In silico molecular modelling of MTHFR protein across eukaryotic species

ASISH KUMAR SWAIN

A dissertation submitted for the partial fulfilment of the BS-MS dual degree in Biological sciences.

Indian Institute of Science Education and Research, Mohali
April 2020

Certificate of Examination

This is to certify that the dissertation titled **"In silico molecular modelling of MTHFR protein across eukaryotic species"** submitted by **Asish kumar Swain** (Reg. No. MS15174) for the fulfilment of **BS-MS dual degree programme** of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr.Shashi Bhusan Pandit	Dr. Anand Kumar Bachhawat	Dr. Monika Sharma
(Committee member)	(Committee member)	(Supervisor)

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Monika Sharma at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography

Asish Kumar Swain

June, 2020

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Monika Sharma

(supervisor)

June, 2020

Acknowledgements

Foremost, I would like to express my deep and sincere gratitude to my supervisor Dr. Monika Sharma for her immense support, kindness, and encouragement during my thesis research. I am thankful to her for giving me the opportunity to work under her guidance.

I would like to thank Muskan Bhatia and Milind kale for their insightful comments during my research work.

Finally, I would like to thank my parents, friends who are always with me in the first place.

List of Tables

List of figures

Abbreviations

MTHFR Methylene tetra hydro folate reductase

FAD Flavin Adenine Dinucleotide

SAM S-Adenosyl Methionine

SAH S-Adenosyl homocysteine

SNP Single Nucleotide Polymorphisim

Abstract

Methylenetetrahydrofolate reductase (MTHFR) is a key regulatory enzyme involves in folate and methionine cycle which are important for the biosynthesis of nucleotide, lipid, and amino acids. Deficiency and mutations in MTHFR lead to hyperhomocysteinemia, vascular diseases, neural tube diseases, diabetes, and various cancer diseases in humans. In other eukaryotes like in plants, it has role in photorespiration, germination, root development, and lignification. In mice, MTHFR accelerates aggregation of unmodified keratin in mice hair, in this way MTHFR retains its core function in various eukaryotes. To study the various pathophysiological role of MTHFR in various species, complete 3D structures of different diverged species were modeled using template-based modelling. As loops play a major role in protein, problematic loops were refined and validated using several tools. Impact of experimentally determined mutations analysed on these models, docking of FAD and SAM to get insight into possible binding modes and how they interact with the enzyme. As identification of SNPs in the human genome growing nowadays, damaging SNPs in human MTHFR gene were analysed using SIFT, PROVEAN, PolyPhen2, Mutpred. Total of 14 SNPs were identified which affect the structure and dynamics of human MTHFR protein. As these mutations occur in the course of evolution these deleterious SNPs may having impact on other eukaryotes also.

Contents

Table of Contents

Chapter 1

Introduction

1.1.Role of MTHFR across eukaryotic phylogeny

Folate is one of the B complex vitamin having crucial role in eukaryotes from DNA synthesis to tissue repair. Methylenetetrahydrofolate reductase (MTHFR) is a key regulatory enzyme involves in folate and methionine cycle which are important for biosynthesis of nucleotide, lipid and aminoacids. MTHFR which irreversibly converts 5,10 methylenetetrahydrofolate to 5 methyltetrahydrofolate which requires FAD as cofactor and NADPH as electron donor that remethylate homocysteine to methionine in humans[1]. Deficiency or polymorphism in MTHFR leads to hyperhomocysteinemia, vascular diseases, neural tube diseases, diabetes and various cancer diseases in humans[2][3]. In plants, MTHFR plays a major role in photorespiration, embryogenesis, germination, hypocotyl and root devlopment[4][5][6][7]. Alteration in MTHFR in plants leads to hypomethylation of DNA, in recent studies shows Cosilencing of MTHFR and COMT(Caffeicacid O-methyltransferase) in plants having a significant impact on lignification[8]. Mutation in homocysteine metabolism generating N-Hcy(N-Homocysteine) which accelerates aggregation of unmodified keratins in mouse hair. Same impact also observed in birds feather generation and also in humans[9].

1.2.Structure and regulation of MTHFR

Mammalian MTHFR is a homodimer of chains, each chain having an N-terminal catalytic domain and C-terminal regulatory domain connected by a linker sequence. Each chain has an N-terminal serine-rich extension region which down regulate the MTHFR activity by protecting thermally unstable SAM[2]. Catalytic domain having binding site for the prosthetic group(FAD), substrates(NADPH & MTHF) and regulatory domain bind to S-AdenosylMethionine (SAM/AdoMet) that allosterically inhibit the enzyme activity. Inhibitory action on SAM can be switch by binding of S-adenosylhomocysteine (SAH) to the regulatory domain[10]. Prokaryotic MTHFR exists as a homotetramer having only catalytic domain which is conserved all across species. E.coli MTHFR have only catalytic domain [Fig.1] consist of 8α/8β TIM barrel which binds to cofactors and NADH as electron donor. Enzyme regulatory mechanisms like SAM binding

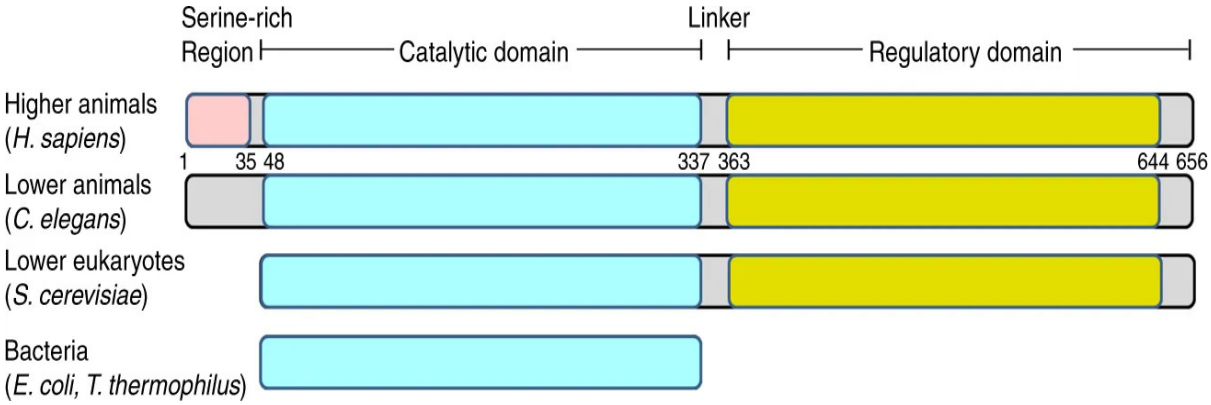

Figure 1:Domain organization of MTHFR orthologs

and phosphorylation absent in prokaryotes[11]. Eukaryote MTHFR catalytic domain comprises of 8α/8β Tim barrel as prokaryotes and the C-terminal regulatory domain consist of 2 five stranded βsheets, there is no direct contact between two domains but the linker region having contacts to both catalytic and regulatory domain. Although plant MTHFR sequence have high homology to other eukaryotes but the enzymatic reaction is different. In mammalian cell high NADPH: NADP ratio results MTHFR reaction irreversible which is inhibited by SAM but in plant cytosolic cells, low NADH: NAD which doesn't require inhibition by SAM this is how plant MTHFR is NADH dependent and SAM insensitive[12].

As MTHFR plays major role in humans, deficiency of folate or mutation in MTHFR can cause severe impact as discussed above. C.elegans, zebrafish and mouse are an appropriate model to study these, as in case of C.elegans methionine synthase have 64% identity to humans[13]. To study the various pathophysiological role of MTHFR in various species requires high resolution of complete structure. Experimental methods such as XRAY crystallography, NMR spectroscopy, electron microscopy was expensive and took years to perform. Modeling eukaryotic MTHFR is incomplete using bacterial MTHFR as a template cause it only homolog to the catalytic domain. Recently almost complete 2.5Å resolution human MTHFR crystal structure resolved in Froese, D. Sean et al[2].To obtain a rational 3D structure of MTHFR across different species comparative protein structure modeling was employed. Predicted homology models can be used for modeling substrate specificity, mutation effects, ligand designing, antigenic epitope binding prediction and more[14][15][16]. Resulting models were refined and validated using different tools. Docking of FAD and SAM also provide insight into possible binding modes and how they interact with enzymes.

Chapter 2

Methods and computational details

2.1. Sequence and structure retrieval

For phylogenetic analysis MTHFR aminoacid sequence of 23 different diverged species obtained from UniProtKB in FASTA format[17]. Human and yeast MTHFR crystallographic structure obtained from protein data bank(PDB:id-6FCX ,6FNU)[18].

2.2.Sequence alignment and phylogenetic analysis

Retrieved sequences were aligned using CLUSTAL Omega saved in phylip format. Clustal omega which generates Multiple sequence alignment using seeded guided trees and HMM profile-profile techniques[19]. Sequence similarity between retrieved sequence and human MTHFR analyzed using EMBOSS Needle[20]. Conserved and consensus positions in the sequence analyzed in CLC sequence viewer 7[21]. To estimating evolutionary relationships between sequences phylogenetic and molecular evolutionary analyses were conducted by MEGA Version7 using maximum likelihood method based on Jones-Taylor-Thorton [JTT] mode with pairwise deletions[22]. Each node of the estimated tree evaluated by the bootstrap method with 1000 replications. Physicochemical properties such as molecular weight, amino acid composition, theoretical pI, extinction coefficient, estimated half-life, aliphatic index, instability index, Grand average of hydropathicity[GRAVY] of the selected sequence computed using protparam tool[23].

