

INTERNSHIP REPORT

**Structural conservation of very distantly
related proteins:
A trip into the twilight zone**

Sumedha J. Gholba
Computational Biosciences,
Arizona State University,
Tempe, AZ 85287-1604 USA

Advisor: Dr. Robert E. Blankenship
Department of Chemistry and Biochemistry
And
Center for the Study of Early Events in Photosynthesis,
Arizona State University,
Tempe, AZ 85287-1604 USA

An internship report presented
In partial fulfillment of requirement for the
Degree of Master of Science.

Report Number: 05-01

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Dr. Robert E. Blankenship for his guidance and support for this research. I could not have asked for a better role model to follow as a researcher and as a person. I'm also deeply appreciative to Dr. Jason Raymond and Dr. Rosemary Renaut for their help and encouragements as I began my career in the field of bioinformatics.

Many thanks to the members of the Blankenship Lab for the countless help they rendered me for the research. I would like to thank Dr. Robert E. Blankenship, Dr. Rosemary Renaut and Dr. Jeffery Touchman for being my graduate committee members.

ABSTRACT

Photosynthesis was established on the earth more than 3 billion years ago. All available evidence suggests that the earliest photosynthetic organisms were anoxygenic and that oxygen-evolving photosynthesis is a more recent development. The reaction center complexes are integral membrane pigment-proteins that span the membrane in vectorial fashion to carry out electron transfer reactions. The origin and extent of distribution of these proteins has been perplexing from a phylogenetic point of view, mostly because of extreme sequence divergence. Our study of both the sequences and structures of reaction center proteins from both the oxygenic photosystems and the anoxygenic reaction centers from bacteria reveal the conservation of certain key amino acids, as well as a surprisingly well conserved overall protein fold. The multiple sequence alignment and phylogenetic trees created from the proteins show high degree of conserved regions in photosystem-II and bacterial reaction center-II, implying common genealogy. Also, the similarity between photosystem-I heterodimers and reaction center I homodimer proteins, indicate them having a single precursor. It is seen that even though L-M and D1-D2 show a similar pattern of evolution with gene duplication, L-M proteins show a step-by-step diversification whereas the other branch bifurcates into D1s and D2s just at the end. The reaction center I homodimer is placed nearly at the center between the photosystem I and II portions of the tree, suggesting it to be perhaps closer to an ancestral type of reaction center. The structural alignment of these proteins depicts five well-

aligned α -helices. Their structures show a striking amount of similarity in the hydrophobic domains forming the transmembrane helices, very strongly suggesting that all photosynthetic reaction centers are derived from a single ancestral form. The overall structural architecture of photosystems was apparently arrived at very early in their evolution and has changed remarkably little, despite extensive sequence divergence and functional specialization of the various types of photosystems.

Keywords: Twilight zone, reaction center proteins, alignments, hydropathy, phylogenetic divergence.

TABLE OF CONTENTS

1. INTRODUCTION	6
2. METHODS	11
2.1 Selection of Sequences	11
2.2 Sequence alignment	12
2.3 Phylogenetic Tree building	13
2.4 Structural alignment.....	14
2.5 Computation of RMS distances	15
2.6 Hydropathy plots.....	17
2.7 Identity and RMSD plots	17
3. RESULTS	21
4. SEQUENCE ALIGNMENTS.....	28
5. REFERENCES	32

1. INTRODUCTION

Photosynthesis – one of the most vital and complex physiological processes; which is brought about million times every second on this planet. The biochemical processes involved in photosynthesis are exceptionally complex, and yet an individual, diminutive cell can perform the complicated tasks many times over. Studies about this wonder-phenomenon started as early as 1772, when an English chemist, Joseph Priestley demonstrated that plants immersed in water give off a gas (oxygen), and that this gas is necessary for animals to live. Despite the discoveries that followed Priestley’s work, we still do not know some of the secrets of photosynthesis, even to this day and photosynthesis remains a mystery in fields like evolutionary studies.

