxin-NADP+ reductase proteins that may span more than one domain. Protein structure classification with CATH and GENE-3D shows two matches with FDPNDO to functional families of Ferredoxin reductase -like domain: superfamily 3.40.50.720 (E-value 1.7Ee-71) and 3.50.50.60 (E-value 2.4Ee-66). Top 5 results of Delta-Blast search of FDPNDO against PDB reveals all 5 sequences contain 3.50.50.60 and all but 1H7W contain 3.40.50.720.

Further comparative proteomics study of FDPNDO along with top 5 PDB matches (Table 2) shows all sequences share the same adrenodoxin reductase family signature as determined by PRINTS protein fingerprint database (Prints ID: PR00419), and are classified in SCOP under the nucleotide-binding domain superfamily (SCOP ID: F51971). InterPro domain Pyridine nucleotide-disulphide oxidoreductase, FAD/NAD(P)-binding domain (IPR023753) is shared between target and 2VDC. InterPro NAD(P)-binding domain (E-value IPR016040) is shared between the top four matches, while adrenodoxin-NADP+ reductase domain (IPR021163) is shared between target FDPNDO and both 1LQT, 1CJC which also share the same E.C. number of 1.18.1.2.

Multiple sequence alignment of top 5 DELTABLAST hits and FDPNDO reveals a significant conservation of glycine residues, which conform to FAD-containing protein sequence motif, *see* Supplementary Figure 1 (7). Overall, FDPNDO contains two distinct domains and all these data suggest FDPNDO is putatively classified as the same E.C. number as the template 1LQT.

3.5. COMPARATIVE MODELING AND ENERGY MINIMIZATION

Alignment, protein threading, backbone generation, loop assignment

Sequence alignment of target and template reveal regions of identity and loop exist outside these regions (Table 5). Main areas of concern are where there is secondary structure discrepancy between known and observed template secondary structure and predicted template structure. Low secondary structure prediction confidence occurs at NADP(H) binding domain target residues 115-126, within this region the template conforms to a short B6 (template residues: 121-123; template seq: Ser-Ile-Ala) while the threaded template omits this short beta sheet. Likewise, a short turn n2 (template 148-152; template seq: PHF) is also not observed. Region of low sequence identity following template B15 to B18 which occur at the target of the FAD-binding domain are also of interest. Secondary structure elements following B15: bridge, B16, TT,B17, turn are not observed within the threaded model and 3_{10} helix (n5) immediately following template a15 is not observed within the template. Although A-423 (φ=-55.82 ψ =-42.86) may fall inside the geometry for a 310 helix, R-424(φ =-62.54 ψ =-16.01), R-425(φ=107.22 ψ =12.94) do not based on the threaded model. Futhermore, regions which contained ISIS sites and weak alignment were used for loops regions. Due to short gap regions and also loop regions a reassessment of the initial alignment was not performed as suggested by Krieger et al. 2009 (65).

Comparative Modeling and Energy Minimization

Modeling of all 4 structures is described under Methods and was followed accordingly. Four models were generated and are henceforth abbreviated as follows: **I** = Insight II model, **P** = Phyre model, **R** = Rosetta model, **Thr** = Threaded Rosetta Model while the template is abbreviated as = **Tem**.

MATRAS Alignment

Following MATRAS Structural and Sequence alignment of all models including the template, a careful accession of all models now follows (44).

Following α2 of β/α/β, all but I adapt β2 which instead adapts a 3 residue 3_{10} helix-TT then $I_{\beta 2}$. The 3-residue 3_{10} helix is adapted by Tem,R,T with residues (GLV) however I and P adapt a regular helix here instead ($I_{\alpha 2}$, $P_{\alpha 3}$). Tem adapts a T-3_{10} helix at following $I_{\alpha 2}$ with sequence PKIKSI whereas aligned 466500 sequence QRIKAV is predicted to be a 4 residue helix by I and P (Iα2, Pα3) while R and T follow templates fold. All models adapt temp β/α/'βTTβ'/α/β/β (tem/I/R/$T_{\alpha 3}$, $P_{\alpha 4}$). Structures β/a follow the same length as tem except for $I_{\beta 3}$ which is shorted to the terminal 2 residues LG while the BTTB motif within the tem is found with residues VGEHV and within the models represented with residues LGRDV, which is 7 residue turn followed by tem_{B6}/$tem_{\alpha 5}$. Within the models, this motif is observed except for the 3 residue beta sheet represented by template residues PGEDLPG/SIS and aligned target residues PGEGIER/AVT. $Tem_{\alpha 5}$/tem_{310}/$tem_{\beta 7}$/$tem_{\alpha 6}$ are also found within the models with generally the same alignment. Tema7 adapted by residues PDVL and aligned target residues RAEL is only predicted by $P_{\alpha 8}$ and $Th_{\alpha 7}$ whereas I and R adapt a bend or turn here instead.

