ISSN 2816-8089

RESEARCH ARTICLE

Molecular and Computational Analysis of Chlorophyll Pigment-binding Protein cp47 from Selected Species of Semi-arid Region of Western India

Hiral V Bumtariya[1,2], Kiran S Chudasama[1], Bhavisha P Sheth[1], Vrinda S Thaker[1,2*]

[1]*Plant Biotechnology and Genetic Engineering Laboratory, Department of Biosciences, Saurashtra University, Rajkot, 360005, Gujarat, India.*
[2]*Vimal Research Society for Agribiotech, 80 feet Road, Aji Area, Rajkot, 360003, Gujarat, India.*

Abstract

Photosynthesis means "synthesis with the help of light", involves the composite functioning of various protein complexes. CP47 is a pigment-binding protein of PSII of a molecular mass of about 56 kDa. CP47, encoded by the chloroplastic *psbB* gene, is an integral part of the oxygen-evolving complex of PS -II centres. In the present study, analysis of a *psbB* gene was performed from various tree, shrub, vine and herb species of Saurashtra region. The genomic DNA was isolated from the 46 samples and *psb*B gene was amplified using specific primers (60R-61F) in PCR. The amplified gene was sequenced from all plant samples and submitted to NCBI database. The length of the amplified sequence was ~300 bp, was translated to the protein sequence. The obtained sequences were analyzed with the help of CPH and Pyre2 tools. The Pyre2 tools showed 40 reliable structure prediction out of 46. ProtParam was used for carrying out the protein physico-chemical analysis of all the proteins showing variations in the protein properties. The number of residues in favored region, as observed in the Ramachandran plot analysis, indicates reliability of the protein structure prediction. The obtained results for the sequence and structure analyses may help to understand the functional application of these proteins.

Key Words: *Pigment-binding protein; psbB; DNA sequence; Protein structure; Physico-chemical properties*

***Corresponding Author**: Vrinda S. Thaker, Plant Biotechnology and Genetic Engineering Laboratory, Department of Biosciences, Saurashtra University, Rajkot 360 005, Gujarat, India; E-mail: thakervs@gmail.com*

***Received Date**: May 25, 2023, **Accepted Date**: July 04, 2023, **Published Date**: July 14, 2023*

***Citation:** Bumtariya HV, Chudasama KS, Sheth BP, et al. Molecular and Computational Analysis of Chlorophyll Pigment-binding Protein cp47 from Selected Species of Semi-arid Region of Western India. Int J Bioinfor Intell Comput. 2023;2(2):161-183.*

1. Introduction

Photosynthetic O_2 production and carbon dioxide assimilation established the composition of the biosphere and provide all life forms with essential food and fuel. Basic biochemical studies showed that chloroplast thylakoid membranes oxidize H_2O, reduce NADP, and synthesize ATP. These reactions are catalyzed by two photosystems (Photosystem I (PSI) and Photosystem II (PSII)) [1]. PSII is a unique complex that reduces water to molecular oxygen, protons, and electrons in the oxygen-evolving complex (OEC) and contains two antenna proteins, CP43 and CP47 [2,3]. CP47 encoded by the *psbB*, are chlorophyll proteins that serve as the proximal antennae for PS II, providing a conduit for excitation energy transfer from the exterior antennae of the Photosystem to the reaction center core [4]. In addition to their role as antennae, these polypeptides may also contribute to the protein environment of the water-splitting apparatus especially since they have been found to play a role in water oxidation [5].

A protein sequence is a linear hetero polymer made up of one of the 20 different amino acids. The 3D structure of proteins can be solved by experimental methods or probable structure prediction using bioinformatics tools. Solving through X-ray crystallography produces reliable results, but it needs to have pure protein sample which must form relatively flawless crystals. Solving through NMR is limited to small soluble proteins. Thus, there is a huge gap between the number of known protein sequences and the number of solved structures. Protein structure prediction aims at reducing this gap. Protein structure prediction is an important area of research in molecular biology, as the generation of enormous nucleotide sequences as a result of genome analysis needs adequate assignment of the physiological functions. The main focus is a prediction of the three-dimensional structure of a protein when only the amino-acid sequence is known. The prediction of protein structure, based primarily on sequence and structure homology using homology models has become more accurate and their range of applicability has increased. These include profile methods for sequence searches, the use of three-dimensional structure information in the sequence alignment, and new homology modeling tools, specifically in the prediction of loop and side-chain conformations [6]. There have also been important advances in understanding the physico-chemical basis of protein stability and the corresponding use of physical chemical potential functions to identify correctly folded from incorrectly folded protein conformations [7-9]. Thus, the researcher's focus on predicting protein structure from sequences remains of great importance to molecular biology. The present study was performed using two bioinformatics tools for protein 3D structure prediction viz. (a) CPH model and (b) Phyre-2 (Homology modeling based). Further, the obtained structures are validated using ANOLEA for energy levels and RAMPAGE for structural configuration.

The CASP (critical assessment of structure prediction) competition shows the progress of the different prediction methods in the last decade. The most accurate prediction method so far is the template, or homology modeling approach, which predicts the structure by comparison to a similar sequence. For two-thirds of sequences, a similar sequence can be found and thus the structure can be predicted by homology modeling with reliable precision for those with less than 300 residues [10]. Homology modeling is based on the fact that if two sequences have a high sequence similarity then they have similar 3D structure. But this is not always the case. Two sequences that don't share much sequence similarity may have similar folds. In the present study, a total of 46 plant species were studied in and around Rajkot city (Gujarat, India) of which 18 are trees, 7 are vines, 12 are shrubs and 9 are herbs (Table 1). They were subjected to DNA isolation, and PCR amplification for the CP47 gene followed by sequencing

of the PCR product. These nucleotide sequences were submitted to NCBI, and the accession numbers obtained are presented in Table 1. These sequences were translated to protein sequences for further analysis for its structure prediction.

2. Materials and Methods

2.1. Plant sampling

A total of 46 plants from different genera representing 14 different plant families were selected for the present study. The plants were collected from various sites in and around Rajkot, Gujarat, India (Table 1).

Table 1: *List of plant sample and accession no. of submitted sequence.*

No.	Plant name	T/V/S/H	Family	Accession No.
1	*Artocarpus heterophyllus*	Tree	Moraceae	JQ435804
2	*Bambusa* sp.1	Tree	Poaceae	JQ435805
3	*Mangifera indica*	Tree	Anacardiaceae	JQ435806
4	*Borassus flabellifer L.*	Tree	Arecaceae	JQ435807
5	*Cassia fistula*	Tree	Fabaceae	JQ435808
6	*Tamarindus indica*	Tree	Fabaceae	JQ435809
7	*Adina cordifolia*	Tree	Rubiaceae	JQ435810
8	*Bambusa* sp.2	Tree	Poaceae	JQ435811
9	*Crataeva nurvula*	Tree	Capparaceae	JQ675557
10	*Ficus longifolia*	Tree	Moraceae	JX141423
11	*Mimusops elengi*	Tree	Sapotaceae	JQ828835
12	*Cassia javanica*	Tree	Fabaceae	JX141424
13	*Balanites aegyptiaca*	Tree	Zygophyllaceae	JX141425
14	*Trifera* sp.	Tree	Vitaceae	JQ828834
15	*Bambusa* sp.3	Tree	Poaceae	JX141427
16	*Aegle marmelos*	Tree	Rutaceae	JX141428
17	*Bambusa* sp.4	Tree	Poaceae	JX141429
18	*Mangifera indica*	Tree	Anacardiaceae	JQ340479
19	*Abrus* sp.2	Vine	Fabaceae	JQ675556
20	*Cucurbita* sp.	Vine	Cucurbitaceae	JQ675558
21	*Abrus* sp.1	Vine	Fabaceae	JQ828836
22	*Vernonia elaeagnifolia*	Vine	Asteraceae	JQ675559
23	*Akebia quinata*	Vine	Larbizabalaceae	JX141426
24	*Abrus* sp.3	Vine	Fabaceae	JX141430
25	*Coccinia grandis*	Vine	Cucurbitaceae	JX141431
26	*Malpighia emarginata*	Shrub	Malphigiaceae	JQ422185
27	*Occimum* sp.1	Herb	Lamiaceae	JQ422188

28	*Occimum* sp.3	Herb	Lamiaceae	JQ422189
29	*Tecoma stans*	Shrub	Bignoniaceae	JQ422190
30	*Duranta erecta*	Shrub	Verbenaceae	JQ422191
31	*Holarrhena antidysenterica*	Shrub	Apocynaceae	JQ422192
32	*Cassia tora*	Shrub	Fabaceae	JQ828838
33	*Indigofera coerulea*	Shrub	Fabaceae	JX141432
34	*Ixora coccinea*	Shrub	Rubiaceae	JQ675553
35	*Ixora* sp.	Shrub	Rubiaceae	JQ675554
36	*Cactus* sp.	Shrub	Cactaceae	JX141433
37	*Artabotrys hexapetalus*	Shrub	Annonaceae	JX141434
38	*Mimosa pudica*	Shrub	Fabaceae	JX141435
39	*Occimum* sp.2	Shrub	Lamiaceae	JX141436
40	*Cleome viscosa*	Herb	Capparaceae	JQ675551
41	*Striga angustifolia*	Herb	Scrophulariaceae	JQ422186
42	*Eclipta alba*	Herb	Asteraceae	JQ422187
43	*Merremia gangetica*	Herb	Convolvulaceae	JQ828839
44	*Martynia diandra*	Herb	Martyniaceae	JQ675552
45	*Bryophyllum pinnatum*	Herb	Crassulaceae	JQ675555
46	*Aerva javanica*	Herb	Amaranthaceae	JQ828837

2.2. DNA isolation

Total genomic DNA was isolated from fresh leaves by grinding 200 mg of tissue to powder in liquid nitrogen with a mortar and pestle, followed by the extraction protocol of [11]. The amount of DNA purity was calculated spectrophotometrically. Then the gel electrophoresis was carried out.