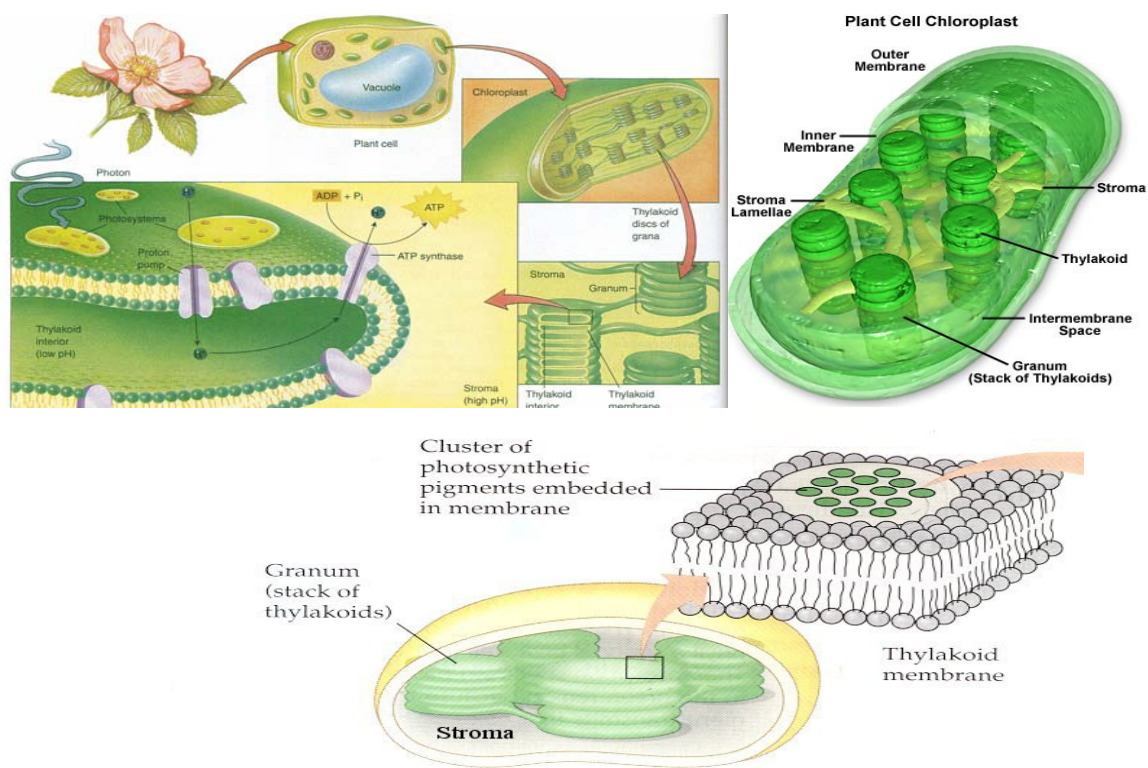


Figure 1. Basic photosynthetic machinery

Adapted from <http://kvhs.nbed.nb.ca/gallant/biology/biology.html>.

In the case of a plant with leaves, the light strikes the leaf, penetrates the outer covering of the plant's cells, and arrives at the first checkpoint of photosynthesis—the chloroplast. Chloroplasts, which specialize in photosynthesis, are often oval or disk-shaped organelles about two to ten micrometers long and have a double-membrane system. In the most common types, the inner membrane is the site where sunlight energy is trapped, and where *adenosine triphosphate (ATP)* is produced. The inner membrane is arranged as a system of stacked disks, called *grana*, which are surrounded by a semifluid matrix called the *stroma*. It is like a miniscule, bean-shaped structure that contains stacks of miniature pancakes inside of it. The individual pancakes that make up the *grana* are called *thylakoid disks* (Figure 1). It is at this location where photosynthesis does much of its work.