Onward, template residue Ala-179 start $tem_{\alpha 8}$ and aligned target residue ala-173 start helix R/$I_{\alpha 7}$, $P_{\alpha 9}$, and $Th_{\alpha 8}$ followed by tem_{B8}, I|P|R|$T_{\alpha 7}$ beta sheet. A consensus 3 residue 3_{10} helix adapted by template residues PLQ and aligned target residues PAQ follows. Template helix $tem_{\alpha 9}$ starts with T-209 whereas within the aligned target, this helix starts with Pro202 for I|$R_{\alpha 8}$, $Th_{\alpha 9}$ while $P_{\alpha 10}$ starts 2 residues later where the prediction is higher with the consensus Glu at template residue 204. Immediately following $tem_{\alpha 9}$ is a 3 residue 3_{10} helix adapted by residues LAD and aligned

target residues LGT and is predicted for all models except for P which predicts $P_{\alpha 7}$ to stretch here instead. Following tem_{A9} a tem 4 residue 3_{10} helix is adapted by tem residues PAEL and aligned target residues PADL. Model I adapts a 4 residue helix ($I_{\alpha 9}$) here instead while a consensus among other models exists with template.

Following template's 3_{10} helix the alignment and prediction becomes generally weak. Template adapts helices $tem_{\alpha 10}$ and $tem_{\alpha 11}$ that are interrupted with a one residue Gly242. Within the primary sequence alignment, there is a 7 residue gap which spans from the last residue in $tem_{\alpha 10}$ to the 5th residue in $tem_{\alpha 11}$. This region is predicted as a mixture of coils/bends by all models either for either the first or last half of aligned $tem_{\alpha 11}$. Template adapts beta sheet tem_{B10}, tem_{B11} with coiled/bend residues LT in the middle and this is only predicted within P model while I adapts a larger 4 residue gap in between. Tem $_{\beta 12/B13/\ \beta\ 14/\ \beta\ 15/\ \beta\ 16/\ \beta\ TTTT/\ \beta\ 17}$ conclude the NADP(H) binding domain and are also found within P model while I and R adapt all but tem $_{\beta\ 17}$. Instead of βTTTT, I adapt a 3 residue 3_{10} helix with the residues KAR. Tem $_{\beta\ 18/\ \beta\ 19/\alpha 12/cScScTT/\alpha 13}$ are generally adapted by P and R models; however, model I adapts a set of turns/bends instead of the aligned beta sheets. $I_{\alpha 11}$, shorter than $tem_{\alpha 13}$, ends with a 3 residue 3_{10} helix with the residues VAD. Template then follows with $tem_{\alpha/14/\beta 20/\alpha 15}$ which is also predicted to be adapted generally with models I,P,R. except for I and P which choose a bend over $tem_{\alpha 20}$. Lastly, C-terminal template helix, $tem_{\alpha 16}$ is conserved within I,P,R models.

3.6. MODEL QUALITY

Methods leading to various criteria are used to check the accuracy of a homology model, and the likelihood of potential errors. Using the methods described in table 8, an atomic and geometrical evaluation of all 4 models was carefully analyzed.

Procheck

Procheck calculates the stereochemical quality of a protein structure by comparing the normalcy of geometry of residues in target model is compared to stereochemical calculations derived from high resolution structures. A Ramachandran plot describes the backbone ϕ, Ψ torsion angles and a good model contains more than 90% residues within the core region as strong models have pronounced clustering of their residues and contain few outliers. Compared to the template model which contains 93.90% core residues, the highest performing model was the threaded model (92.60%) which is otherwise noise since this model is essentially an un-minimized template and failed Errat. Rosetta and Phyre model with core regions 84.10% and 89.80% respectively scored almost above the ~90% cutoff while I model scored the lowest with 59.20% of core residues.

Verify3D

Verify3D assesses the compatibility of a 3D atomic model against its own amino acid sequence where each residue is assigned a structural identified based on its location and biophysical environmental properties. Then a reference collection of good structures is used against the model to obtain a score for each of the 20 amino acids within structural environmental class which above all gives a good indication of structural packing quality. Of the residues which have averaged 3D-1D score greater than 0.2, in order of highest percentage are Insight II model (94.62%), Phyre (93.96%) and Rosetta (92.38%). For comparison, template model scored 99.56% and threaded model scored lowest with 63.68%. A higher score for I and Thr model was expected because of the use of only one template while P and R use fragments of multiple templates.

Errat

An overall quality factor for non-bonded atomic interactions, Errat is defined as a percentage of the amino acid sequence that is errously below the 95% statistical rejection limit for each chain in the input structure. Both Rosetta models (threaded and final structure) have failed errat indicating issues with backbone conformation and non-bonded interactions. An Errat score greater than 50 is considered good and is found within models Phyre (77.854%) and Insight II (64.183%). For comparison, template scored highest with Errat score of 92.045%.

Prove

PROtein Volume Evaluation or PROVE assess model volume-based structure validation by first calculating Voronoi spherical volume of model atoms then comparing against a highly refined and resolved structures within the PDB. For high-resolution structures, the mean Z-RMSD is ~1.0 and the score increases with lower resolution structures. Insight Model provided lowest Z-Score RMS of 2.037 with 13.6% outliers but still exceeded a Z-score and %outliers of ~1.0. Phyre, Rosetta Model and the template scored exceptionally high while threaded Rosetta model failed this test.