2.3. Amplification of *psbB* gene and DNA sequencing

The extracted DNA was used to amplification *psbB gene* using primer 60F (5'-ATG GGT TTG CCT TGG TAT CGT GTT CAT AC-3') and 61R (5'-TCC CAA TAY ACC CAA TGC CAG ATA G-3') (Graham and Olmstead 2000). DNA was amplified in a total volume of 25 µl. The reaction mixture contained 2.5µl 10X buffer (10mM Tris-HCl pH 9.0, 50mM KCl, 0.1% Trion X100), 1.5mM $MgCl_2$, 200µM each deoxynucleoside triphosphate, 10 µM primer and 1U of *Taq* DNA polymerase, 200 ng DNA. PCR reactions were performed using the Viriti™ Thermal Cycler with 35 cycles of denaturation at 94°C temperature for 1 min., annealing 52°C to 56 °C for 45s, and extension was done at 70°C for 2 min, and final extension at 72°C for 7 min. Amplified DNA fragments were electrophoresis through a 1.5% agarose gel. The purified PCR products were sequenced using a Big Dye Terminator V 3.1 Cycle Sequencing Kit using ABI 3130 genetic analyzer. The nucleotide sequences determined in this study have been submitted to the NCBI GenBank database.

2.4. Sequence analysis

The sequences obtained using the ABI 3130 Data Collection Software were further analyzed using ABI Sequencing Analysis Software v5.1. Sequence editing was performed using Bioedit. Nucleotide sequence converts to amino acid sequence using Bioedit. The amino acid composition of this sequence was commutated using the ExPASy's ProtParam tool. All 46 protein sequences were analyzed in different structure prediction software. The 3-D structure of the protein was predicted by the following CPH models and Phyre2 [12].

2.4.1. CPHmodels-3.0 server

The CPH model-3.0 server is used for the protein structure prediction for all plant species studied. It is based on a homology modelling algorithm. This includes a PsiBlast search against a reduced non-redundant protein sequence database (nr), profile-profile alignment including predicted local structure information obtained from NetSurfP [13], and a double-sided Z-score evaluation. Once the appropriate template has been found, Ca-atom coordinates are extracted according to the sequence alignment and used as a starting point for the homology-modeling process. The modelling methodology uses an optimized alignment scoring function that beyond secondary structure includes predicted relative surface accessibility. It employs a double-sided Z-score to rank individual template hits. This Z-score ranking attempts to reduce the bias imposed by the composition and length of the query and template database sequences on the alignment score, significantly improving the overall prediction accuracy. A Z-score threshold value is 3.8. Z-score is above 3.8 which means the protein model is 'reliable' and Z-score is below 3.8 which means the protein model is 'not-reliable'.

2.4.2. Phyre2

The second online software is phyre2 which also uses a homology modeling algorithm [14]. However, the Phyre server uses a library of SCOP (structure classification of protein) database and is augmented with newer depositions in the Protein Data Bank (PDB). The sequence of each of these structures is scanned against a nonredundant sequence database and a profile is constructed and deposited in the 'fold library'. The known and predicted secondary structure of these proteins is also stored in the fold library. A user-submitted sequence, henceforth known as the 'query', is similarly scanned against the non-redundant sequence database, and a profile is constructed.

Following profile construction, the query secondary structure is predicted. Three independent secondary structure prediction programs are used in Phyre: Psi-Pred13, SSPro14, and JNet15. Each of these three programs provides a confidence value at each position of the query for each of the three secondary structure states. These confidence values are averaged and a final, consensus prediction is calculated and displayed beneath the individual predictions. In addition, the program Disopred16 is run to calculate a two-state prediction of which regions of the query are most likely to be structurally ordered (o) and which are disordered (d). Usually, a high sequence identity will be indicative of a high-accuracy model.

Validations of these models were done by Ramachandran plot [15]. The protein energy level was calculated using ANOLEA.

2.4.3. ANOLEA (Atomic Non-Local Environment Assessment) server (http://melolab.org/anolea)

ANOLEA is a program used to calculate the NL profile of a protein structure containing one or more chains. The energy of each pairwise interaction in this non-local environment is taken from a distance-dependent knowledge based mean force potential that has been derived from a database of 147 non-redundant protein chains with a sequence identity below 25% and solved by X-Ray crystallography with a resolution lower than 3 Å. The method uses a very sensitive and accurate atomic mean force potential (AMFP) to calculate the non-local energy profile (NL-profile) of the structure of a protein. The AMFP-derived energy profiles can correlate high scores with point errors and misalignments in the models. Point errors are frequently found in loops or regions of structural differences between the template and the target protein. The misalignments are detected with very high scores. The performance of the method was also tested for the assessment of X-ray-solved protein structures. First, the NL profile of a protein structure refined in the incorrect space group has very high scores in several regions. One region has already been described to be out-of-register with the density map of the structure. The NL profile of the re-refined structure with the correct space group is vastly improved. In the second case, the method can accurately point out disordered residues, even if the atoms of these residues do not violate the sum of the van der Waals. ANOLEA calculates the energy level of the protein chain. This software is based on the color in the protein sequence, where if most of the part is red indicates that the protein structure level is reliable and if a large yellow portion exists, it indicates that the protein structure level is not reliable.

2.5. Analysis of physicochemical parameters

The different physicochemical properties of all the 46 *psbB* protein sequences were computed using the ExPASy's ProtParam tool. The ProtParam includes the following computed parameters: Molecular weight (M. Wt), theoretical pI, instability index (II), aliphatic index (AI), and grand average of hydropathicity (GRAVY). The computed isoelectric point (pI) will be useful for developing buffer systems for purification by the isoelectric focusing method [16]. The instability index provides an estimate of the stability of our protein. A protein whose instability index is smaller than 40 is predicted as stable; a value above 40 predicts that the protein may be unstable [17]. The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of the thermostability of globular proteins [18].

3. Results and Discussion

It is interesting to note that there is no protein structure for CP47 for these studied higher plants available in PDB (protein database). Therefore, an attempt is made to predict protein structure for 46 plants in this study.

3.1. Protein structure prediction using homology modelling tools

3.1.1. CPH model

In the present study, 46 plant species which includes 18 trees, 7 vines, 12 shrubs, and 9 herbs are preceded in CPH software and the respective Z scores according to the software are recorded in Table 2. Out of 18 tree species 10 tree species have above 3.8 Z-score and the remaining 8 had below this. In the 10 tree species highest Z-score was observed in *Balanites aegyptiaca* (7.2) followed by *Artocarpus heterophyllus* (5.2), *Bambusa* sp. (5.5), *Borassus flabellifer L.* (4.3), *Cassia fistula* (4.5), *Ficus longifolia* (4), *Mimusops elengi* (6.6), *Cassia javanica* (6), *Trifera* sp. (4.2) and *Bambusa* sp.3 (4.1). The significant Z-score for these tree species indicated that their protein structures are good or predicted properly.

From 7 vine species, 5 showed a significant Z-score while 2 species were non-significant. In 5 species the highest score was observed in *Abrus* sp.2 (5.1) followed by *Cucurbita* sp. (3.9), *Abrus* sp.1 (4.4), *Vernonia elaeagnifolia* (4.6) and *Akebia quinata* (4.2) and hence indicates that they are good protein models.

Table 2: *Protein structure prediction using CPH model (the threshold value of Z-score >3.8 indicate an appropriate 3-D structure*).