2.3. 3D model building, refinement and validation

The three-dimensional model of Mus musculus, Arabidopsis analyzing, Danio rerio(zebrafish), Gallus gallus, Acanthaster planci, C.elegans with sequence similarity higher than 35% indicating strong structural conservation was modeled by homology modeling alignment of these sequences shown in figure.5. Modeller9.22 program performed to build 3D structure of the sequences. Modeller can perform comparative protein structure modelling by satisfaction of special restraints[24]. Crystal structure of human MTHFR(PDB ID:6FCX) taken as a template for modeling, 10 potential structures were generated, conformation with the lowest DOPE(discrete optimized protein energy) score was selected for further refinement.

To increase the accuracy of models, refinement of problematic loops in the protein needed. Loops with high dope score were selected from residue-by-residue energy profile.

Selected loops were refined in analyzing by using loop model class. Modeled structures were further refined in Modrefiner webserver where human MTHFR crystal structure used as a reference structure for C-alpha trace. Modrefiner refines models to their native stare in terms of hydrogen bonds, backbone topology, and side-chain modeling[25]. Rplot summary of models before and after refinement shown in table.1.. Stereochemical properties and accuracy of model analysed by Ramchandran plot in RAMPAGE server[26].Statistics of nonbonded interaction between different atom types determined in ERRAT server.Z score of backbone and sidechain contact estimated in WHAT IF server[27]. Modeled structures are further checked in Verify3D, PROCHECK, QMEAN web interface[28][29].

	R.plot summary before refinement			R.plot summary after refinement		
Organism name	A	B	C	A	B	C
Mus musculus	95.4	2.9	1.7	94.4	5.5	0.2
Gallus gallus	93.3	6.2	0.5	93.0	7.0	0
Danio rerio	97.2	2.3	0.5	96.3	2.8	0.9
Acanthaster planci	91.4	8.5	0.2	92.4	7.1	0.5
Arabidopsis thaliana	90.3	8.5	1.2	91.3	8.1	0.6
C.elegans	93.9	4.1	2.0	93.7	6.3	0.0

Table 1:Summary of Ramchandran plot of models before and after refinement of models

2.4.Mutational impact on models

As MTHFR one of the key regulatory enzyme involves in folate and methionine cycle, mutations in MTHFR leads to abnormal enzymatic functions that affect folate and methionine cycle which can cause severe diseases like Schizophrenia, neural tube defects, and various cancers. Most studies allelic variant C677T which leads to A222V substitution having high frequency throughout the world can cause reduced enzymatic activity, increase in thermolability and vascular diseases. Similarly, missense mutations E429A, G149V, A116T, R157Q, N324S are well studied can cause schizophrenia, homocystinuria in humans[30][31][32][33]. As MTHFR gene highly conserved all across the species, these mutations may affect enzymatic abnormality in other species also. To predict these mutational effects in modeled 3D structures FoldX algorithm was used to calculate $\Delta\Delta G$ value by substrating energy value from WT to mutant. As some residues in models involve in

Vanderwall clash or bad torsion angles so RepairPDB in FoldX used to optimize the energy of models before mutagenesis[34].

2.5.Cofactor addition and active site identification

Role of FAD, SAM in folate and methionine cycle is essential in MTHFR proteins. Coordinate of FAD extracted from the template PDB file map it into modelled proteins using Pymol. Majority of contacts to FAD are from catalytic domain W95, H127, R157, G158, A175, Y197, K2217 are the major residues interacting to FAD in human MTHFR. Replacing SAM in the position of SAH is more realistic as in the event of enzyme catalyzation, SAM in the regulatory domain who allosterically inhibits the enzyme. Molecular docking of SAM in MTHFR receptor performed using Autodock Vina program[35]. Receptor molecule prepared by removing all water molecules added with polar hydrogen atoms also Kollman charge added to the molecule. 3D conformer of SAM retrieved from Pubchem, Grid box for docking was set such that all residues interacting to the SAH should come in that 3D grid box. Binding affinity between different conformers of SAM and receptor (MTHFR) analyzed via negative Gibbs free energy score(Kcal/mol)which was calculated based on Vina's different scoring function. The most energetically favorable conformer of SAM selected for Further analysis which is shown in figure.2. Residues interacting with SAH(in crystal structure) and SAM(best conformer) analyzed via Ligplot+[36]. Result obtained from Ligplot+ shows, those residues interacting with SAH also interact with SAM. Coordinate of SAM extracted and map in into all modelled structures. After the addition of both cofactors to the models, conserved residues that are interacting with FAD, SAM in all models analyzed in Ligplot+.

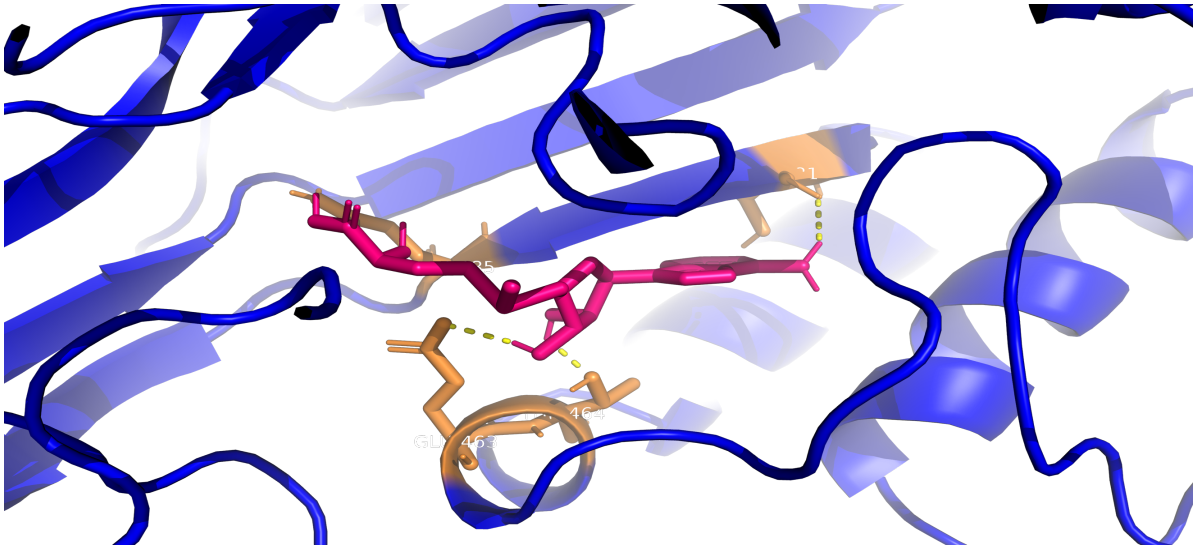

Figure 2:Surface contact of 4 crucial residues(E463, T464, T481, Q485[Orange]) to SAM in the predicted docking site of human MTHFR protein

2.6.Prediction of Deleterious nsSNPs in human MTHFR gene

Substitution of a nucleotide at some specific position in the genome called single nucleotide polymorphism(SNP). SNPs that mutate amino acids in the protein called nonsynonymous single nucleotide polymorphism(nsSNP).These nsSNPs in the coding region are related to most human diseases as they influence protein structure , dynamics, and stability of the protein. As these mutations occur in the course of evolution, deleterious SNPs found in the human MTHFR gene can have significant impact on other eukaryotic species also. To find the impact of deleterious nsSNPs in human MTHFR gene via experimental work is costly and took years to perform. Insilico analysis of nsSNPs can be a constructive alternative to experimental study. SNP information of the MTHFR gene retrieved from NCBI dbSNP database[37]. Out of 6390 SNPs 4872 SNPs in intronic region,1492 SNPs in 3 prime UTR region,602 in 5 prime UTR region and 915 SNPs were nsSNPs(missense mutation). Deleterious effect of 915 missense mutation predicted using SIFT, PROVEAN, PolyPhen2, and mutpred webserver. SIFT(Sorting Intolerant from tolerant) predict AA substitution effect based on sequence homology and physical properties of amino acids[38]. SIFT score below 0.05 considered as damaging. PROVEAN(Protein Variation effect analyser) predict impact of AA substitution by using pairwise sequence alignment, scores below -2.5 consider as deleterious substitution[39]. PolyPhen2 predicts the nsSNP effect by analyzing multiple sequence alignment and protein 3D structure. Scores between 0.85-1.00 are more confidently predicted to be damaging[40]. Mutpred predicts pathogenicity of amino acid substitution and molecular mechanisms associated with it[41]. Out of 915 missense mutation after filter out 40 nsSNPs carried out for further analysis by finding out $\Delta\Delta G$ value via Yasara Foldx tool.