ProSA

ProSA provides a quick screening methods for good quality models and does this by evaluating the distance-based pair potential and a solvent exposure residue potential then calculates a Z-score which is reflective of overall model quality. Best models when compared to other high resolution and viable (range of values characteristic of native proteins) structures have lowest Z-scores. Highest and most un-viable structure as predicted was the threaded Rosetta model with a Z-score of 20.44. Other three target models (Insight: -8.14, Phyre: -9.05 and Rosetta: -7.6) and

template models (1LQT_A: -10.57) are clustered. Figure 8 shows the plot of sequence position vs. Knowledge-Based energy in a ProSA assessment of all 3 target models and the template. It is evident that for all 3 models, the energy increases to approach 0 and then climbs more positively generally at two peaks around residue ~110 and ~270: within the alignment these regions represent low sequence identity and gap insertions.

3.7. MODEL SUPERPOSITIONS AND DETAILED STRUCTURAL STUDY

Chimera MatchMaker

To construct a careful structural superposition based on the initial alignment, Chimera MatchMaker was utilized.The highest RMSD model resulted with InsightII (RMSD = 1.281 with 180 atom pairs). Rosetta (RMSD = 0.197 with 394 atom pairs) and Phyre (RMSD = 0.642 with 361 atom pairs) both scored relatively low RMSD values but with a high atom pair alignment. As expected, threaded model results in an RMSD of 0 with 431 atom pairs aligned.

TM-SCORE

TM-Score is a novel approach to determining structural similarity between two protein structures; thereby, if necessary, the accuracy of protein structure predictions. Unlike RMSD calculations, TM-score weighs close atom-pairs with more heavily which proposed to solve local error with RMSD (75). A TM-Score > 0.5 indicates correct topology within the local fold while a TM-Score < 0.17 recommends random similarity (75). All results are outlined in Table 9. Here we see that none of the models using TM-Score, scored above 0.5 and generally all cluster around the range of 0.27-0.32. However, the highest score is with P model (0.3261).

4. DISCUSSION

Of all of the 4 models generated, the threaded model is generally the least favorable and this was predicted. The I model would have been the top contender as the final model; however, it contains the least amount of core residues among all models including threaded model. Perhaps this is an interesting point, as all model quality programs offer a final score which are all in disagreement with other scores. In other words, it is inconclusive which model can be the final model. However, it is very clear that an all-atom energetic study and refinement is necessary for understanding where the models have performed well and not so well and this is outside the scope of this paper but nevertheless is initiated in Supplemental Figure 2.

Following superposition of template and target, the ligand of the template was identified and a space of within 3A, 38 water molecules exist. As mentioned, template 1LQT undergoes a catalytic triad like mechanism to construct a FprA:NADPO moiety that is not yet found in 1LQT homologues or anywhere else. The two main catalytic residues Glu214 and His57 are conserved within the alignment between template and target and further inspection can take place to identify if target undergoes the same complex formation hydroxylated NADPH (9). This is evident in Figure 7 A and B although more refined models are required for such assumptions. Regardless, it is very notable to state that such an interaction between residues and its ligand can only be determined by looking at 3D modeled structure and definitely not primacy amino acid sequence. Annotation transfer between template and target E.C number perhaps may apply because there is a high sequence identity of 41% where Rost et al. 1999 suggest a sequence identity of greater than 40% which will transfer annotation based on FSSP database (37). For this reason, it is then assumed that target and template undergo the same catalytic function.

Metagenomics of OHRB and *A. dehalogenans* strain 2CP-C-like populations are increasing due to the increase in sequencing projects and elucidate the diversity of the ecogenomic toolkit. Furthermore, classification of the function of the seminal members of the ecogenomic toolkit require understanding of the catalytic mechanism involved in redox reactions with substrates and organic or metal cofactors metal (5). Homology modeling of such organisms allows for greater understanding of which genomic features were kept through HGT events and within proteins, the seminal catalytic residues and their interaction with ligands.

Gene disruption of target protein through techniques in biotechnology can also be used to understand metabolic importance. In one example, A homolog of 1LQT and also FDPNDO, yeast *Arh1p* is which is lethal when inactivated through gene disruption (10). If such is the case, it would be very interesting to see whether protein 466500 is necessary for viability. Likewise, techniques in protein engineering such as directed evolution can also be used to design NADPO analogs (9).

Rosetta modeling of template structure with the provided input data forms constraints that can be looked into under future applications (20). Rosetta is currently mostly successful in predicting small globular proteins (>150 residues) with mostly α-helices and β-strands. Currently, the template and target models contain 3_{10} helices which caused irregularities in prediction. Another point with Rosetta is that more models generated determine the quality of models. Since only 40 models were generated for this paper, a longer and more detailed Rosetta run would prove more useful while ProSa could be used to quickly avoid bad models, as it is way quicker than SAVES and provides one number Z-score that is proportional to ideal energy. Another trick with Rosetta is to utilized fragments from one or really closely aligned structural templates to construct a

curated fragment library. Such a library perhaps would choose which templates to use such as clade members from Figure 2: although their crystal structures aren't determined, a more careful phylogenetic analysis of all determined structures against target protein can also be conducted using more tree generation iterations. Picking fragments from this library instead of the uncurated library determined using PSI-BLAST iterating and e-value cut-offs would mean more models would be generated that are energetically favorable and from evolutionary and biologically significant structures. For this reason, Phyre model outperformed when it came to choose 3_{10} helices (see section 3) as it used multiple templates. On the same note, Insight II used one template and it chooses coils over structures due to low sequence identity with the only template model. Overall, the point here is that it would be beneficial to use multiple templates but not too many which Phyre and Rosetta have done although of the two (actually all 3) only Rosetta can be modified to run its homology modeling method differently as it is open source and run locally.