Plant name	AA	Template found	Z-score	Alignment length	ID	Coverage	Model Mw
Artocarpus heterophyllus	83	2DXR.C	**5.2**	38	34.2	34.9	2902
Bambusa sp.1	89	2EP8.A	**5.5**	52	21.2	58.4	6090
Mangifera indica	84	1D0B.A	3.5	61	14.8	72.6	6833
Borassus flabellifer L.	90	1JJD.A	**4.3**	35	37.1	32.2	3066
Cassia fistula	82	1MTP.B	**4.5**	22	31.8	26.8	2418
Tamarindus indica	66	1S5L.B	-	-	30	89.4	6251
Adina cordifolia	78	1CHL.A	3.3	15	46.7	19.2	1577
Bambusa sp.2	83	3BZ1.B	-	-	56.7	79.5	7317
Crataeva nurvula	90	1EM8.D	3.3	39	20.5	43.3	4444
Ficus longifolia	50	2GW4.A	**4**	46	21.7	92	5287
Mimusops elengi	52	1EYF.A	**6.6**	38	26.3	73.1	4763
Cassia javanica	86	1BL8.A	**6**	36	11.1	41.9	3945
Balanites aegyptiaca	71	1DLO.A	**7.2**	37	24.3	52.1	3984
Trifera sp.	87	2YTE.A	**4.2**	42	28.6	47.1	4774
Bambusa sp.3	40	2AXT.Z	**4.1**	31	19.4	77.5	3324
Aegle marmelos	39	1NJ3.A	3.1	21	28.6	53.8	2448
Bambusa sp.4	49	2FOT.C	3.1	21	23.8	42.9	2573
Mangifera indica	52	1W1H.A	3.3	21	33.3	40.4	2532
Abrus sp.2	63	2CRC.A	**5.1**	58	19	90.5	6085
Cucurbita sp.	76	1WVE.C	**3.9**	40	22.5	52.6	4318
Abrus sp.1	77	2JX1.A	**4.4**	27	33.3	32.5	2622

Vernonia elaeagnifolia	57	1EBU.A	**4.6**	31	6.5	54.4	3576
Akebia quinata	75	1UB4.A	**4.2**	80	13.8	96	8287
Abrus sp.3	52	1MMO.A	3.4	11	45.5	21.2	1273
Coccinia grandis	85	2DJ6.A	3.1	60	15	70.6	6761
Malpighia emarginata	69	2F6A.E	**5.1**	34	29.4	37.7	2825
Occimum sp.1	69	2F6A.E	**5.1**	34	29.4	37.7	2825
Occimum sp.3	91	15SL.B	-	-	48	54.9	5313
Tecoma stans	87	1Y96.B	3.5	57	24.6	58.6	5744
Duranta erecta	64	2AXT.1	3.7	23	30.4	35.9	2432
Holarrhena antidysenterica	90	1EMA.D	3.3	39	20.5	43.3	4444
Cassia tora	57	2IYB.E	**4.2**	24	25	38.6	2640
Indigofera coerulea	52	1W1H.A	3.3	21	33.3	40.4	2532
Ixora coccinea	90	1PM7.A	**3.7**	77	19.5	83.3	8259
Ixora sp.	87	2J00.N	**4**	36	27.8	40.2	3736
Cactus sp.	70	1BMR.A	**5**	37	-	-	-
Artabotrys hexapetalus	46	1LQB.D	3.4	8	37.5	15.2	773
Mimosa pudica	63	2F49.C	**3.9**	7	14.3	11.1	761
Occimum sp.2	39	2I3S.B	3	32	18.8	82.1	3860
Cleome viscosa	85	ZZ3R.T	3.6	54	14.8	63.5	5827
Striga angustifolia	63	2A49.C	**3.9**	7	14.3	11.1	761
Eclipta alba	91	155L.B	-	-	48	54.9	5312
Merremia gangetica	59	1IGL.A	**4.7**	39	10.3	64.4	4191
Martynia diandra	58	1S5L.B	-	-	56.2	81	5219
Bryophyllum pinnatum	71	2HFG.R	3.5	4	50	5.6	473
Aerva javanica	65	1QFW.B	**4**	40	25	61.5	4095

Out of 12 shrub species, 5 shrub species showed above a 3.8 Z-score which indicates that these all species have a reliable protein model. The remaining 7 species have a below 3.8 Z-score and thus are poor protein models. Highest Z-scores were observed in *Malpighia emarginata* (5.1) and have higher score than other protein models. The other 4 shrubs also have relatively less value for the protein models are *Cassia tora* (4.2), *Ixora* sp. (4), *Cactus* sp. (5) and *Mimosa pudica* (3.9). From 9 herb species 5 species have significant Z-score value. Amongst the 3 species highest Z-score was observed in *Occimum* sp.1 (5.1), *Merremia gangetica* (4.7) followed by *Striga angustifolia* (3.9) and *Aerva javanica* (4) and hence indicated that the protein models are of good quality.

Thus, a total of 24 plants out of 46 showed reliable protein prediction.

3.1.2. Phyre 2

In the present study, out of 18 tree species only 1 tree species i.e., *Borassus flabellifer L.* has a poor protein model because of a low sequence identity (25%). The other parameters of the same model are: confidence score is 57.3%, coverage is 18%, the disorder is 20%, alpha helix is 21%, beta strand is 14%. Among the 7 vine species only one, i.e., *Coccinia grandis* has a poor

model. This species has a sequence identity of 25%, Confidence score-12. 8%, coverage-28%, disorder-22%, alpha helix-79%, and beta strand-9% (Table 3).

Table 3: *Protein structure prediction using phyre 2 (the threshold value of sequence identifies >20 % indicates an appropriate 3-D structure).*

Plant name	Confidence	Coverage	Disordered	alpa helix	beta strand	%ID
Artocarpus heterophyllus	18.4	19	18	16	24	38
Bambusa sp.1	9.1	7	33	47	13	86
Mangifera indica	20.6	18	15	8	50	60
Borassus flabellifer L.	57.3	18	20	21	28	25
Cassia fistula	99.9	98	20	46	14	43
Tamarindus indica	99.6	68	20	20	36	40
Adina cordifolia	9.2	19	24	62	0	27
Bambusa sp.2	100	92	14	45	14	49
Crataeva nurvula	17.4	0	36	24	24	44
Ficus longifolia	34.2	52	20	0	66	27
Mimusops elengi	20.3	19	23	33	19	80
Cassia javanica	11	19	23	56	10	56
Balanites aegyptiaca	79.8	39	18	24	34	32
Trifera sp.	11.8	21	29	33	44	44
Bambusa sp.3	17.8	28	38	82	0	45
Aegle marmelos	10.5	15	36	0	49	83
Bambusa sp.4	24.7	29	12	90	0	50
Mangifera indica	38.6	15	38	56	0	63
Abrus sp.2	10.8	16	35	6	22	60
Cucurbita sp.	36.6	8	30	14	24	100
Abrus sp.1	22.6	40	44	55	8	32
Vernonia elaeagnifolia	46.8	32	14	49	16	56
Akebia quinata	17.5	31	13	43	24	35
Abrus sp.3	14.9	44	38	60	12	26
Coccinia grandis	12.8	28	22	79	9	25
Malpighia emarginata	17.4	10	36	24	24	44
Occimum sp.1	35.3	46	23	23	23	25
Occimum sp.3	99.7	55	16	43	10	48
Tecoma stans	26.6	52	25	31	32	33
Duranta erecta	14.5	52	12	56	27	18
Holarrhena antidysenterica	17.4	10	36	24	24	44
Cassia tora	25	25	14	7	48	15
Indigofera coerulea	64.4	16	10	16	43	36
Ixora coccinea	38.6	15	38	56	0	63

Ixora sp.	23.5	11	33	56	21	60
Cactus sp.	20.7	27	31	11	47	26
Artabotrys hexapetalus	24.9	72	24	0	78	24
Mimosa pudica	20.8	14	30	24	21	44
Occimum sp.2	18.3	21	23	21	54	63
Cleome viscosa	25	25	14	7	48	23
Striga angustifolia	20.8	14	20	24	21	44
Eclipta alba	99.7	55	16	43	10	48
Merremia gangetica	19.5	25	24	24	31	53
Martynia diandra	99.7	83	22	0	47	56
Bryophyllum pinnatum	15.1	25	31	38	18	50
Aerva javanica	30.4	20	20	22	54	54

Among the shrubs out of 12 species, three have a poor-quality protein model. These species are, *Duranta erecta*, *Cassia tora,* and *Cleome viscosa.* In *Duranta erecta* the parameters sequence identity-18%, confidence score-14.5%, coverage-52%, disorder-12%, alpha helix-56% and beta strand-27%. In *Cassia tora* sequence identity-15%, confidence score-25%, coverage-25%, disorder-14%, alpha helix-7%, and beta strand-48%. In C*leome viscosa* the different parameters sequence identity-23%, confidence score-25%, coverage-25%, disorder-14%, alpha helix-7%, and beta strand-48% (Table 3).

Out of 9 herb species, 8 species have a good protein model which means all species have <20% sequence identity except *Occimum* sp.1. For *Occimum* sp.1 the different parameters are sequence identity-25%, confidence score-35.3%, coverage-46%, disorder-23%, alpha helix-23% and beta strand-23%.

Thus, a total of 40 plants out of 46 showed reliable protein prediction.