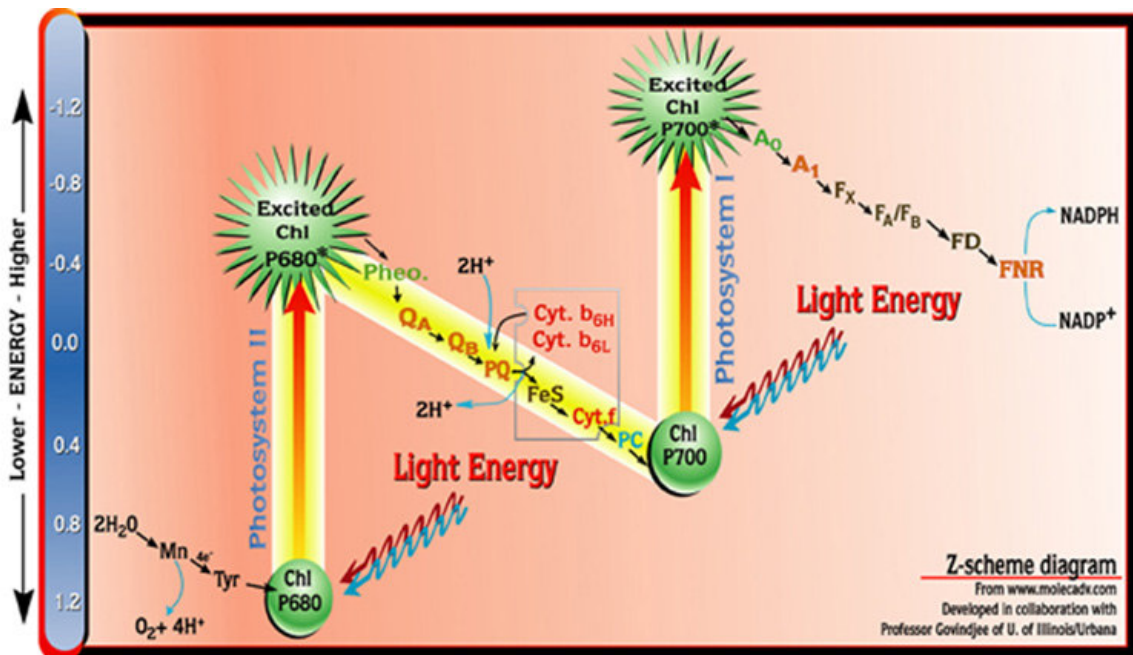


Figure 2. Basic photosynthetic light reactions, also called as Z-scheme.

Adapted from <http://www.life.uiuc.edu/govindjee/ZSchemeG.html>

The outer portion of the thylakoid disks is where light becomes useful to the plant. There are many structures that function in absorbing the energy and converting it to a form that is useful to the cells. The complexity of energy synthesis soon becomes almost inconceivable. Once the pure light energy reaches the thylakoid membrane, it starts a chain reaction that ultimately will result in the production of energy for the cell. The beginning of these reactions constitutes the electron transport chain. The light reaction begins with photosystem II, when light excites a specific chlorophyll molecule known as P680 (Figure 2). The molecule absorbs the light energy and becomes excited, meaning it has more energy than usual. An electron “jumps” to a higher energy level in the molecule. Normally, the electron would immediately lose its additional energy and drop back down to its original position. However, it is transferred to an electron acceptor, Q, which sends the excited electron down a series of molecules known as an electron transport chain. As the electron is passed from one molecule to the next, a series of coupled redox reactions occurs. Some of the energy is eventually used to make ATP. The unexcited electron settles into a different chlorophyll molecule, known as P700, leaving behind a “hole” (electron deficiency) in P680.

Photosystem I begins when a different molecule in P700 also absorbs light energy and becomes excited. It too jumps to a higher energy level, where it is met by the electron acceptor which is iron-sulfur cluster and it further sends the excited electron down a ferredoxin chain, where coupled redox reactions occur. Meanwhile, back in photosystem II, the hole in P680 formed supplies the oxidizing energy used to split some molecules into their component hydrogen ions, electrons and oxygen molecules. The oxygen is

released into the air. One of its electrons is used to “plug” the hole in P680. The hydrogen atoms are concentrated into the lumen forming proton gradient and this gradient is useful in synthesis of ATP. The reaction is mediated by ATP synthase enzyme. The ATP (*adenosine triphosphate*) molecules and the NADPH (*reduced nicotine amide adenine dinucleotide*) molecules are used to produce glucose in another set of reactions that do not require direct sunlight i.e. the dark reaction.