Addition of experimental data such as NMR, chemical shift and distance data can bluster predictions and is suggest by Combs et al. 2013 (20). Ligand docking also would provide a more accurate illustration of whether putative proposed mechanism occurs (20). However, the issue with docking is that the model must be of good quality which has not been fully achieved yet.

REFERENCES

1. Dworkin, Martin. Myxobacteria. John Wiley & Sons, Ltd, 1993.
2. Thomas, S. H., Wagner, R. D., Arakaki, A. K., Skolnick, J., Kirby, J. R., Shimkets, L. J., ... & Löffler, F. E. (2008). The mosaic genome of Anaeromyxobacter dehalogenans strain 2CP-C suggests an aerobic common ancestor to the delta-proteobacteria. PLoS One, 3(5), e2103.
3. Maphosa, Farai, Willem M. de Vos, and Hauke Smidt. "Exploiting the ecogenomics toolbox for environmental diagnostics of organohalide-respiring bacteria." Trends in biotechnology 28.6 (2010): 308-316.
4. Maphosa, Farai, Willem M. de Vos, and Hauke Smidt. "Exploiting the ecogenomics toolbox for environmental diagnostics of organohalide-respiring bacteria." Trends in biotechnology 28.6 (2010): 308-316.
5. Futagami, Taiki, Masatoshi Goto, and Kensuke Furukawa. "Genetic System of Organohalide-Respiring Bacteria." Biodegradative Bacteria. Springer Japan, 2014. 59-81.
6. Kim, J. D., Senn, S., Harel, A., Jelen, B. I., & Falkowski, P. G. (2013). Discovering the electronic circuit diagram of life: structural relationships among transition metal binding sites in oxidoreductases. Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1622).
7. Mathews, K., van Holde and Ahern. (2000) Biochemistry Third Edition. Benjamin Cummings.
8. Dym, Orly, and David Eisenberg. "Sequence-structure analysis of FAD-containing proteins." Protein Science 10.9 (2001): 1712-1728.
9. Bossi, R. T., Aliverti, A., Raimondi, D., Fischer, F., Zanetti, G., Ferrari, D., ... & Mattevi, A. (2002). A covalent modification of NADP+ revealed by the atomic resolution structure of FprA, a Mycobacterium tuberculosis oxidoreductase.Biochemistry, 41(28), 8807-8818.
10. Manzella, Liliana, Mário H. Barros, and Francisco G. Nobrega. "ARH1 of Saccharomyces cerevisiae: a new essential gene that codes for a protein homologous to the human adrenodoxin reductase." Yeast 14.9 (1998): 839-846.

11. Dym, Orly, and David Eisenberg. "Sequence-structure analysis of FAD-containing proteins." Protein Science 10.9 (2001): 1712-1728.
12. Gromiha, M. Michael, Motohisa Oobatake, and Akinori Sarai. "Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins." Biophysical chemistry 82.1 (1999): 51-67.
13. Ding, Y., Cai, Y., Zhang, G., & Xu, W. (2004). The influence of dipeptide composition on protein thermostability. FEBS letters, 569(1), 284-288.
14. Ceccarelli, E. A., Arakaki, A. K., Cortez, N., & Carrillo, N. (2004). Functional plasticity and catalytic efficiency in plant and bacterial ferredoxin-NADP (H) reductases. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics,1698(2), 155-165.
15. Sancho, J. "Flavodoxins: sequence, folding, binding, function and beyond."Cellular and Molecular Life Sciences CMLS 63.7-8 (2006): 855-864.
16. Kurisu, Genji, et al. "Structure of the electron transfer complex between ferredoxin and ferredoxin-NADP+ reductase." Nature Structural & Molecular Biology 8.2 (2001): 117-121.
17. Yeom, Jin Ki. "Biochemical characterization of ferredoxin-NADP+ reductase interaction with flavodoxin in Pseudomonas putida." Biochemistry and Molecurar Biology Reports 45.8 (2012): 476-481.
18. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–3402
19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.
20. Combs, S. A., DeLuca, S. L., DeLuca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., ... & Meiler, J. (2013). Small-molecule ligand docking into comparative models with Rosetta. Nature protocols, 8(7), 1277-1298.
21. Kelley, Lawrence A., and Michael JE Sternberg. "Protein structure prediction on the Web: a case study using the Phyre server." Nature protocols 4.3 (2009): 363-371.