3.2. Protein physico-chemical analysis using ProtParam

3.2.1. Amino acid composition

The amino acids percentage in the *psbB* of 46 different plant samples was found by using the ProtParam Tool from the Expasy Proteomic Server (Figure 1). Among all different amino acids the maximum percentage was observed in Leucine (L) (3-17%), Proline (P) (0-16%), Glycine (G) (0-10%) and the minimum percentage in Methionine (M) (0-5%), Asparagine (N) (0-6%), Aspartic acid (D) (0-6%), The percentage of amino acids observed in each of the organisms is as follows:

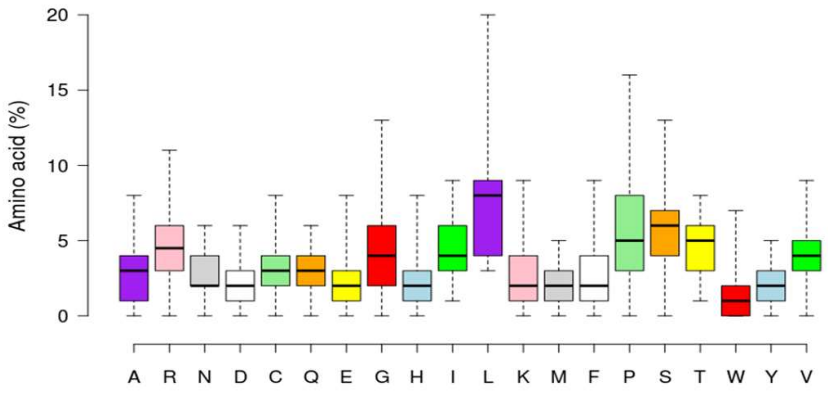


Figure 1: *BOX plot showing the amino acid frequency (%) information for the 46 psbB protein sequences of plants all taken together.*

Alanine (A) was found to be highest in *Abrus* sp.3 (8%) and lowest in *Bambusa* sp.1, *Ficus longifolia*, *Malpighia emarginata, Occimum* sp.1, *Tecoma stans, Cleome viscosa,* and *Bambusa* sp.4 (1%). Arginine (R) was found to be highest in *Cactus* sp. (11%) and lowest in *Abrus* sp.2, *Balanites aegyptiaca, Striga angustifolia,* and *Aegle marmelos* (1%). Asparagine (N) was found to be highest in *Bambusa* sp.1 (6%) and lowest in *Artocarpus heterophyllus, Mimusops elengi, Duranta erecta, Bryophyllum pinnatum, Coccinia grandis, Bambusa* sp.3 and *Aegle marmelos* (1%). Aspartic acid (D) was found to be highest in *Cassia fistula* (6%) and lowest in *Mangifera indica, Tamarindus indica, Mangifera indica, Crataeva nurvula, Malpighia emarginata, Occimum* sp.1, *Tecoma stans, Holarrhena antidysenterica, Ixora coccinea, Mimosa pudica, Occimum* sp.2, *Martynia diandra, Aerva javanica, Ficus longifolia, Balanites aegyptiaca, Cucurbita* sp. and *Akebia quinata* (1%). Cysteine (C) was found to be highest in *Balanites aegyptiaca* (8%) and lowest in *Ficus longifolia, Tecoma stans, Mimosa pudica, Cleome viscosa, Martynia diandra, Bryophyllum pinnatum, Aegle marmelos* and *Bambusa* sp.4 (1%). Glutamine (Q) was found to be highest in *Borassus flabellifer* (6%) and lowest in *Mangifera indica, Cassia fistula, Tamarindus indica, Bambusa* sp.2, *Abrus* sp.2, *Crataeva nurvula, Cassia javanica, Vernonia elaeagnifolia, Malpighia emarginata, Occimum* sp.1, *Duranta erecta, Cassia tora, Aerva javanica, Cleome viscosa, Aegle* marmelos and Bambusa sp.4 (1%).Glutamic (E) was found to be highest in *Trifera* sp. (8%) and lowest in *Artocarpus heterophyllus, Borassus flabellifer L., Malpighia emarginata, Occimum* sp.1, *Merremia gangetica, Aerva javanica, Occimum* sp. 2, *Tamarindus indica, Balanites aegyptiaca, Abrus* sp.1 and *Abrus* sp.3 (1%). Glycine (G) was found to be highest in *Ixora* sp. (13%) and lowest in *Mangifera indica, Ixora coccinea, Cleome viscosa, Mimusops elengi* and *Akebia quinata* (1%). Histidine (H) was found to be highest in *Borassus flabellifer* (8%) and lowest in *Tamarindus indica, Haldina cordifolia,* in *Cassia tora, Ixora* sp., *Aerva javanica, Jobra, Mimosa pudica, Cleome viscosa, Bambusa* sp.2, *Ficus longifolia, Abrus* sp.1, *Vernonia elaeagnifolia, Bambusa* sp.3 and *Bambusa* sp.4 (1%). Isoleucine (I) was found to be highest in *Artocarpus heterophyllu* and *Coccinia grandis* (9%) and lowest in *Cassia tora*, *Duranta erecta*, *Ixora coccinea*, *Cactus* sp., *Artabotrys hexapetalus* and *Akebia quinata* (1%). Leucine (L) was found to be highest in *Cassia javanica* (17%) and lowest in *Tamarindus indica, Abrus* sp.2, *Mimusops elengi, Bambusa* sp.3, *Mangifera indica, Merremia gangetica, Martynia diandra, Ixora coccinea, Bambusa* sp.4 and *Akebia quinata* (3%). Lysine (K) was found to be highest in *Ixora* sp. (9%) and lowest in *Bambusa* sp.2, *Striga angustifolia, Duranta erecta, Indigofera coerulea, Aerva javanica, Trifera sp.* and *Abrus sp.3* (1%). Methionine (M) was found to be highest in *Bambusa* sp.2 (5%) and lowest in *Borassus flabellifer, Abrus* sp.2, *Striga angustifolia, Indigofera coerulea, Martynia diandra, Cactus* sp., *Jobra, Occimum* sp.2, *Cleome viscosa,*

Mimusops elengi and *Aegle marmelos* (1%). Phenylalanine (F) was found to be highest in *Ixora* sp. (9%) and lowest in *Haldina cordifolia, Ficus longifolia, Striga angustifolia, Tecoma stans, Martynia diandra, Coccinia grandis, Balanites aegyptiaca, Cucurbita* sp. and *Vernonia elaeagnifolia* (1%). Proline (P) was found to be highest in *Artocarpus heterophyllus* (16%) and lowest in *Trifera* sp., *Bambusa* sp.3 and *Bambusa* sp.4 (1%). Serine (S) was found to be highest in *Eclipta alba* and *Occimum* sp.3 (13%) and lowest in *Cleome viscosa* (1%). Threonine (T) was found to be highest in *Cassia fistula, Mangifera indica*, *Tecoma stans*, *Martynia diandra*, *Ixora coccinea*, *Coccinia grandis,* and *Akebia quinata* (8%) and lowest in *Indigofera coerulea, Ficus longifolia* (1%). Tryptophan (W) was found to be highest in *Eclipta alba* and *Occimum* sp.3, (7%) and lowest in *Artocarpus heterophyllus, Mangifera indica, Haldina cordifolia, Crataeva nurvula, Mimusops elengi, Coccinia grandis, Duranta erecta, Holarrhena antidysenterica, Ixora* sp., *Bryophyllum pinnatum, Occimum sp. 2, Trifera* sp., *Cucurbita* sp., *Abrus* sp.1 and *Bambusa* sp.4 (1%). Tyrosine (Y) was found to be highest in *Borassus flabellifer, Mimusops elengi, Cassia javanica* (5%) and lowest in *Artocarpus heterophyllus, Mangifera indica*, *Striga angustifolia*, *Eclipta alba*, *Occimum* sp.3, *Duranta erecta*, *Ixora* sp., *Cleome viscosa, Bambusa* sp.3, *Aegle marmelos* and *Akebia quinata* (1%). Valine (V) was found to be highest in *Duranta erecta* (9%) and lowest in *Cucurbita* sp. *Merremia gangetica* and *Aegle marmelos* (1%) (Figure 1).

3.2.2. pI value

pI values vary from 11.53 (*Ixora* sp.) to 3.91 (*Bambusa* sp.) (Figure 2). In 39 plants pI values ranged 7.11 to 11.52 and remains plants pI value was less than 7 and the minimum value was 3.91 (Figure 2). This variation in the pI values may be the result of changes in amino acid composition in their proteins (Figure 2).