Chapter 3

Results

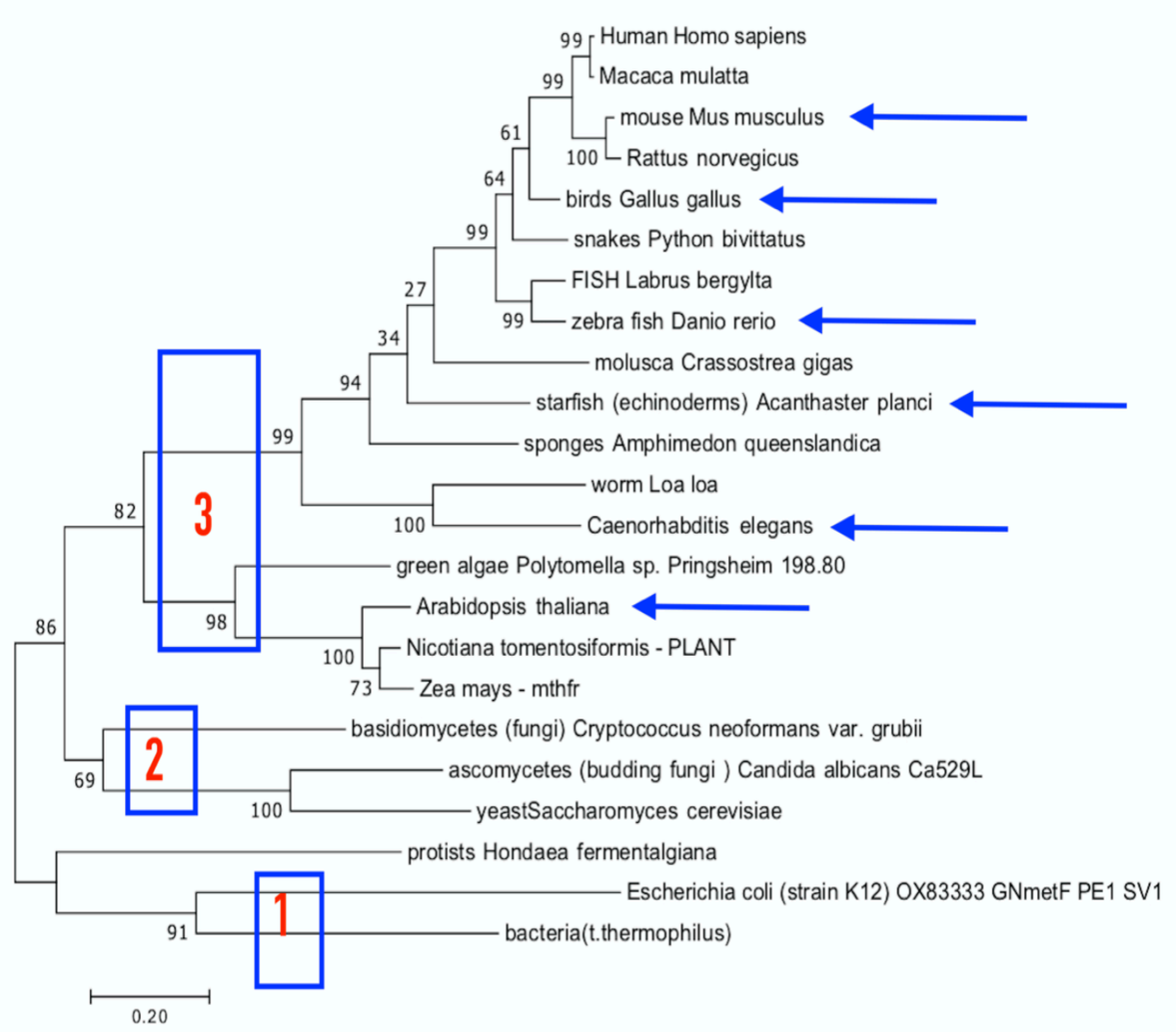

Figure 3.:Phylogenetic tree of MTHFR proteins all across species based on Jones-Taylor-Thorton mode, maximum likelihood tree with 1000 bootstrap replication.Bootstrap value above the branch suggest confidence/robustness of branch to be form.clades were labelled with boxes ,blue arrow marked sequences are considered to 3D modelling .

3.1.Sequence Alignment And Phylogenetic Analysis

Maximum likelihood tree derived from multiple sequence alignment of 23 diverged species shown in fig.3. Sequences of prokaryotes aligned to the catalytic domain of other species. Evolutionary tree of selected organisms shows delineation of MTHFR into 3

major clades. First clade consists of prokaryotes as they have the only catalytic domain and sequence similarity to human MTHFR is only 49%[to the catalytic domain of human MTHFR]. Second clade consist of fungi species, they have regulatory domain but lack serine-rich phosphorylation region. Clade 3 consists of eukaryotic species ranging from plants to advanced mammal-like humans. Nodes separating major clades having bootstrapping values [60-80] and bootstrapping values of internal nodes [>90] suggest tree is accurate and reliable. In clade 3 there is a subclade comprise of plants and algae as their sequence similarity [60% to human MTHFR] distant from other eukaryotes.

3.2 Structure Based Sequence Alignment

In the sequence alignment, high conservation of Leucine, aspartic acid, glutamic acid, proline residues in all across the species predicts these residues maybe took part in active catalytic activities, as through the process of evolution MTHFR still retain its function from prokaryotes to advanced mammals. Sequence of Mus musculus, Arabidopsis 24nalyzin, Danio rerio(zebrafish), Gallus gallus, Acanthaster planci, C.elegans having sequence similarity greater than 50% with reference to human MTHFR. These sequence were further carried out for modelling where human MTHFR crystal structure (PDB ID-6FCX)choose as template for generating homology models.

3.3.Physiochemical Parameters Of The Sequences

organism name	M.weight	sequence length	pI	EC(assuming all pairs of Cys residues form cystines)	EC(assuming all Cys residues are reduced)	Half life(hrs)	II	GRAVY	AI
Human	74596.57	656	5.22	119915	119290	1d 6h	49.02	-0.418	80.72
Mus musculus	78924.40	695	5.32	127030	126280	1d 6h	51.17	-0.416	77.17
Gallus gallus	74212.98	651	5.47	119915	119290	1d 6h	47.04	-0.450	80.11
Danio rerio	74655.33	656	5.23	126905	126280	1d 6h	44.36	-0.425	79.21
Acanthaster planci	78138.48	689	5.93	120375	119750	1d 6h	50.89	-0.467	79.97
Arabidopsis thaliana	66803.18	594	5.33	101925	101300	1d 6h	37.31	-0.279	82.88
c.elegans	75486.72	663	5.26	119010	118260	1d 6h	37.97	-0.373	79.25

Table 2:Physiochemical properties of human, Mus musculus, Gallus gallus, Danio rerio, Acanthaster Planci, Arabidopsis thaliana, C.elegans MTHFR sequence where EC=Extinction coefficient, II=Instability Index, GRAVY= Grand Average of hydropathy, AI=Aliphatic Index

ExPASy protparam tool was used to predict different physical and chemical properties of aminoacid sequences which was summarized in table.2..The result suggested that average molecular weight of MTHFR protein in eukaryotes varies from 66000-79000Da and theoreotical pI predicted to be between 5.22-5.93 which indicates that protein is negatively charged and acidic in nature.ExPASy protparam tool predict pH according to the pKa values of the aminoacid sidechains.Extinction coefficient(EC) indicates rate of transmitted light via scattering and absorption medium .EC of MTHFR protein lies between 119000-128000$m^{-1}cm^{-1}$ at 280nm.However EC of Loa loa was 135635 due to higher number of tyrosine,tryptophan and cysteine compare to other sequences. Aliphatic index(AI) of a protein is defined as the relative volume ocuupied by aliphatic sidechains like alanine,valine,isoleucine and leucine. AI of selected sequences predicted between 77-83 which is stable for wide range of temprature. Hydropathy is the hydrophobic or hydrophilic properties of the aminoacids in the protein. GRAVY(Grand Average of hydropathy)value of protein calculated as sum of hydropathy values of aminoacids divided by number of residues in the sequences. GRAVY value of selected sequences lies between -0.467to -0.279 indicates plausible interaction with water. Higher GRAVY value in C.elegans, Arabidopsis thaliana indicate presence of some hydrophobic residues.Instability index(II)predict stability of protein invitro.Instability index less than 40 predict as stable protein.Sequence of human,Mus musculus,Gallus gallus,Danio rerio and Acanthaster planci(Star fish) rich in PEST[proline(P),glutamic acid(E),serine(S),threonine(T)] may be the reason of their II more than 40. Factos such as some specific dipeptides in protein,disulphide bridge,protease recognition mechanism also influence instability index,In the sequence of C.elegans, Arabidopsis thaliana contain more cysteine(disuphide bridge)than other sequence so this may be reason their II fall below 40.

3.4.Validation Of Models

Refined models result in RMSD value of 0.972-2.380 Å to the template human MTHFR structure which is summarized in table.3. RAMPAGE server and PROCHECK suite evaluation of models reveals that in Ramchandran plot (normal distribution of points where Phi(**φ**) angels restricted to a negative value) of all 3D structures 90-95% of residues fall in the most favourable region, 5-8% residues fall in additionally allowed region and 0.7-1.5% of residues in generously allowed region. Only 1-6 residues(0-1.2%) per monomer present in the disallowed region, however, none of these residues fall in any binding site region indicates overall structures to be reasonably good which is summarized in fig.4. Verify 3D