22. Bennett-Lovsey, R. M., Herbert, A. D., Sternberg, M. J., & Kelley, L. A. (2008). Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins: Structure, Function, and Bioinformatics, 70(3), 611-625.
23. Laaksonen, L. (1992). A graphics program for the analysis and display of molecular dynamics trajectories. Journal of molecular graphics, 10(1), 33-34.
24. Tutorial#1, Kahn, 20xx
25. Tutorial#2, Kahn, 20xx
26. Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., ... & Baker, D. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins: Structure, Function, and Bioinformatics, 77(S9), 89-99.
27. Feng, Y., Kloczkowski, A., & Jernigan, R. L. (2010). Potentials' R'Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. BMC bioinformatics, 11(1), 92.
28. Pagani, I., Liolios, K., Jansson, J., Chen, I. M. A., Smirnova, T., Nosrat, B., ... & Kyrpides, N. C. (2012). The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic acids research, 40(D1), D571-D579.
29. Bernal, A., Ear, U., & Kyrpides, N. (2001). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Research, 29(1), 126-127.
30. Bajorath J, Stenkamp R, Aruffo A. 1993. Knowledge-based model build-ing of proteins: Concepts and examples. Protein Sci 2:1798-1810.
31. Krieger, Elmar, Sander B. Nabuurs, and Gert Vriend. "Homology modeling."Methods of biochemical analysis 44 (2003): 509-524.
32. Zvelebil, Marketa J., and Jeremy O. Baum. Understanding bioinformatics. New York: Garland Science, 2008. Print.
33. Stormo, G. D. 2009. An Introduction to Sequence Similarity (“Homology”) Searching. Current Protocols in Bioinformatics. 27:3.1.1–3.1.7.
34. Felsenstein, J. (2004). Inferring phylogenies. Sunderland, Mass: Sinauer Associates.
35. Felsenstein, Joseph. "Phylogenies And The Comparative Method." The American Naturalist 125.1 (1985): 1. Print.

36. Brändén, Carl, and John Tooze. Introduction to protein structure. New York: Garland Pub., 1991. Print.
37. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999 Feb;12(2):85–94.
38. –
39. –
40. Wiederstein & Sippl (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407-W410.
41. Sippl, M.J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. Proteins 17, 355-362
42. --
43. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., & Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. BMC bioinformatics, 7(1), 339.
44. Kawabata T. "MATRAS: a program for protein 3D structure comparison" (2003). Nucleic Acids Res. Vol 31, 3367-9.
45. Heinig, M., Frishman, D. (2004). STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins. Nucl. Acids Res. , 32, W500-2.
46. F.C.Bernstein, T.F.Koetzle, G.J.Williams, E.E.Meyer Jr., M.D.Brice, J.R.Rodgers, O.Kennard, T.Shimanouchi, M.Tasumi, "The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures," J. of. Mol. Biol., 112 (1977): 535.
47. Velankar, S., Best, C., Beuth, B., Boutselakis, C. H., Cobley, N., Da Silva, A. S., ... & Kleywegt, G. J. (2010). PDBe: protein data bank in Europe. Nucleic acids research, 38(suppl 1), D308-D317.
48. Kim, David E., Dylan Chivian, and David Baker. "Protein structure prediction and analysis using the Robetta server." Nucleic acids research 32.suppl 2 (2004): W526-W531.
49. Combet, C., Jambon, M., Deleage, G., & Geourjon, C. (2002). Geno3D: automatic comparative molecular modelling of protein. Bioinformatics, 18(1), 213-214.

50. Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.
51. Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology, 247(4), 536-540.
52. Pearl, F. M. G., Bennett, C. F., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., ... & Orengo, C. A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Research, 31(1), 452-455.
53. Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J., & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. Nucleic acids research, 34(suppl 1), D257-D260.
54. Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., ... & Bryant, S. H. (2005). CDD: a Conserved Domain Database for protein classification. Nucleic acids research, 33(suppl 1), D192-D196.
55. Zdobnov, E. M., & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17(9), 847-848.
56. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., ... & Eddy, S. R. (2004). The Pfam protein families database. Nucleic acids research, 32(suppl 1), D138-D141.
57. Pal, D., & Eisenberg, D. (2005). Inference of protein function from protein structure. Structure, 13(1), 121-130.
58. Shen, H. B., & Chou, K. C. (2007). EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. Biochemical and biophysical research communications, 364(1), 53-59.
59. Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In The proteomics protocols handbook (pp. 571-607). Humana Press.
60. D. Charif and J. Lobry, SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis, in Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, U. Bastolla, M.

Porto, H. Roman, and M. Vendruscolo, eds., Springer Verlag, New York, 2007, pp. 207–232.