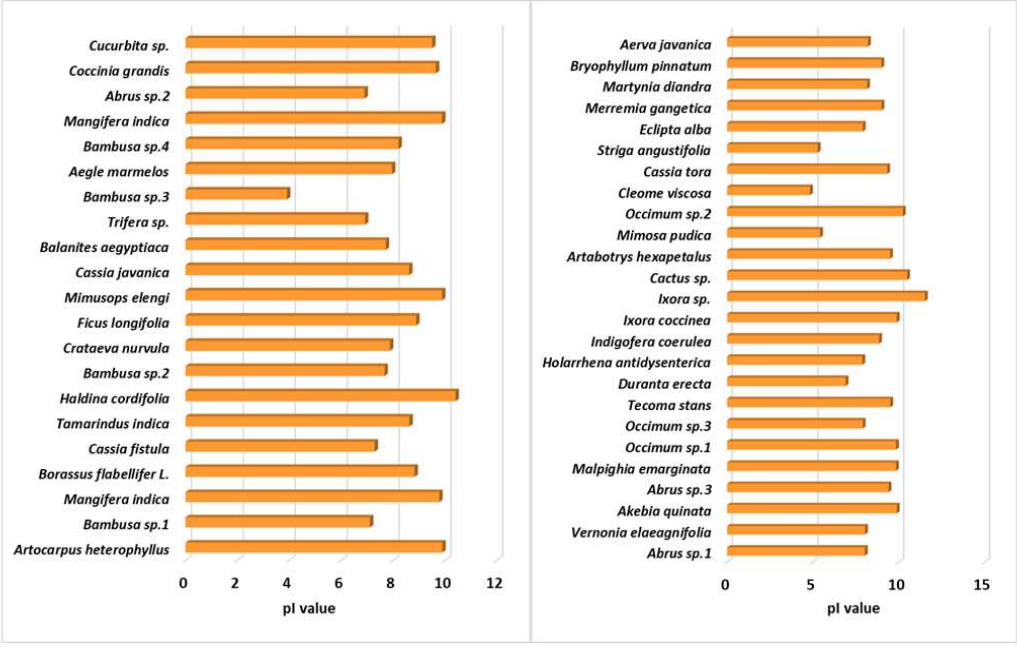


Figure 2: *pI value of 46 psbB protein sequence of selected plants.*

3.2.3. Aliphatic index

The aliphatic index in 46 different plant samples was found to be in the range of 45-137.44. In tree aliphatic index ranged from 45-116.74. The highest value was observed in *Cassia javanica* and the lowest in *Mimusops elengi.* In vine aliphatic index ranged from 58.08-114.27. The highest value was observed in *Vernonia elaeagnifolia* and the lowest in *Akebia quinata.* In the shrub aliphatic index ranged from 58.08-137.44. The highest value was observed in *Cleome viscosa* and the lowest in *Ixora coccinea.* In herb, aliphatic index ranged from 64.58-112.86 (Figure 3). The highest value was observed in *Striga angustifolia* and the lowest in *Merremia gangetica.* The aliphatic index (AI) which is defined as the relative volume of a protein occupied by aliphatic side chains is regarded as a positive factor for the increase of thermal stability of globular proteins. The very high aliphatic index of all sequences indicates it may be stable for a wide temperature range.

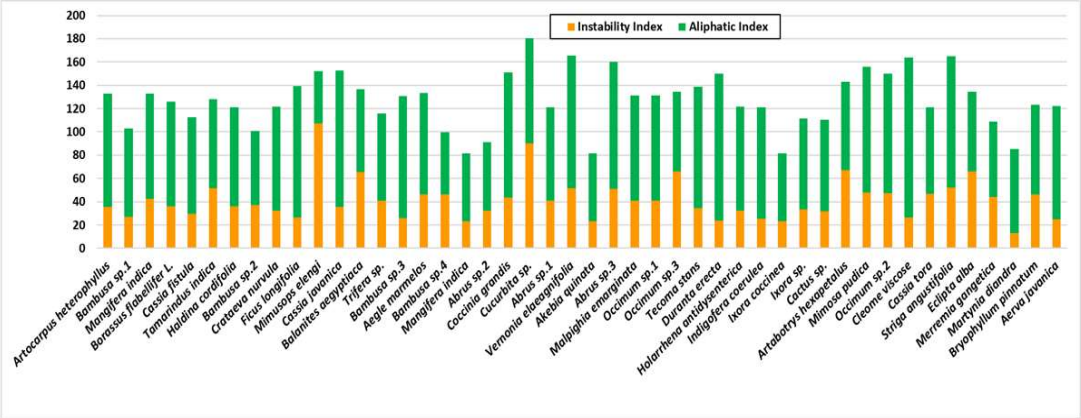


Figure 3: *Instability index and aliphatic index of psbB proteins sequences.*

3.2.4. Instability index

The instability index from all 46 different plants was found to be between 107.39 (unstable) to 13.16 (stable). The instability index value above 40 confirms the instability nature but the value of less than 40 reveals the stability nature of the overall 3D structure of the protein [11]. In tree, instability index was found from 23.34 to 107.39. The highest value was observed in *Mimusops elengi* and the lowest in *Mangifera indica.* In vine species instability index was found from 23.34 to 90.33. The highest value was observed in *Cucurbita* sp. and the lowest in *Akebia quinata.* In shrub instability index was found from 23.34 to 66.87. The highest value was observed in *Artabotrys hexapetalus* and the lowest in *Ixora coccinea.* In herb instability index was found from 13.16 to 66.14. The highest value was observed in *Eclipta alba* and the lowest in *Martynia diandra* (Figure 3).

3.2.5. Grand average of hydropathicity (GRAVY)

The Grand average of hydropathicity (GRAVY) in 46 plant samples was found to be in the range of -0.073-0.858. In tree, GRAVY ranged from -0.767 to 0.728. The highest value was observed in *Bambusa* sp.3 and the lowest in *Mangifera indica.* In vine, GRAVY ranged from -0.073 to 0.133. The highest value was observed in *Abrus* sp.1 and the lowest in *Vernonia*

elaeagnifolia. In shrub GRAVY ranged from -0.767 to 0.858. The highest value was observed in *Duranta erecta* and the lowest in *Ixora coccinea.* In herb, GRAVY ranged from -0.504 to 0.638 (Figure 4). The highest value was observed in *Aerava javanica* and the lowest in *Bryophyllum pinnatum.* The low value indicates the possibility of better interaction with water [10].

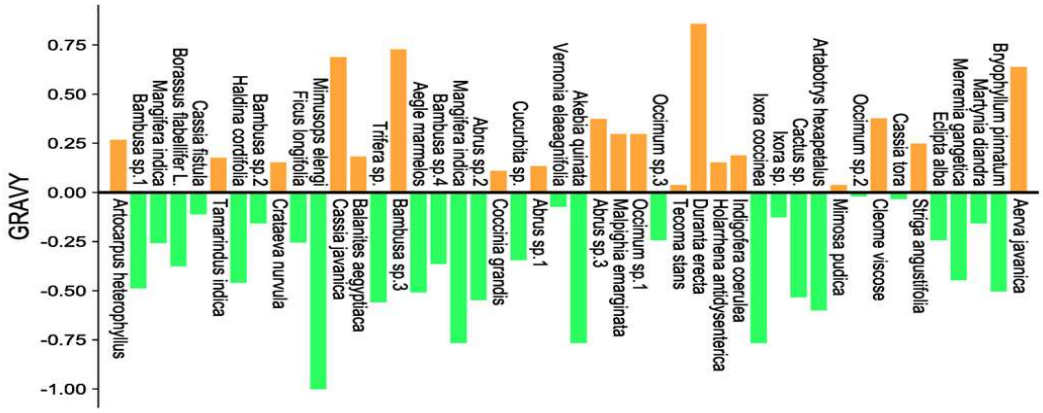


Figure 4: *Grand Average of Hydropathicity (GRAVY) of psbB proteins sequences.*

3.3. Ramachandran plot analysis using rampage

The Ramachandran plot indicates the acceptable quality of a protein model (Table 4). Ramachandran created two-dimensional (2D) scatter plots of Ø and Ψ pairs, comparing them to a predicted distribution. These scatter plots are now commonly known as Ramachandran plots [9]. Simple polymer physics models can be used to make a predicted distribution of pairs of Ø and Ψ angles using a volume exclusion model: no two non-bonded atoms can overlap. Results of such calculations show considerable conformational freedom in the two torsion angles but with many clear restrictions. Deviations from the expected distribution for a new protein structure can now be used to judge the quality of that structure. Rampage derives Phi/Psi plots for Gly, Pro, Pre-Pro, and other residues and the plot is divided into three regions that are favored, allowed, and outlier regions. The number of residues in the favored region (>90%) is the measure of the good quality of a model.

Table 4: *Ramachandran plot value obtained through RAMPAGE (the threshold value of favorured region >90 % indicate appropriate 3-D structure).*

Plant Name	No. of residues in favoured region	No. of residues in allowed region	No. of residues in outliner region
Artocarpus heterophyllus	86.4 %	6.2 %	7.4 %
Bambusa sp.1	88.5 %	4.6 %	6.9 %
Mangifera indica	76.8 %	11 %	12.2 %
Borassus flabellifer L.	78.4 %	9.1 %	12.5 %
Cassia fistula	80 %	12.5 %	7.5 %
Tamarindus indica	90.6%	3.1 %	6.2 %

Haldina cordifolia	85.5 %	14.5 %	0%
Bambusa sp.2	84.7%	8.6 %	3.7 %
Crataeva nurvula	81.8%	12.5 %	5.7 %
Ficus longifolia	85.3 %	10.4 %	6.2 %
Mimusops elengi	84 %	10 %	6 %
Cassia javanica	79.8 %	7.1 %	13.1 %
Balanites aegyptiaca	85.5 %	10.1 %	4.3 %
Trifera sp.	88.2 %	8.2%	5.9 %
Bambusa sp.3	100 %	0 %	0 %
Aegle marmelos	78.4 %	13.5 %	8.1 %
Bambusa sp.4	97.9 %	2.1%	0 %
Mangifera indica	84 %	8 %	8 %
Abrus sp.2	68.9 %	18 %	13.1 %
Coccinia grandis	95.2 %	3.6 %	1.2 %
Cucurbita sp.	68.9 %	13.5 %	17.6 %
Abrus sp.1	94.7 %	4 %	1.3 %
Vernonia elaeagnifolia	83.6 %	10.9 %	5.5 %
Akebia quinata	84.9 %	12.3 %	2.7 %
Abrus sp.3	77 %	13.1 %	9.8 %
Malpighia emarginata	77.6 %	13.4 %	9 %
Occimum sp.1	77.6 %	13.4 %	9 %
Occimum sp.3	88.8 %	10.1 %	1.1 %
Tecoma stans	88.2 %	10.6 %	1.2 %
Durantaerecta	88.7 %	6.5 %	4.8 %
Holarrhena antidysenterica	88.6 %	4.5 %	6.8 %
Indigofera coerulea	84 %	8 %	8 %
Ixora coccinea	81.8 %	11.4 %	6.8 %
Ixora sp.	85.9 %	8.2 %	5.9 %
Cactus sp.	77.9 %	10.3 %	11.8 %
Artabotrys hexapetalus	79.5 %	13.6 %	6.8 %
Mimosa pudica	77 %	11.5 %	11.5 %
Occimum sp.2	94.6 %	2.7 %	2.7 %
Cleome viscose	73.5 %	12 %	14.5 %
Cassia tora	87.3 %	7.3 %	5.5 %
Striga angustifolia	77 %	13.1 %	9.8 %
Eclipta alba	88.8 %	10.1 %	1.1 %
Merremia gangetica	91.2 %	3.5 %	5.3 %
Martynia diandra	76.8 %	16.1 %	7.1 %
Bryophyllum pinnatum	82 %	5 %	13.5 %
Aerva javanica	76.2 %	14.3 %	9.5 %