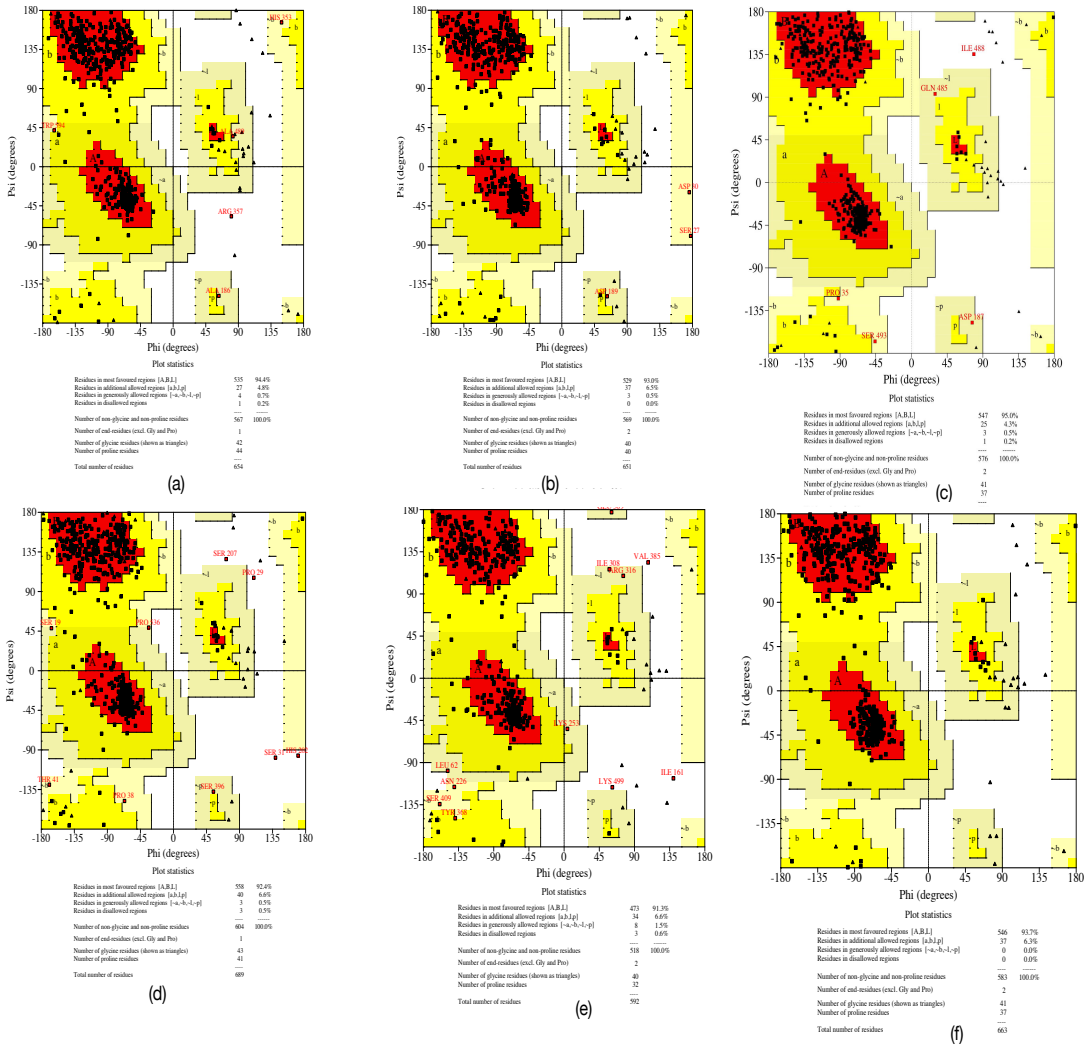

Figure 4Ramchandran plot of – (a) Mus musculus,(b) Gallus gallus,(c) Danio rerio,(d) Acanthaster planci, (e)Arabidopsis thaliana, (f) C.elegans

showed all models have 88-96% of the residues have average 3D-1D score >= 0.2[table.3] shows the quality of the model reliable. QMEAN Z score varies from 0 to 1, higher Z score relates to more favourable model and better agreement to predicted feature. QMEAN Z score of the predicted structure varies from 0.68-0.80 indicate greater degree of nativeness. In ProSA-web analysis models lie in the space of protein determined by X-RAY crystallography. Z score value of models ranges from -10 to -13 which is close to Z score of template[PDB ID-6FCX]-11.58 suggests models are so close to experimentally determined structure. WHAT IF fine packing quality control Z score of backbone and sidechain contact differs from -1.93 to -2.94 indicates models are reliable and accurate which all are summarized in table.3.

		Verify 3D	Q MEAN	ProSA	WHAT IF (Fine packing quality control)	
organism name	**Template-target RMSD(in Å)**	**Residues have average 3D-1D score >= 0.2**	**global score**	**Z score**	**All contacts score**	**z**
Mus musculus	**1.481**	90.21%	0.78±0.05	-12.39	-0.292	-1.93
Gallus gallus	**0.972**	92.78%	0.79±0.05	-11.87	-0.339	-2.23
Danio rerio	**0.988**	95.12%	0.78 ± 0.05	-12.27	-0.450	-2.94
Acanthaster planci	**1.517**	91.58%	0.70±0.05	-10.83	-0.410	-2.69
Arabidopsis thaliana	**2.380**	90.54%	0.68±0.05	-12.02	-0.323	-2.13
C.elegans	**1.574**	88.69%	0.72±0.05	-12.32	-0.548	-2.75

Table 3:Structural validation of all models, details about the tools already explained .

3.5.Structural Analysis Of Models

The catalytic domain of all models consist of α8-β8 Tim barrel with three α-helices(α9,α10,α11) and regulatory domain comprise of two five stranded βsheets(β9-β19) and 8 α-helices which confirmed from ENDScript2.0. In the structure of Arabidopsis thaliana where regulatory domain lacks an α-helix(α19). In all models, 38-40% of elements comprise of alpha-helix,12-17% of element comprises of extended strand ,5-9% of elements comprise of beta-turn and 37-43% of element comprises of random coil which all are summarized in table.6. Residue interaction diagram of all models to FAD and SAM showed in fig. 6, fig.7. In the SAM interaction diagram of Arabidopsis thaliana less number of residues interacting to it, as plant MTHFR doesn't require SAM for inhibition. A lot of conserved residues interacting to FAD and SAM in all models, these residues may be the building block of the MTHFR protein in all across the species. T94, H127, T129, M129, L156, R157, Y174, A175, A195, H213, K217 and Y321 (numbering according to human MTHFR) residues in approx all models interacting to FAD. All these residues fall in the catalytic domain. P348, A368, T464, T481, I482, N483, S484, T560 and T575 residues in approx all models interacting to SAM. All these residues fall in the regulatory domain except P348 which lies in the linker region. As these residues conserved and interacting with cofactors, mutation or single nucleotide polymorphism impact on these residues may affect the function of the protein. Mutational impact(ΔΔG) on all the models having the approx same value maybe the structures are not changing much as MTHFR one of the conserved gene all across the species. However, mutations G149V, N324S, W339G may cause instability(ΔΔG>2.0) in wide range of species. ΔΔG value obtained from FoldX summarize in table.4.

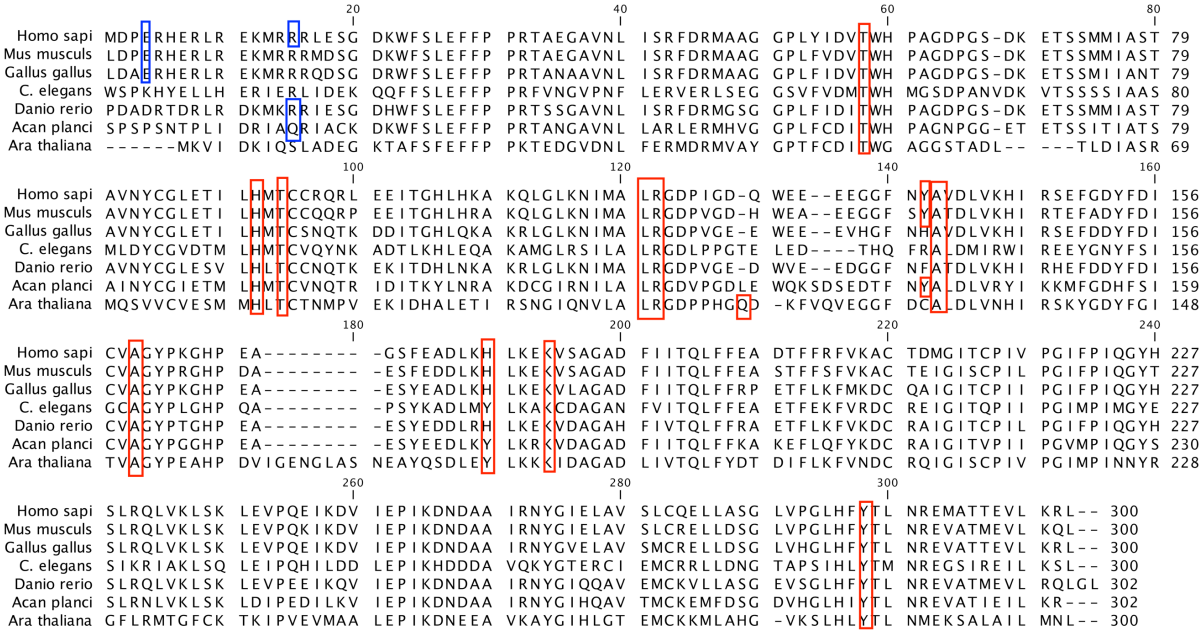

Fig.8.Sequence alignment of catalytic domain of all models. Residues inside red box interacting to FAD and residues inside blue box are the dimeric interface residues [Homo sapi(Homo sapiens), Danio rerio(zebrafish), C.elegans(Caenorhabditis elegans), Acan planci(Acanthaster planci), Ara thaliana(Arabidopsis thaliana)]