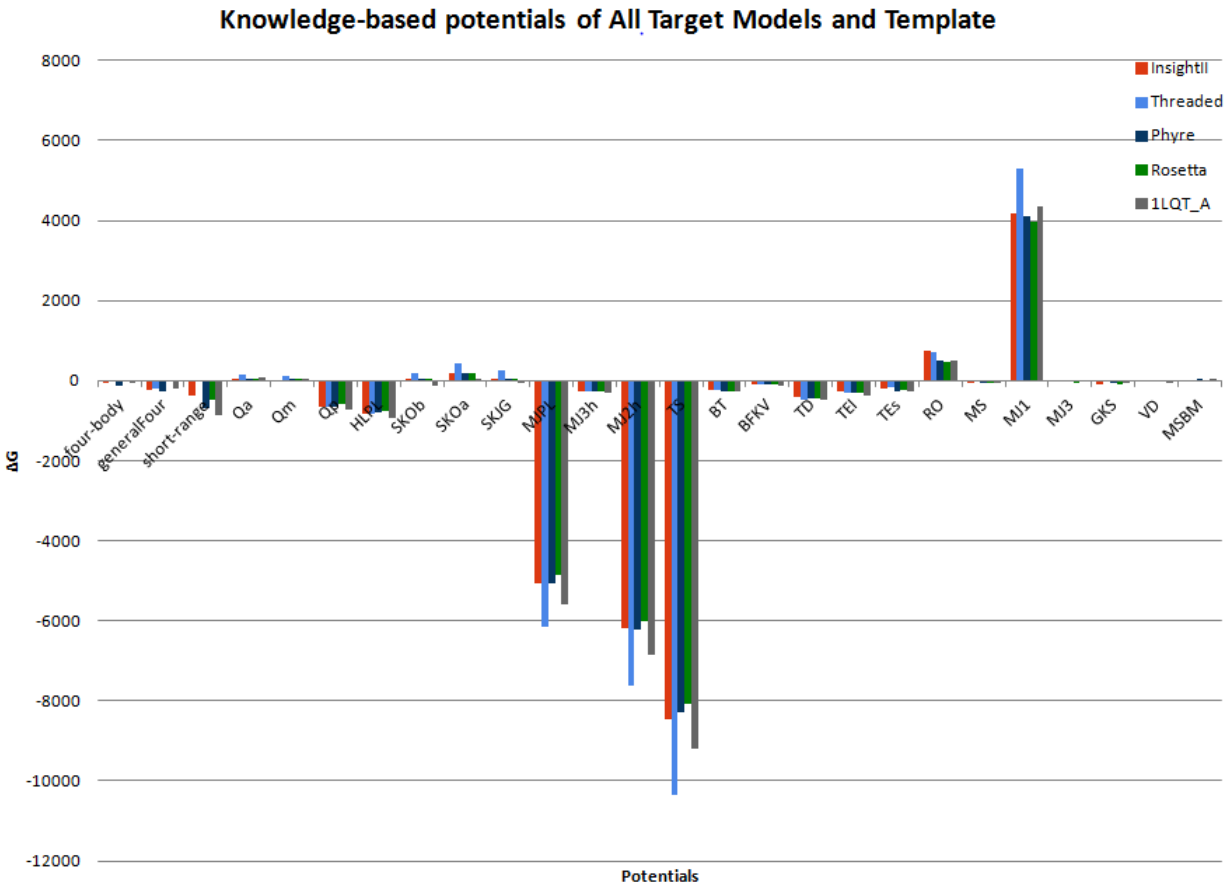
61. Y. Wei, J. Thompson, C.A. Floudas. CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization
62. 62 Wickham, H. ggplot2: Elegant Graphics for Data Analysis (Use R) p 224 (Springer, 2009).
63. Frishman, Dmitrij, and Patrick Argos. "Knowledge-based protein secondary structure assignment." Proteins: Structure, Function, and Bioinformatics 23.4 (1995): 566-579.
64. Ofran, Y., & Rost, B. (2007). ISIS: interaction sites identified from sequence.Bioinformatics, 23(2), e13-e16.
65. –
66. –
67. –
68. –
69. –
70. Laskoswki RA, MacArthur MW, Moss DS & Thorton JM. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26, 283-291.
71. Luthy R, Bowie JU, Eisenberg D. (1992). Assessment of protein models with three-dimensional profiles. Nature 356, 83-85.
72. Colovos C, Yeates TO. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 2, 1511-1519
73. Pontius J, Richelle J, Wodak SJ. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol. 264, 121-136.
74. Wiederstein & Sippl (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407-W410.
75. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality, Proteins, 2004 57: 702-710

SUPPLEMENTAL INFORMATION

Ding et. al (2004) analyzed and outlined characteristic dipeptides that correlate highly to protein thermostability: proteins found from mesophilic to hyperthermophilic archaea, dipeptide composition of **VK, KI, YK, IK, KV, KY** increased significantly and DA, AD, TD, DD, DT, HD, DH, DR, and DG decreased while in mesophilic to hyperthermophilic bacterial proteins **KE, EE, EK, YE, VK, KV, KK, LK, EI, EV, RK, EF, KY, VE, KI, KG, EY, FK, KF, FE, KR, VY, MK, WK, and WE** increased significantly and WQ, AA, QA, MQ, AW, QW, QQ, RQ, QH, HQ, AD, AQ, WL, QL, HA, and DA compositions decreased (13). Amino acid sequence of FDPNDO shows top 9 2-mer based on frequency are 12: AR, 11: AA, 10: VA, 9: LE, RR, 8: VD, 7: LE, PA, RL, 6: AV, EV, GE, GL, LG, RG, RP, VV. 5: AD, AG, DV, VG, VL, VR and 4: DA, DP, GN, LA, LP, LV, PD PG, RA and RV (13).