In the present study, in the trees, from a total of 18 species, only 3 showed a significant value. Species that have a significant value are *Tamarindus indica* (90.60%), *Bambusa* sp.3 (100%), and *Bambusa* sp.4 (97.90%). Other species *Artocarpus heterophyllus* (86.40%), *Bambusa* sp.1 (88.50%), *Adina cordifolia* (85.50%), *Bambusa* sp.2 (87.70%), *Crataeva nurvula* (81.80%), *Ficus longifolia* (83.30%), *Mimusops elengi* (84%), *Balanites aegyptiaca* (85.50%), *Trifera* sp. (88.20%) and *Mangifera indica* (84%) have a favored region value and are near the threshold value.

Of the 7 vine species, 2 showed a significant favored region. From these two vine species, the highest favored region was observed in *Coccinia grandis* (95.20%) and *Abrus sp.1* (94.70%) and 2 species had a favored region value near the threshold value like *Vernonia elaeagnifolia* (83.60%) and *Akebia quinata* (84.90%).

Out of 12 shrub species, only 1 species showed a significant favored region and hence concluded that the protein model is good which was observed in *Occimum* sp.2 (94.60%). Other species *Occimum* sp.3 (88.80%), *Tecoma stans* (88.20%), *Duranta erecta* (88.70%), *Holarrhena antidysenterica* (88.60%), *Cassia tora* (87.30%), *Indigofera coerulea* (84%), *Ixora coccinea* (81.80%) and *Ixora* sp. (85.90%) had a favored region value are near the threshold value.

Out of 9 herb species, 1 showed a significant value in *Merremia gangetica* (91.20%) then *Eclipta alba* (88.80%) and *Bryophyllum pinnatum* (82%) had near the significant favored region.

3.4. Calculation of protein energy level using Anolea

In the present study, out of 18 tree species, 14 species have a good protein structure level all species are *Artocarpus heterophyllus* (60), *Bambusa* sp.1 (69), *Mangifera indica* (65), *Borassus flabellifer L.* (57), *Cassia fistula* (67), *Tamarindus indica* (42), *Adina cordifolia* (45), *Bambusa* sp.2 (61), *Crataeva nurvula* (67), *Ficus longifolia* (45), *Mimusops elengi* (41), *Cassia javanica* (52), *Balanites aegyptiaca* (40) and *Trifera* sp. (66). Same as above, out 7 vine species all have a good protein structure level viz. *Abrus* sp.2 (47), *Cucurbita* sp. (52), *Abrus* sp. (43), *Vernonia elaeagnifolia* (38), *Akebia quinata* (47), *Abrus* sp.3 (28) and *Coccinia grandis* (47) (Figure 5).

Out of 12 shrub species, 10 species have a good protein structure level which is *Malpighia emarginata* (59), *Tecoma stans* (48), *Holarrhena antidysenterica* (56), *Cassia tora* (46), *Ixora coccinea* (67), *Ixora* sp. (46), *Cactus* sp. (48), *Artabotrys hexapetalus* (41), *Mimosa pudica* (35) and *Cleome viscosa* (66). Out of 6 herb species, 4 species have a good protein structure level all species are *Eclipta alba* (62), *Martynia diandra* (37), *Bryophyllum pinnatum* (38), and *Aerva javanica* (57).

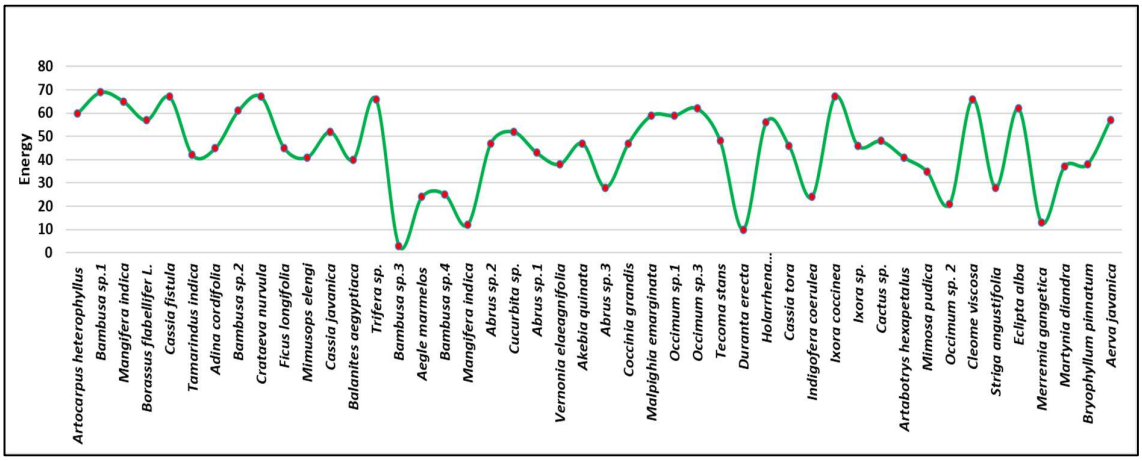
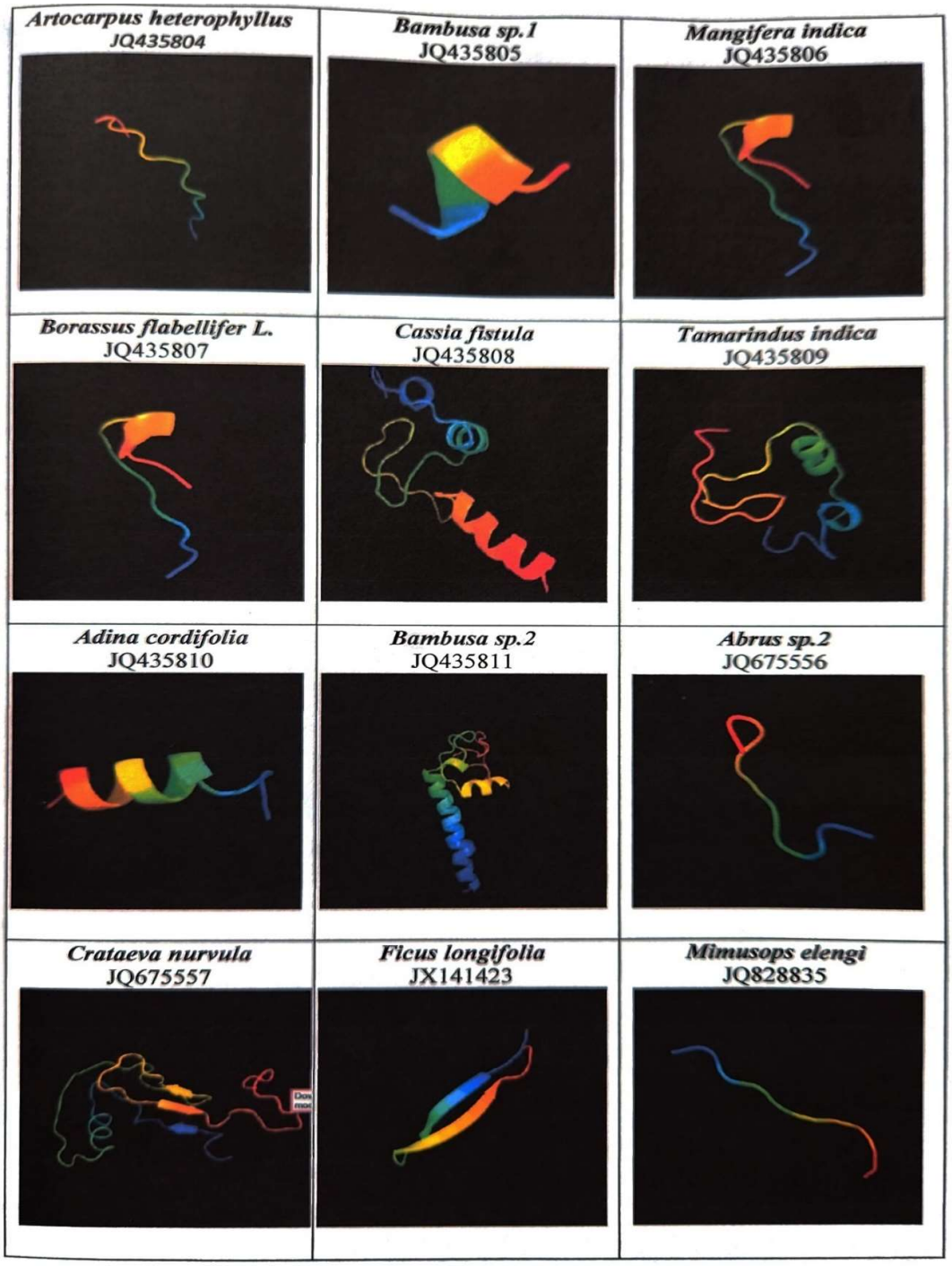
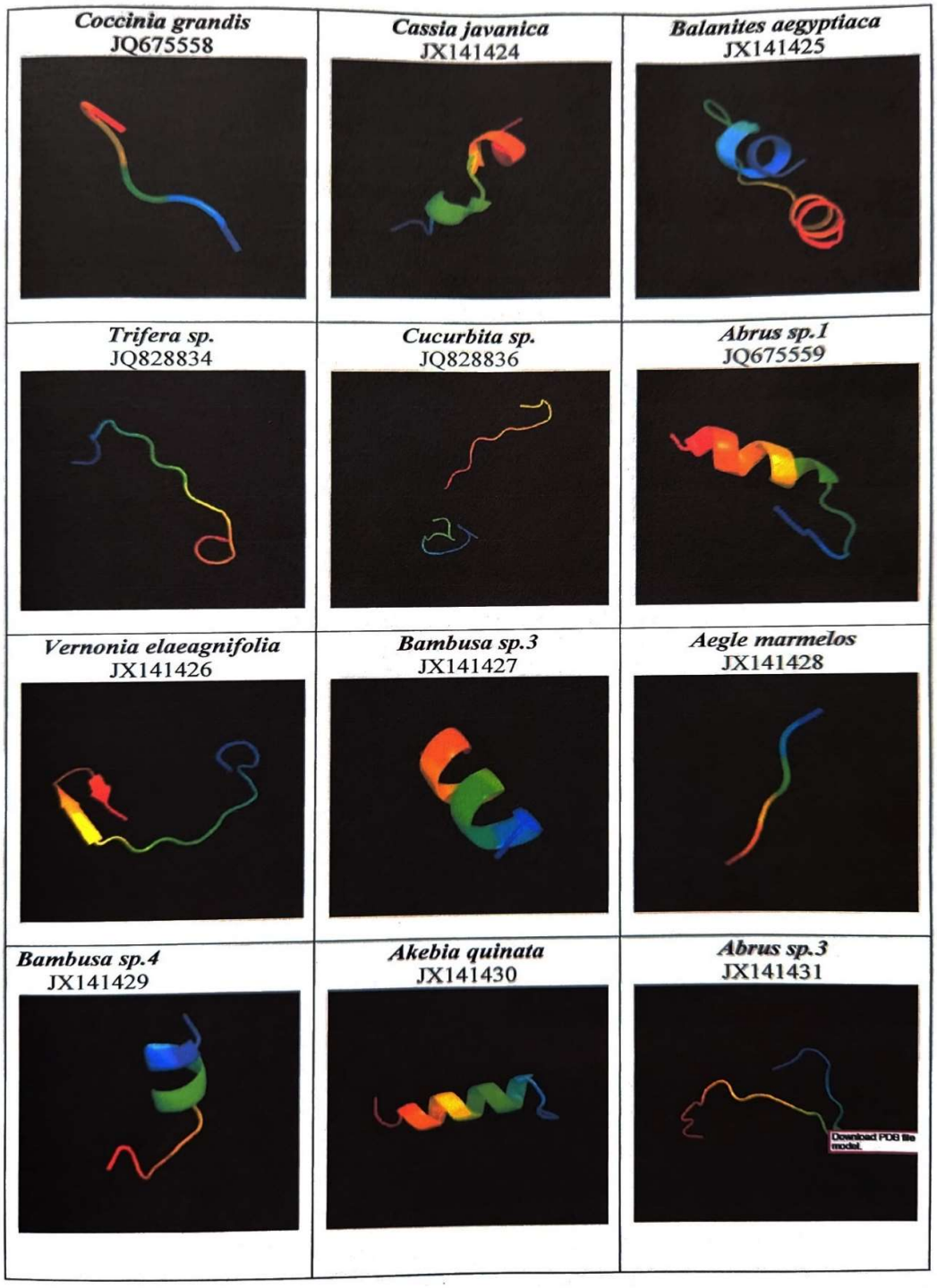