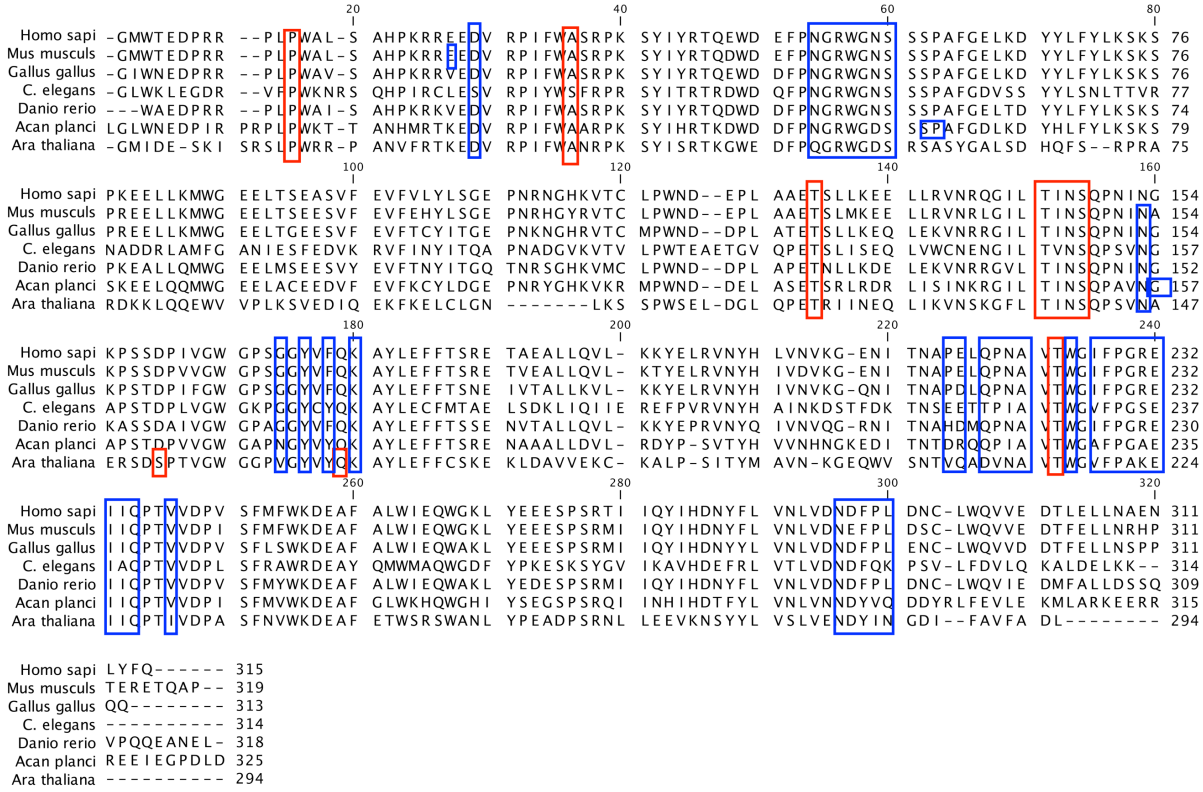

Fig.9.Sequence alignment of regulatory domain on all models. Residues inside red box are interacting to SAM and residues inside blue box are the dimeric interface residues. [Homo sapi(Homo sapiens), Danio rerio(zebrafish), C.elegans(Caenorhabditis elegans), Acan planci(Acanthaster planci), Ara thaliana(Arabidopsis thaliana)]

Sequence alignment of both catalytic and regulatory domain shown in fig.8 and fig.9. Approx all dimeric interface residues present in regulatory domain and they are conserved all across the species.

In the structure of Mus musculus catalytic domain ranges from 1-335 residue, linker region ranges from 336-360 and the regulatory domain ranges from 361 to 654. Main residues interacting with FAD are T93, W94, H95, R156, G157, Y196, D202, Q227. Main residues interacting with SAM are A367, T480, I481, N482, S483. From well-studied mutations in human MTHFR A116T, G149V, R157Q, A222V, N324S(numbers according to human MTHFR) present in catalytic domain W339G present in linker region and E429A present in regulatory domain of mouse . R157Q(in mouse R156) among the main residue interacting with FAD.

In the structure of Gallus gallus(birds) catalytic domain ranges from 1-338 residue, linker region ranges from 339-365 and the regulatory domain ranges from 366 to 651. Main residues interacting with FAD are D94, W97, H129, R159, H176, A177, D179, H203, H215, K219. Main residues interacting with SAM are P380, N458, A463, E465. From well-studied mutations in human MTHFR A116T, G149V, R157Q, A222V, N324S present in catalytic domain W339G present in linker region and E429A present in regulatory domain. R157Q(in birds R159) among the main residue interacting with FAD.

In the structure of Danio rerio(Zebrafish) catalytic domain ranges from 1-336 residue, linker region ranges from 337-363 and the regulatory domain ranges from 364 to 656 which is identical to human MTHFR. Main residues interacting with FAD are W95, H127, R157, G158, A175, W197, H201, A204, D210. Main residues interacting with SAM are S369, N456, A461. From well-studied mutations in human MTHFR G149V, R157Q, A222V, N324S present in catalytic domain, W339G present in linker region, E429A present in regulatory domain and A116T absent in Zebrafish. R157Q(in Zebrafish also R157) among the main residue interacting with FAD.

In the structure of Caenorhabditis elegans catalytic domain ranges from 1-349 residue, linker region ranges from 350-377 and the regulatory domain ranges from 378 to 663. Main residues interacting with FAD are T107, H141, R171, P175, P176, H184, F186, Y210, H214, K230. Main residues interacting with SAM are P361, F383, Q477, E479. From well-studied mutations in human MTHFR G149V, R157Q, A222V, N324S present in the catalytic domain, W339G present in the linker region, E429A present in regulatory domain and A116T absent in C.elegans. R157Q(in C.elegans R171) among the main residue interacting with FAD.

In the structure of Acanthaster planci catalytic domain ranges from 1-337 residue, linker region ranges from 338-360 and the regulatory domain ranges from 361 to 689. Main residues interacting with FAD are W93, R155, D211, K218, Y214. Main residues interacting with SAM are T484, S487, Q488, T563, T576. From well-studied mutations in human MTHFR A116T, G149V, R157Q, A222V, N324S present in catalytic domain W339G present in linker region and E429A present in regulatory domain. R157Q(in Starfish R155) among the main residue interacting with FAD.

In the structure of Arabidopsis thaliana catalytic domain ranges from 1-300 residue, linker region ranges from 301-328 and the regulatory domain ranges from 329 to 594. Main residues interacting with FAD are Q118, A131, K182. Main residues interacting with SAM are T438, W411, S441, Q466, T516. As compare to models of other species fewer residues in Arabidopsis thaliana interacting with FAD. From well-studied mutations in human MTHFR G149V, R157Q, A222V, N324S present in the catalytic domain but like other models R157(R111 in A.thaliana)not interact to FAD. A116T, W339G, E429A are absent in plants.

Mutations	Gallus gallus	z.fish	C.elegans	Ara.thaliana	Mus.musculus	A.planci(Star fish)
A116T	0.0505	absent	absent	absent	0.0346	0.0542
G149V	15.0062	12.3015	15.5432	13.0041	11.8753	12.9865
R157Q	0.7466	1.8404	0.8321	0.7635	1.6543	0.7653
A222V	1.1462	-0.1047	0.9843	1.2341	0.6742	1.2341
N324S	2.0623	2.0428	2.4143	2.8890	2.2135	1.9854
W339G	4.7338	3.7752	3.3954	absent	3.6843	4.6753
E429A	0.0594	-0.0521	-0.1654	absent	-0.0342	-0.0321

Table 4: Mutational impact of common mutations occur in human MTHFR where these values represent the ΔΔG value obtained from FoldX

3.6. Deleterious nsSNPs in human MTHFR gene and their impact on other species

AA variant	ΔΔG	SIFT SCORE	SIFT PREDICTION	PROVEAN SCORE	PROVEAN PREDICTION	PolyPhen2 SCORE	POLY_PHEN2_PRED	MUTPRED SCORE
R157Q	2.1012	0.006	Damaging	-3.93	Deleterious	1	probably damaging	0.912
A220V	3.8608	0.008	Damaging	-3.76	Deleterious	1	probably damaging	0.849
G221R	5.8128	0.001	Damaging	-7.91	Deleterious	1	probably damaging	0.954
F239L	1.9047	0.008	Damaging	-5.64	Deleterious	1	probably damaging	0.848
I256N	3.3092	0	Damaging	-6.61	Deleterious	1	probably damaging	0.951
L323P	4.5113	0.001	Damaging	-6.08	Deleterious	1	probably damaging	0.968
N324S	3.3112	0	Damaging	-4.76	Deleterious	0.983	probably damaging	0.812
R377C	3.3124	0	Damaging	-7.91	Deleterious	1	probably damaging	0.886
G387S	2.3416	0	Damaging	-5.93	Deleterious	1	probably damaging	0.891
W421C	5.8527	0.001	Damaging	-11.35	Deleterious	1	probably damaging	0.941
W421L	3.0653	0.006	Damaging	-11.14	Deleterious	0.968	probably damaging	0.882
G422R	2.6659	0.005	Damaging	-7.17	Deleterious	1	probably damaging	0.871
W500C	6.0318	0	Damaging	-12.39	Deleterious	1	probably damaging	0.956
P572L	4.5507	0	Damaging	-9.81	Deleterious	0.996	probably damaging	0.919

Table 5:Functional consequences of nsSNPs in human MTHFR

Total of 915 missense SNPs were analysed using SIFT. Out of 915 SNPs 89 were predicted deleterious (sift score < 0.05),348 were predicted tolerated and 478 SNPs not found in the database. These 89 SNPs further analysed via various tools, out of 89 SNPs 29 were predicted neutral mutation by PROVEAN, and 20 were predicted no pathogenicity by MutPred. After filtering out, 40 SNPs carried out in FoldX to quantitatively predict change in free energy upon mutagenesis. Further analyzing all the data, 14 SNPs were found damaging and structural impact on protein in all tools and ΔΔG value greater than 1.8 Kcal/mol shown in table.5. R157Q may have serious impact on various eukaryotic species as this was conserved residue that bind to FAD. Out of 14 damaging SNPs first 7 lies in catalytic domain and another 7 lies in regulatory domain.