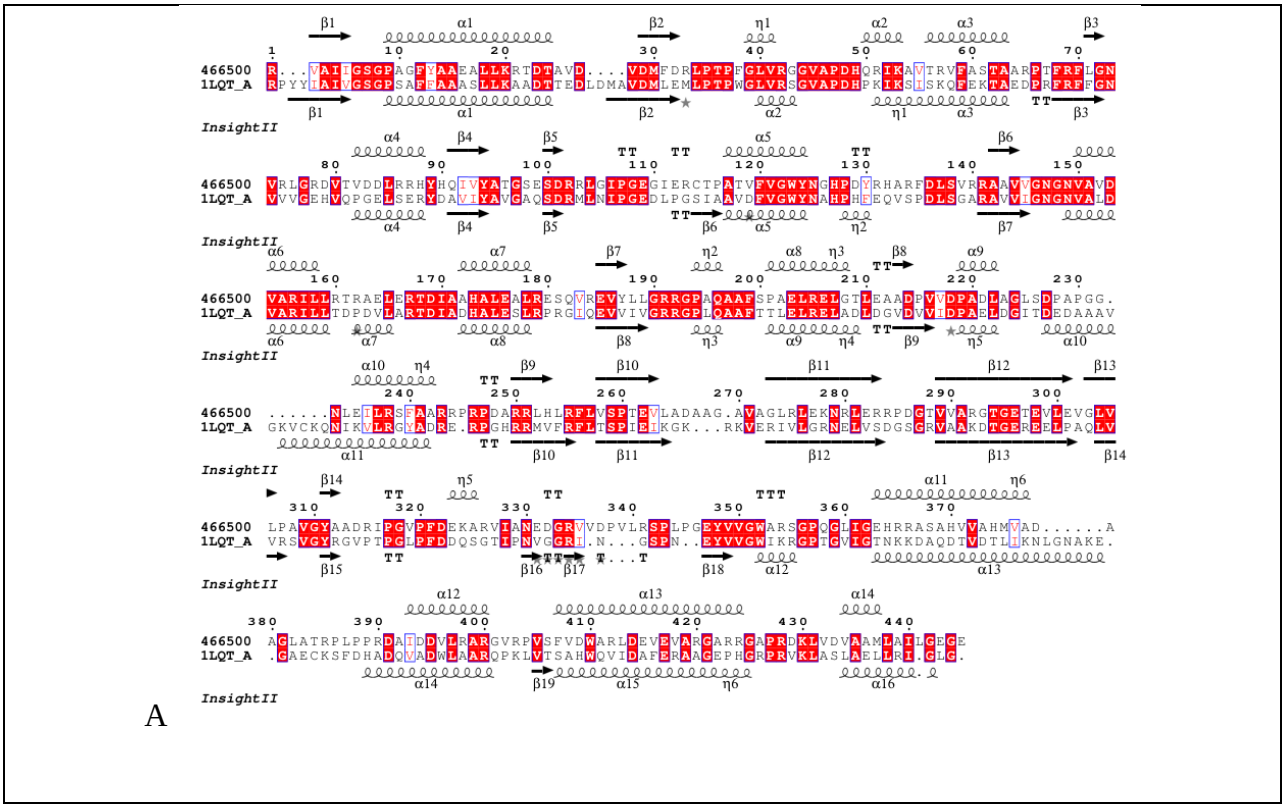
SUPPLEMENTAL FIGURES

SF1. An all protein PSI-BLAST protein alignment using BLOSUM-90 scoring matrix. Regions of identity are colored. Insight model Observed Secondary Structure is the graphic on the first row. This figure can be found under Figures folder.

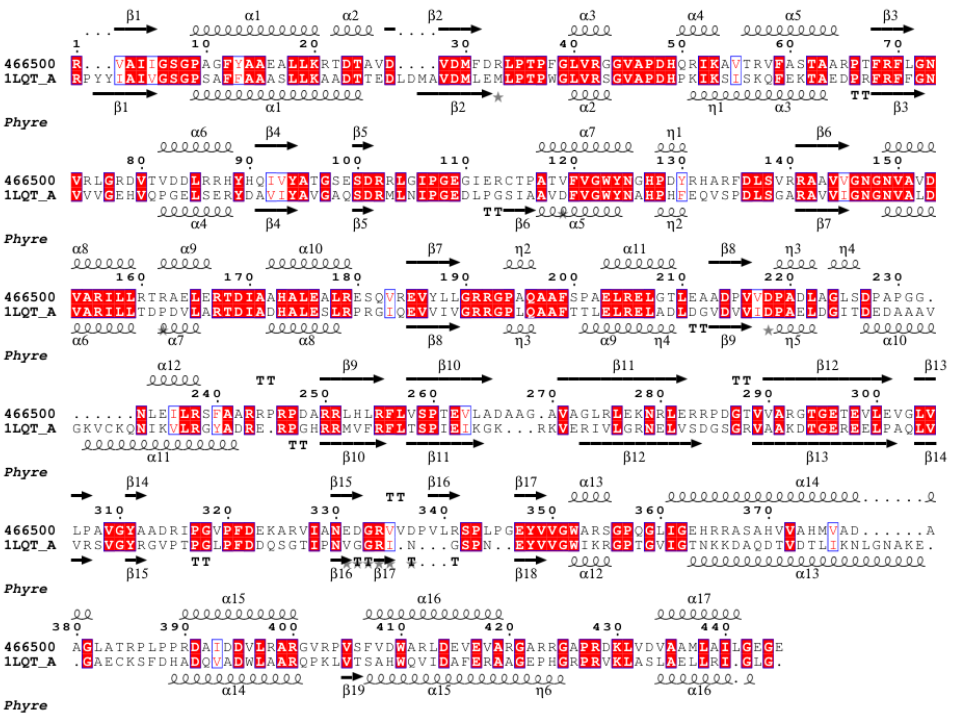


SF2. Potentials ‘R Us database Energies output with all 4 models and the template.