Figure 5: *Calculate the energy level of the protein chain using Anolea.*

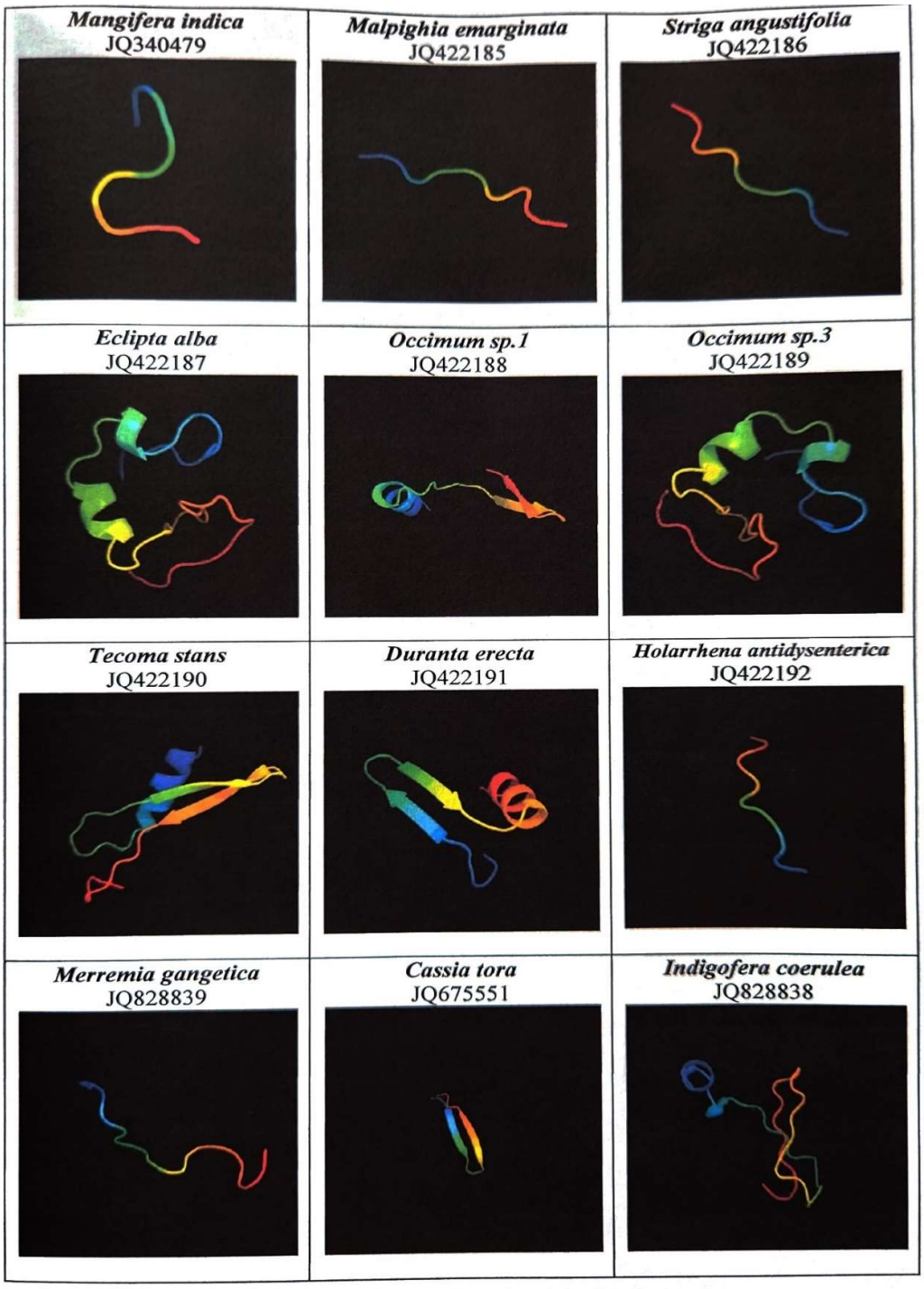
The protein structure prediction by using two different algorithms; Pyre2 showed more numbers of the sequences prediction and its confirmation using Rampage than that of CPM. The predicted structures are given in Figure 6. *psbB* gene is a key molecule of water splitting process of photosynthesis and thus understanding its structural and function aspects may play a decisive role in future of renewable energy [19]. In this study, structural variations in the predicted proteins are observed irrespective of their habitat i.e., tree, vine, shrub or herb. The protein folding mechanism involves very complex dynamics [20] and unknown energy factors [21-23]. The homology modeling approach of protein structure prediction in this study offers similarity index with the known protein(s); and hence to evaluate functional ability and mutation or any other variations for biological performance. Further, the basic properties of these proteins vary in all habitats e. g. its composition, instability index, suggesting that the turnover rate of this protein may vary in these plants (Figures 1-4). The observed changes in the protein *psbB* help design and comprehension of protein function.