Chapter 4

Conclusion

Role of MTHFR in eukaryote phylogeny is crucial, our modeled 3D MTHFR structure of Mus musculus, Gallus gallus, Danio rerio, Acanthaster planci, Arabidopsis thaliana, C.elegans can be used for mutational effects, ligand designing, antigenic epitope binding prediction and more. Models with cofactors can be accessed through https://doi.org/10.5281/zenodo.3781396 . T94, H127, T129, M129, L156, R157, Y174, A175, A195, H213, K217, Y321 are the conserved residues interacting to FAD and P348, A368, T464, T481, I482, N483, S484, T560, T575 are the conserved residues interacting to SAM. Mutations in these residues can cause abnormality in protein function in all across the species. R157Q is one of the most studied mutation in human MTHFR, as this is the one the conserved residue interacts with FAD this mutation may affect protein function in other species also. Predicted 14 nsSNPs are the most deleterious SNPs among MTHFR nsSNPs as their conservation score is very high, their impact on eukaryote phylogeny can't negligible.

Biblography

[1] "P. Goyette, A. Pai, R. Milos, P. Frosst, P. Tran, Z. Chen, M. Chan, and R. Rozen, 'Gene structure of human and mouse methylenetetrahydrofolate reductase (mthfr),' Mammalian Genome, vol. 9, no. 8, pp. 652–656, 1998."

[2] "D. S. Froese, M. Huemer, T. Suormala, P. Burda, D. Coelho, J.-L. Guéant, M. A. Landolt, V. Kožich, B. Fowler, and M. R. Baumgartner, 'Mutation update and review of severe methylenetetrahydrofolate reductase deficiency,' Human mutation, vol. 37, no. 5, pp. 427–438, 2016."

[3] "S.-C. Liew and E. D. Gupta, 'Methylenetetrahydrofolate reductase (mthfr) c677t polymorphism: epidemiology, metabolism and the associated diseases,' European journal of medical genetics, vol. 58, no. 1, pp. 1–10, 2015."

[4] "B. J. Reyes-Hernández, A. C. Srivastava, Y. Ugartechea-Chirino, S. Shishkova, P. A. Ramos-Parra, V. Lira-Ruan, R. I. D13̈053'faz de la Garza, G. Dong, J.-C. Moon, E. B. Blancaflor, et al., 'The root indeterminacy-to-determinacy developmental switch is operated through a folate-dependent pathway in a rabidopsis thaliana,' New phytologist, vol. 202, no. 4, pp. 1223–1236, 2014."

[5] "M. E. Stokes, A. Chattopadhyay, O. Wilkins, E. Nambara, and M. M. Campbell, 'Interplay between sucrose and folate modulates auxin signaling in arabidopsis,' Plant physiology, vol. 162, no. 3, pp. 1552–1565, 2013."

[6] "B. Gambonnet, S. Jabrin, S. Ravanel, M. Karan, R. Douce, and F. Rébeillé, 'Folate distribution during higher plant development,' Journal of the Science of Food and Agriculture, vol. 81, no. 9, pp. 835–841, 2001."

[7] "V. Gorelova, L. Ambach, F. Rébeillé, C. Stove, and D. Van Der Straeten, 'Folates in plants: research advances and progress in crop biofortification,' Frontiers in chemistry, vol. 5, p. 21, 2017."

[8] "Liu, Sijia, et al. 'Simultaneous downregulation of MTHFR and COMT in switchgrass affects plant performance and induces lesion-mimic cell death.' Frontiers in plant science 8 (2017): 982."

[9] "Borowczyk, Kamila, et al. 'Mutations in homocysteine metabolism genes increase keratin N-homocysteinylation and damage in mice.' International journal of genomics 2018 (2018)."

[10] "C. Kutzbach and E. Stokstad, 'Mammalian methylenetetrahydrofolate reductase partial purification, properties, and inhibition by s-adenosylmethionine,' Biochimica et Biophysica Acta (BBA)-Enzymology, vol. 250, no. 3, pp. 459–477, 1971."

[11] "B. D. Guenther, C. A. Sheppard, P. Tran, R. Rozen, R. G. Matthews, and M. L. Ludwig, 'The structure and properties of methylenetetrahydrofolate reductase from escherichia coli suggest how folate ameliorates human hyperhomocysteinemia,' Nature Structural & Molecular Biology, vol. 6, no. 4, p. 359, 1999."

[12] "A. D. Hanson, D. A. Gage, and Y. Shachar-Hill, 'Plant one-carbon metabolism and its engineering,' Trends in plant science, vol. 5, no. 5, pp. 206–213, 2000."

[13] "Ortbauer, Martina, et al. 'Folate deficiency and over-supplementation causes impaired folate metabolism: Regulation and adaptation mechanisms in Caenorhabditis elegans.' Molecular nutrition & food research 60.4 (2016): 949-956."

[14] "M. A. Mart13˘053'f-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Šali, 'Comparative protein structure modeling of genes and genomes,' Annual review of biophysics and biomolecular structure, vol. 29, no. 1, pp. 291–325, 2000."
[15] "C. S. Ring, E. Sun, J. H. McKerrow, G. K. Lee, P. J. Rosenthal, I. D. Kuntz, and F. E. Cohen, 'Structure-based inhibitor design by using protein models for the development of antiparasitic agents.,' Proceedings of the national academy of Sciences, vol. 90, no. 8, pp. 3583–3587, 1993."
[16] "J. Westbrook, Z. Feng, L. Chen, H. Yang, and H. M. Berman, 'The protein data bank and structural genomics,' Nucleic acids research, vol. 31, no. 1, pp. 489–491, 2003."
[17] "U. Consortium, 'Uniprot: a hub for protein information,' Nucleic acids research, vol. 43, no. D1, pp. D204–D212, 2014."
[18] "H. M. Berman, P. E. Bourne, J. Westbrook, and C. Zardecki, 'The protein data bank,' in Protein Structure, pp. 394–410, CRC Press, 2003."
[19] "F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al., 'Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega,' Molecular systems biology, vol. 7, no. 1, 2011."
[20] "S. B. Needleman and C. D. Wunsch, 'A general method applicable to the search for similarities in the amino acid sequence of two proteins,' Journal of molecular biology, vol. 48, no. 3, pp. 443–453, 1970."
[21] "Knudsen, B., et al. 'Clc sequence viewer.' A/S Cb, version 6.2 (2011)."
[22] "S. Kumar, G. Stecher, and K. Tamura, 'Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets,' Molecular biology and evolution, vol. 33, no. 7, pp. 1870–1874, 2016."
[23] "E. Gasteiger, C. Hoogland, A. Gattiker, M. R. Wilkins, R. D. Appel, A. Bairoch, et al., 'Protein identification and analysis tools on the expasy server,' in The proteomics protocols handbook, pp. 571–607, Springer, 2005."
[24] "B. Webb and A. Sali, 'Comparative protein structure modeling using modeller,' Current protocols in bioinformatics, vol. 47, no. 1, pp. 5–6, 2014."
[25] "D. Xu and Y. Zhang, 'Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization,' Biophysical journal, vol. 101, no. 10, pp. 2525–2534, 2011."
[26] "S. C. Lovell, I. W. Davis, W. B. Arendall III, P. I. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, 'Structure validation by cα geometry: ϕ, ψ and cβ deviation,' Proteins: Structure, Function, and Bioinformatics, vol. 50, no. 3, pp. 437–450, 2003."
[27] "C. Colovos and T. O. Yeates, 'Verification of protein structures: patterns of nonbonded atomic interactions,' Protein science, vol. 2, no. 9, pp. 1511–1519, 1993."
[28] "G. Vriend, 'What if: a molecular modeling and drug design program,' Journal of molecular graphics, vol. 8, no. 1, pp. 52–56, 1990."
[29] "D. Eisenberg, R. Lüthy, and J. U. Bowie, '[20] verify3d: assessment of protein models with three-dimensional profiles,' in Methods in enzymology, vol. 277, pp. 396–404, Elsevier, 1997."
[30] "K. Yamada, Z. Chen, R. Rozen, and R. G. Matthews, 'Effects of common polymorphisms on the properties of recombinant human methylenetetrahydrofolate reductase,' Proceedings of the National Academy of Sciences, vol. 98, no. 26, pp. 14853–14858, 2001."

[31] “Lewis, S. J., Zammit, S., Gunnell, D., & Smith, G. D. (2005). A meta-analysis of the MTHFR C677T polymorphism and schizophrenia risk. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 135(1), 2-4.”
[32] “Hyland, K., Smith, I., Bottiglieri, T., Perry, J., Wendel, U., Clayton, P. T., & Leonard, J. V. (1988). Demyelination and decreased S-adenosylmethionine in 5.10-methylenetetrahydrofolate reductase deficiency. Neurology, 38(3), 459-459.”
[33] “Burda, P., Schäfer, A., Suormala, T., Rummel, T., Bürer, C., Heuberger, D., ... & Kožich, V. (2015). Insights into severe 5, 10-methylenetetrahydrofolate reductase deficiency: molecular genetic and enzymatic characterization of 76 patients. Human mutation, 36(6), 611-621.”
[34] “Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. Nucleic acids research, 33(suppl_2), W382-W388.”
[35] “AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading - Trott - 2010 - Journal of Computational Chemistry - Wiley Online Library.” https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.21334?casa_token=KjTRmdZ2FKsAAAAA%3AaPmCUing1-5vp60sA5jmI5WRN2dIRxuYV9L2pTfuCPwxzJfj3XiFTWTGTLQsIqKyLfa_-14YFdFmdX8 (accessed Nov. 12, 2019).
[36] “LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery | Journal of Chemical Information and Modeling.” https://pubs.acs.org/doi/abs/10.1021/ci200227u (accessed Nov. 12, 2019).
[37] " Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, *29*(1), 308-311. "
[38] " Ng, Pauline C., and Steven Henikoff. "SIFT: Predicting amino acid changes that affect protein function." *Nucleic acids research* 31.13 (2003): 3812-3814."
[39]" Choi, Yongwook, and Agnes P. Chan. "PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels." *Bioinformatics* 31.16 (2015): 2745-2747."
[40]" Adzhubei, Ivan, Daniel M. Jordan, and Shamil R. Sunyaev. "Predicting functional effect of human missense mutations using PolyPhen-2." *Current protocols in human genetics* 76.1 (2013): 7-20."
[41]" Mort, Matthew, et al. "MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing." *Genome biology* 15.1 (2014): R19."