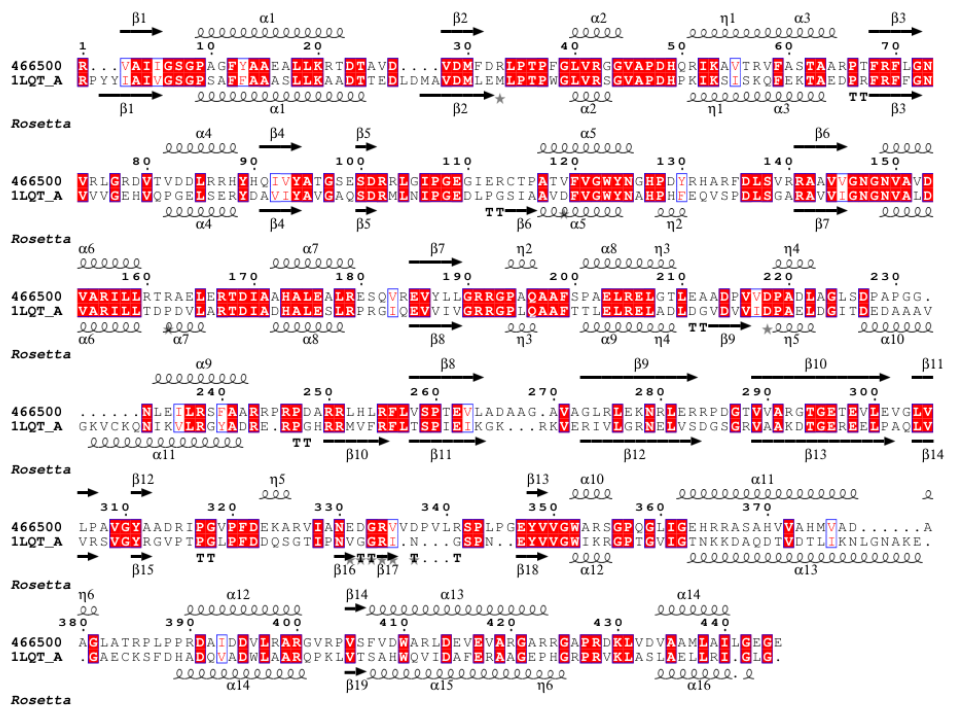
SF3. Observed secondary structure of all 4 models against observed template secondary structure. **A.** InsightII model. **B.** Phyre model. **C.** Rosetta model**. D.** Threaded Rosetta Model



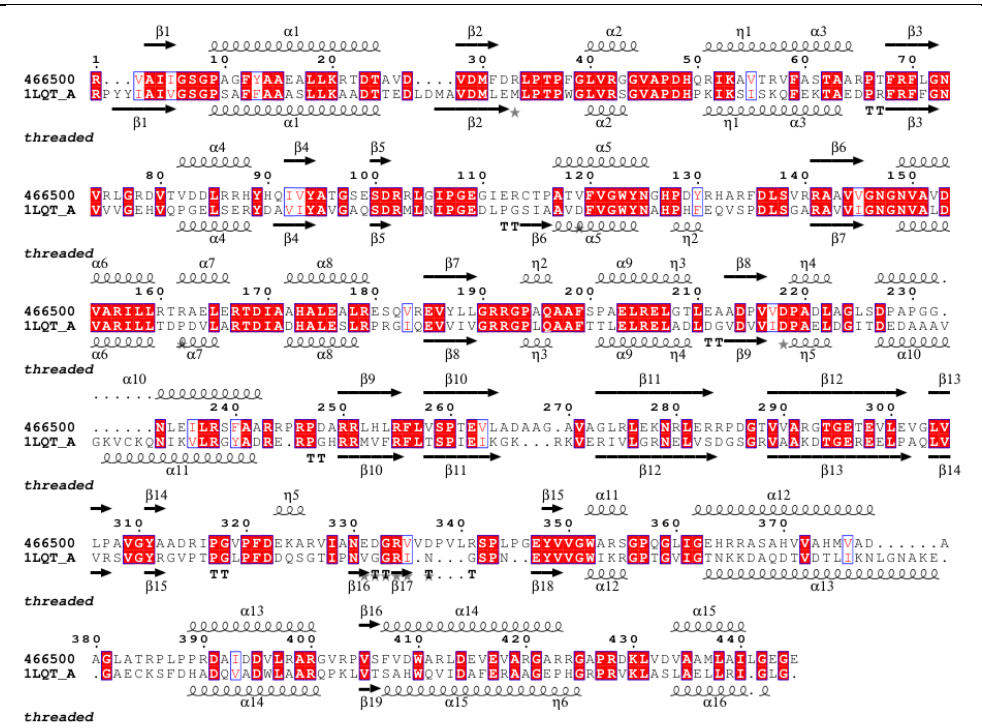
A



B



C



D

SF4. Observed secondary structure of all models side by side for better comparison.