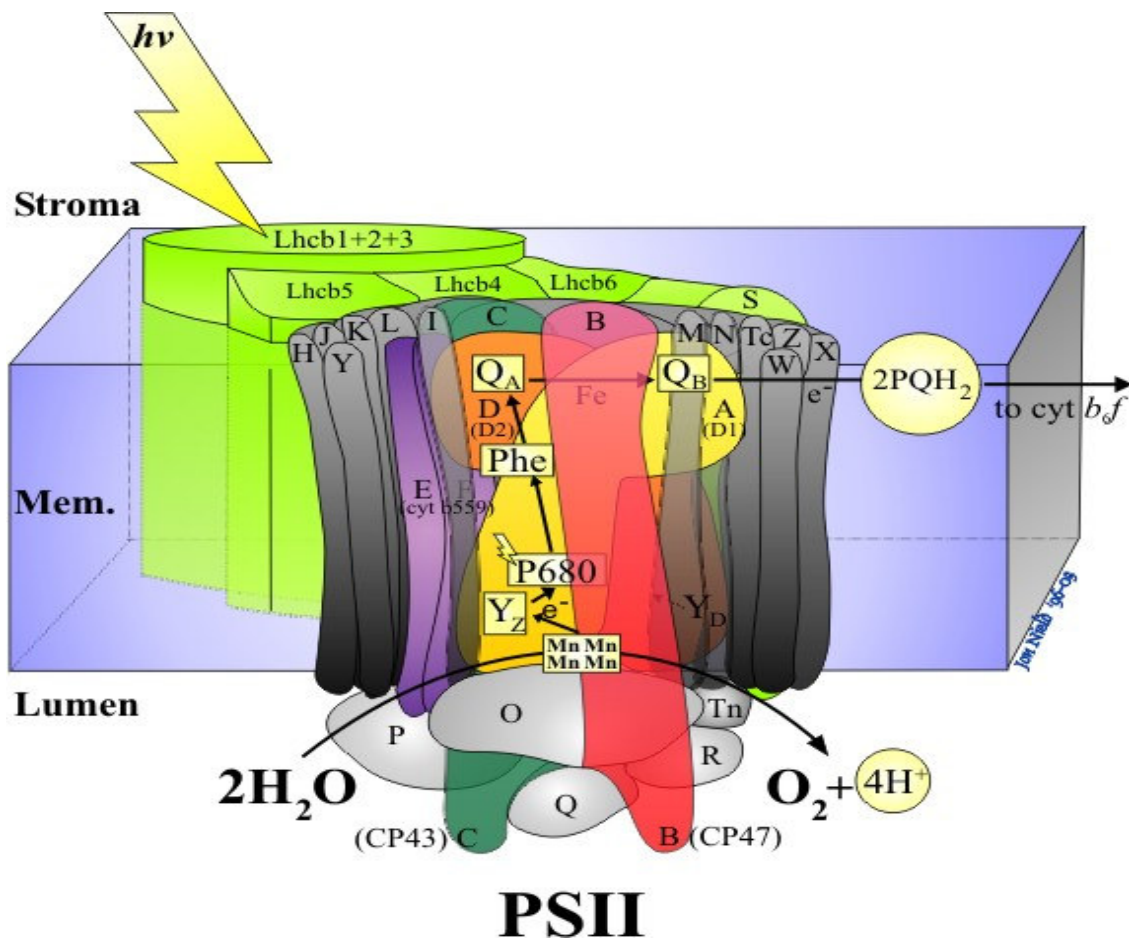


Figure 3. Reaction Center Proteins.

Adapted from <http://www.bio.ic.ac.uk/research/barber/psIIimages/PSII.html>

The protein complex that is involved in the electron transfer mechanism is called reaction center protein complex (Figure 3). The reaction center complexes are integral membrane pigment-proteins that span the membrane in vectorial fashion to carry out electron transfer reactions. The origin and extent of distribution of these proteins has been perplexing from a phylogenetic point of view, mostly because of extreme sequence divergence. Photosynthesis was carried out in non-oxygenic manner by primitive organisms; evolution has resulted in oxygenic mode of photosynthesis. However the reaction center proteins in all the organisms are somewhat similar and hence intriguing from a phylogenetic and evolutionary viewpoint. The bacterial reaction centers (non – oxygenic organisms other than *Cyanobacteria*) are similar either to photosystem I or II.

2. METHODS

2.1 Selection of Sequences

The apoproteins psaA (A1) and psaB (A2), of photosystem I, and the protein heterodimers psbA (D1) and psbD (D2), of photosystem II, were chosen from

- Cyanobacteria : (a)*Synechocystis sp. PCC 6803*,
(b)*Nostoc sp. PCC 7120*
(c)*Gloeobacter violaceus PCC 7421*
(d)*Prochlorococcus marinus str. MIT 9313*
- Rhodophyta : *Cyanidium caldarium*
- Chlorophyta : *Chlorella vulgaris*
- Streptophyta : (a)*Physcomitrella patens*
(b)*Arabidopsis thaliana*

The reaction center I protein homodimers were obtained from *Chlorobium tepidum TLS* (Chlorobi) and *Heliobacillus mobilis* (Firmicutes) while reaction center II proteins pufL and pufM were from