Appendix

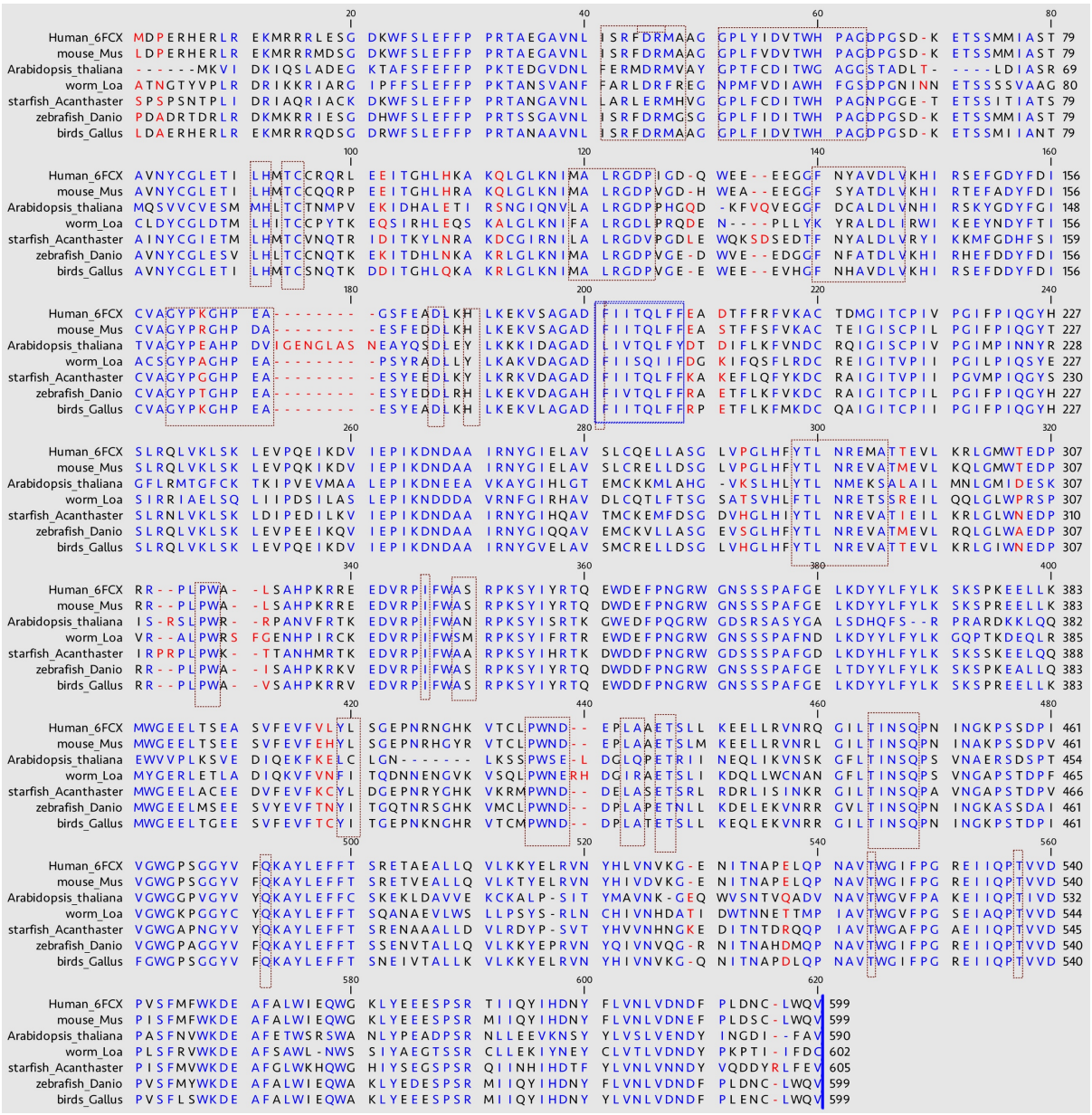

*Figure 5*Sequence alignment of Mus musculus, Gallus gallus,Danio rerio, Acanthaster planci, Arabidopsis thaliana, C.elegans. Important conserve resdidues are in boxes.

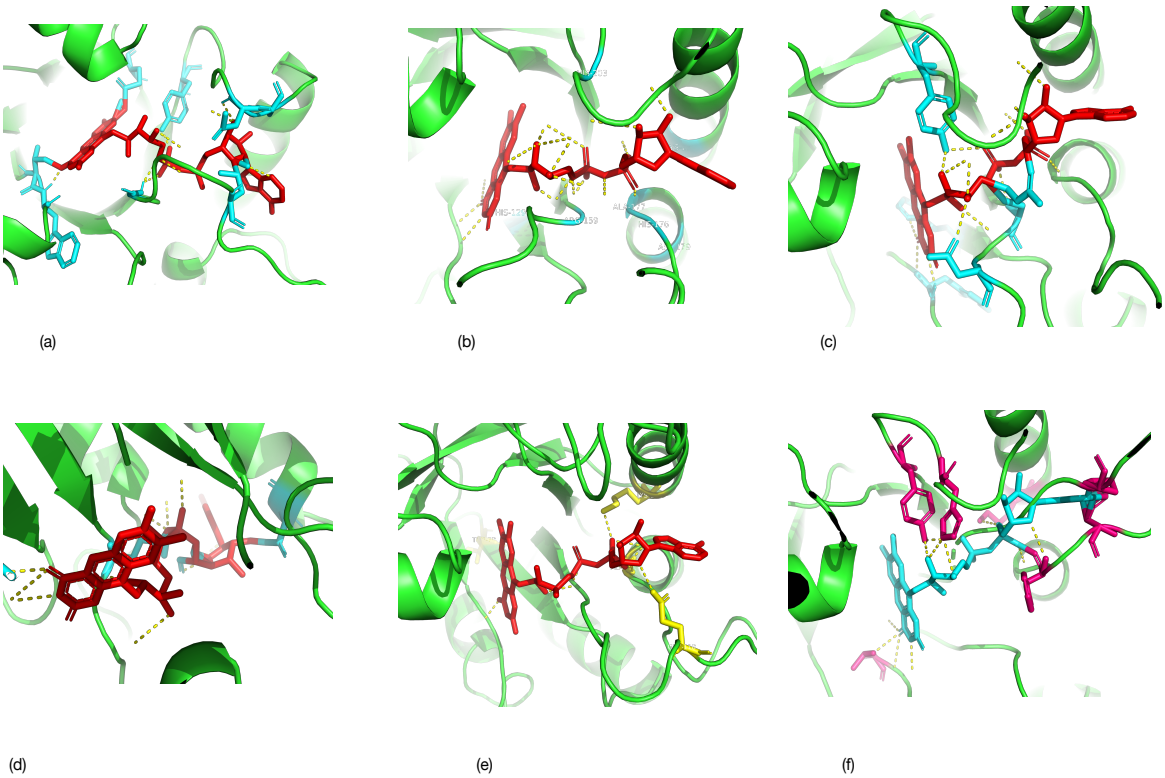

Figure.10.3D interaction digram of FAD in (a) Mus musculus,(b) Gallus gallus,(c) Danio rerio,(d) Acanthaster planci, (e)Arabidopsis thaliana, (f) C.elegans

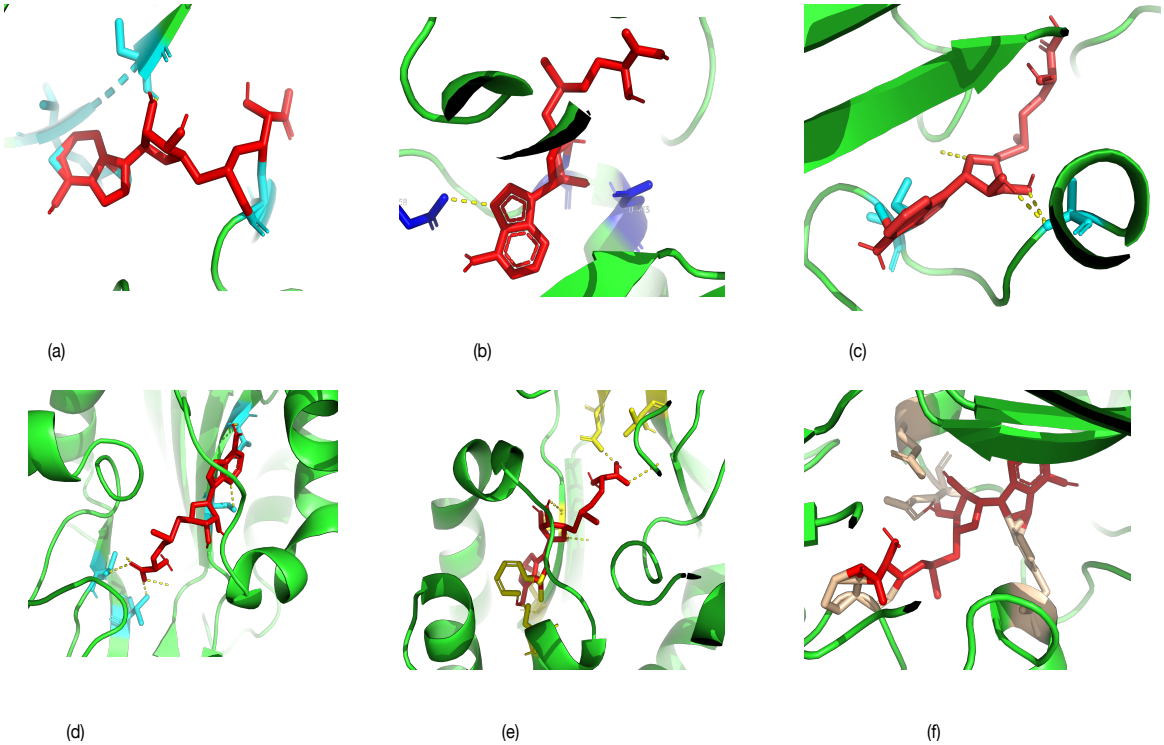

Figure.11.3D interaction digram of SAM in (a) Mus musculus,(b) Gallus gallus,(c) Danio rerio,(d) Acanthaster planci, (e)Arabidopsis thaliana, (f) C.elegans