Artocarpus heterophyllus
JQ435804
Bambusa sp.1
JQ435805
Mangifera indica
JQ435806
Borassus flabellifer L.
JQ435807
Cassia fistula
JQ435808
Tamarindus indica
JQ435809
Adina cordifolia
JQ435810
Bambusa sp.2
JQ435811
Abrus sp.2
JQ675556
Crataeva nurvula
JQ675557
Ficus longifolia
JX141423
Mimusops elengi
JQ828835

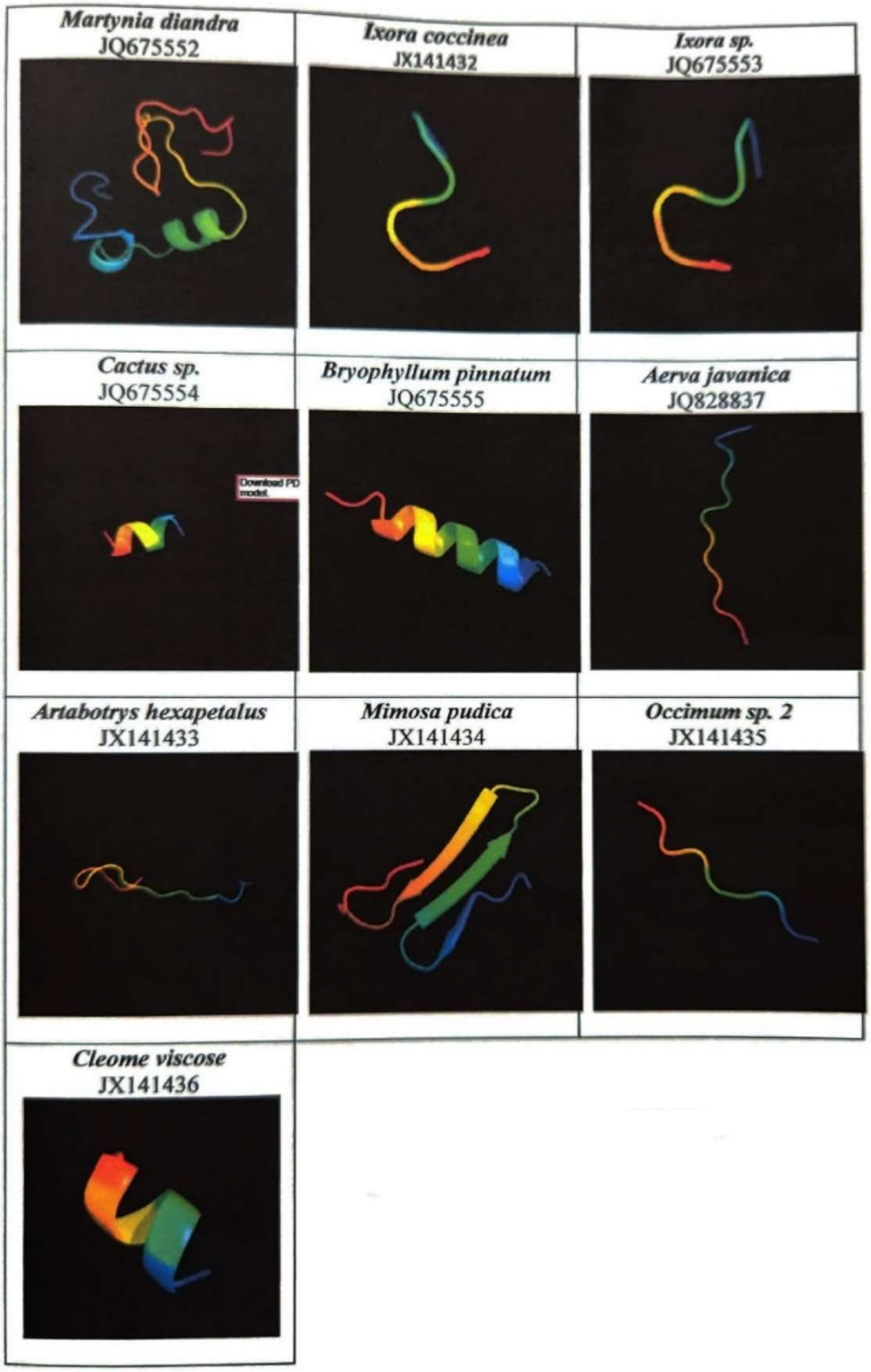
A


Coccinia grandis
JQ675558
Cassia javanica
JX141424
Balanites aegyptiaca
JX141425
Trifera sp.
JQ828834
Cucurbita sp.
JQ828836
Abrus sp.1
JQ675559
Vernonia elaeagnifolia
JX141426
Bambusa sp.3
JX141427
Aegle marmelos
JX141428
Bambusa sp.4
JX141429
Akebia quinata
JX141430
Abrus sp.3
JX141431

B


Mangifera indica
JQ340479
Malpighia emarginata
JQ422185
Striga angustifolia
JQ422186
Eclipta alba
JQ422187
Occimum sp.1
JQ422188
Occimum sp.3
JQ422189
Tecoma stans
JQ422190
Duranta erecta
JQ422191
Holarrhena antidysenterica
JQ422192
Merremia gangetica
JQ828839
Cassia tora
JQ675551
Indigofera coerulea
JQ828838

C



D

Figure 6: *Protein structure prediction using Pyre2 online bioinformatics tool for the 46 plants studied from semi-arid region of western India.*

4. Conclusion

Although it is an approach for the computational protein structure prediction energy levels evaluated using ANOLEA showed the majority of the proteins with better structure configuration; however, it should need to be verified biologically prior to their usage as references or standards. It is important to note that structure similarity studies revealed that there is no influence of plant habit (tree/herb/shrub/vine) on protein structure and any two-group way has more relatedness in the *psbB* proteins. Further from the above study, it is clear that protein Pyre2 methods showed more numbers of reliable protein structure for the plant species studied. In general, it is important to evaluate protein properties and structure for the functional assignment at the physiological level. The availability of various tools with a range of threshold values helps researchers to understand and predict protein fold. Validation of these models using the Ramachandran plot helps in the understanding of predicted protein structure.

Acknowledgement

Authors are thankful to Higher Education Department, Gandhinagar and Department of Biosciences, Saurashtra University, Rajkot, Gujarat, India for providing lab facilities.

References

1. Stirbet A, Lazár D, Guo Y, et al. Photosynthesis: basics, history and modelling. Ann Bot. 2020;126:511-37.

2. Blankenship RE. Molecular Mechanisms of Photosynthesis. Wiley-Blackwell, New Jersey, USA. 2002.

3. Schelvis JPM., van Noort PI, Aartsma TJ, et al. Energy transfer, charge separation and pigment arrangement in the reaction center of Photosystem II. Biochim Biophys Acta-Bioenerg. 1994;1184:242-50.

4. Bricker TM, Frankel LK. The structure and function of CP47 and CP43 in photosystem II. Photosynth Res. 2002;72:131-46.

5. Bricker TM. The structure and function of CPa-1 and CPa-2 in photosystem II. Photosynth Res. 1990;24:1-13.

6. Martí-Renom MA, Stuart AC, Fiser A, et al. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29:291-325.

7. Al-Lazikani B, Jung J, Xiang Z, et al. Protein structure prediction. Curr Opin Chem Biol. 2001;5:51-6.

8. Petersen TN, Lundegaard C, Nielsen M, et al. Prediction of protein secondary structure at 80% accuracy. Proteins. 2000;41:17-20.

9. Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics.2005;21:951-60.

10. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol. 2003;334:793-802.

11. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11-5.

12. Kelley LA, Mezulis S, Yates CM, et al. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10:845-58.

13. Petersen B, Petersen TN, Andersen P, et al. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol. 2009;9:51.

14. Bennett-Lovsey RM, Herbert AD, Sternberg MJ, et al. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins. 2008;70:611-25.

15. Ramachandran GN. Protein structure and crystallography. Science.1963;141:288-91.

16. Sivakumar K, Balaji S, Gangaradhakrishna. In silico characterization of antifreeze proteins using computational tools and servers. J Chem Sci. 2007;119:571-9.

17. Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence. Protein Eng. 1990;4:155-61.

18. Gasteiger E, Hoogland C, Gattiker A, et al. Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed), The Proteomics Protocols Handbook. Humana Press, New Jersey, USA. 2005;pp.571-607.

19. Sundaram S, Tripathi A, Gupta V. Structure prediction and molecular simulation of gases diffusion pathways in hydrogenase. Bioinformation. 2010;5:177-83.

20. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct. 2001;30:173-89.

21. Rohl CA, Strauss CE, Misura KM, et al. Protein structure prediction using Rosetta. Methods Enzymol. 2004;383:66-93.

22. Lee J, Wu S, Zhang Y. Ab initio protein structure prediction. In: Rigden DJ (ed), From Protein Structure to Function with Bioinformatics. Springer, Dordrecht. 2009;pp.3-25.

23. Xia Y, Huang ES, Levitt M, et al. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol. 2000;300:171-85.