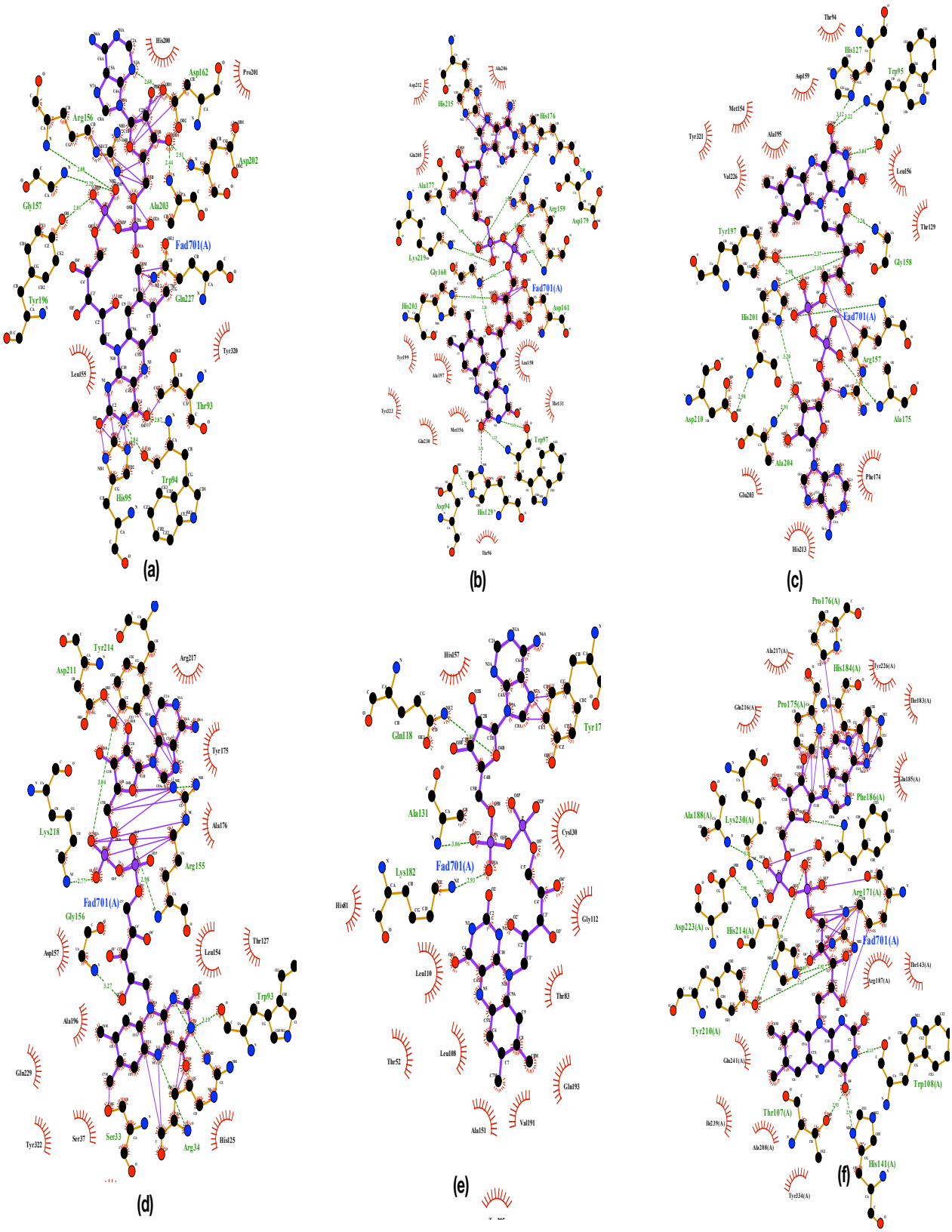

Figure 6: 2D Interaction diagram of residues to FAD – (a) Mus musculus,(b) Gallus gallus,(c) Danio rerio,(d) Acanthaster planci, (e)Arabidopsis thaliana, (f) C.elegans

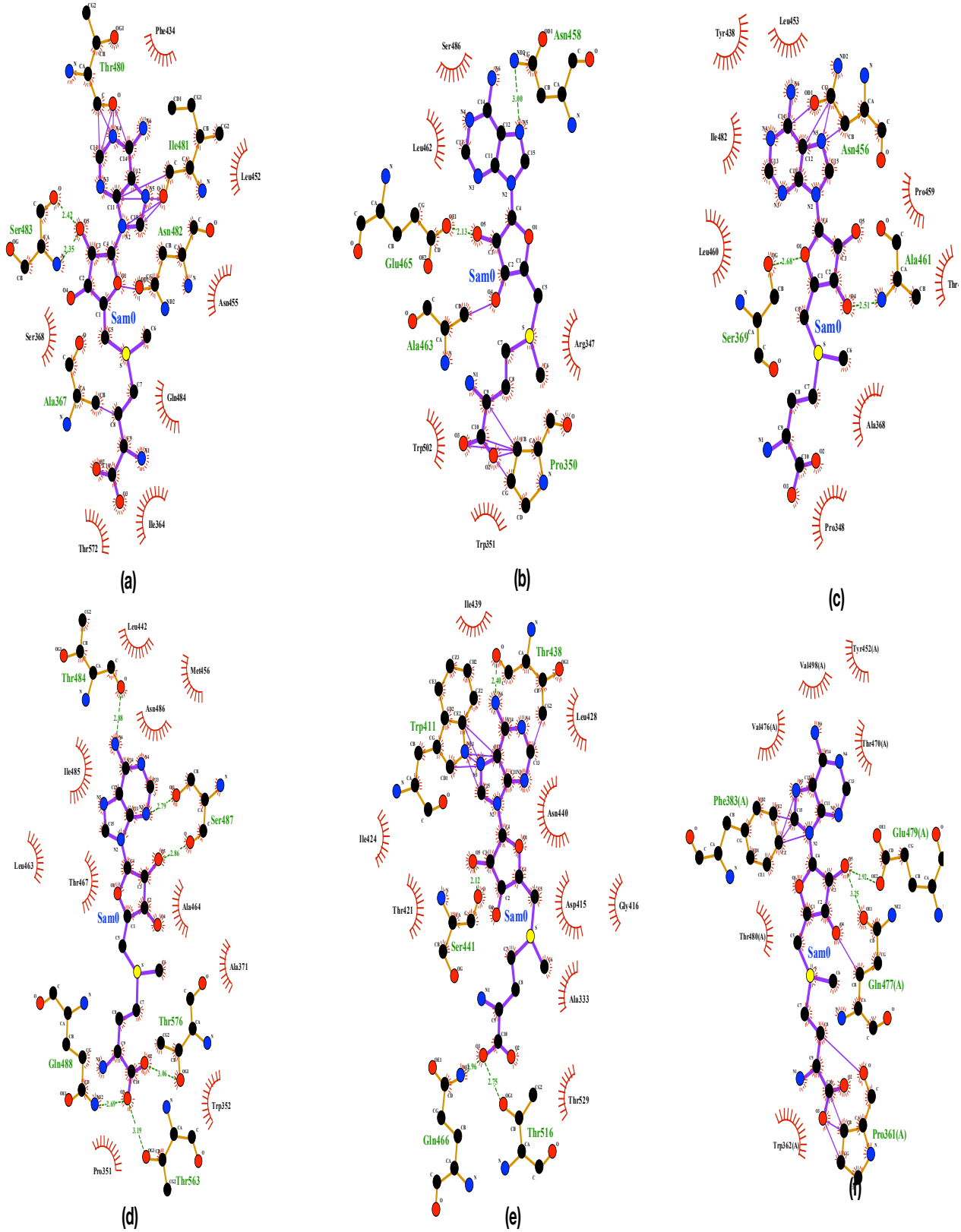

Figure7: 2D interaction digram of residues to SAM –(a) Mus musculus,(b) Gallus gallus,(c) Danio rerio,(d) Acanthaster planci, (e)Arabidopsis thaliana, (f) C.elegans

	alpha helix (%)	extend strand	beta turn	random coil	number of residues
human	38.72	12.35	5.79	43.1	656
Mus musculus	39.91	14.68	8.26	37.16	654
z.fish	38.57	13.41	5.64	42.38	656
c.elegans	39.82	15.23	7.09	37.84	663
Arabidopsis thaliana	38.72	17.34	9.34	34.51	594
Acanthaster planci	38.17	14.51	6.10	41.22	689
Gallus gallus	39.38	12.14	5.53	42.86	651

Table 6:Percentage of secondary structural elements in models