- Two green non-sulfur bacteria (Chloroflexi): *Chloroflexus aurantiacus*, *Roseiflexus castenholzii*
- Three α -proteobacteria: *Blastochloris viridis*, *Rhodobacter sphaeroides*, *Erythrobacter sp.*
- One β -proteobacteria: *Rubrivivax gelatinosus*
- One γ -proteobacteria: *Allochromatium vinosum*

Roseiflexus castenholzii has a single gene coding for pufL as well as pufM proteins. The total sequence is broken into these two proteins and individual sequence is used in this study. The C-terminal end of the psaB protein from *Gloeobacter violaceus* is cleaved (around 155 amino acids) [1]. This domain is associated with the outer membrane, not electron transfer mechanism and relates to the fact that the membrane organization in *Gloeobacter violaceus* is different from other organisms included in this study. Also, the first six N-terminal helices of psaA and psaB are cleaved off since they code for the antenna binding complex, not the electron transfer domain. The homodimeric proteins from *Chlorobium tepidum* and *Heliobacillus mobilis* have eleven helices, the last six similar to those in psaA or psaB. Only this portion is retained for analysis in this study. The investigation methodology is explained below.

2.2 Sequence alignment

In this study, multiple sequence comparison is the first step to recognize the sequence similarities. Subtle identities between sequences generally imply structural, functional, and evolutionary relationships among protein and DNA sequences. The protein sequences are much easier to align than the DNA sequences and hence are preferred for analyzing distantly related proteins. The sequences are obtained from protein database maintained at the National Center for Biotechnology Information (Bethesda, Maryland). The protein accession numbers are tabulated in Table 1. Pairwise alignments are computed using GONNET series matrix, to get pairwise scores. Further, for multiple alignments, proteins from same protein family are first aligned with each other. Then, protein families forming dimers in same photosystem, like

pufL-pufM, D1-D2, or A1-A2, are aligned forming protein family pairs. Alignment of these pairs from photosystem I and reaction center I was done next. This step is repeated for photosystem II and reaction center II. Finally, proteins from both systems were aligned to yield the overall similarity picture. Multiple alignments are done with Clustal W [2] using GONNET series matrix, as shown in the sequence alignment section.

	psbA (D1)	psbD (D2)	psaA(A1)	psaB(A2)	pufL	pufM	RC-I
Synechocystis sp. PCC 6803	CAA31899.1	PS0097	BAA17437.1	CAA41630.1	--	--	--
Nostoc sp. PCC 7120	BAB75441.1	BAB75989.1	BAB76853.1	BAB76854.1	--	--	--
Gloeobacter violaceus PCC 7421	NP_926090.1	BAC90264.1	BAC91379.1	BAC91380.1	--	--	--
Prochlorococcus marinus str. MIT 9313	CAE20594.1	CAE21354.1	CAE21945.1	CAE21944.1	--	--	--
Anabaena variabilis	F2A117	S42646	I39615	AAA18489.1	--	--	--
Arabidopsis thaliana	CAA56270.1	AAO13251.1	BAA84385.1	BAA84384.1	--	--	--
Chlorella vulgaris	BAA57842.1	BAA57876.1	BAA57926.1	BAA57928.1	--	--	--
Cyanidium caldarium	AAB82694.1	CAA44459.1	AAF12880.1	AAF12881.1	--	--	--
Physcomitrella patens	NP_904209.1	NP_904207.1	BAC05488.1	BAC05489.1	--	--	--
Chloroflexus aurantiacus	--	--	--	--	CAA33102.1	CAA30694.1	--
Roseiflexus castenholzii	--	--	--	--	BAC76414.1	BAC76414.1	--
Allochromatium vinosum	--	--	--	--	BAA32740.1	BAA32741.1	--
Rubrivivax gelatinosus	--	--	--	--	2123297A	BAA04101.1	--
Rhodobacter sphaeroides	--	--	--	--	CAA44999.1	WNRFFMS	--
Blastochloris viridis	--	--	--	--	CAA27550.1	CAA27551.1	--
Erythrobacter sp.	--	--	--	--	CAA40818.1	CAA40819.1	--
Chlorobium tepidum TLS	--	--	--	--	--	--	NP_662895.1
Heliobacillus mobilis	--	--	--	--	--	--	T31454

Table 1. Table showing accession numbers of photosystem-I and II and reaction center I and II proteins

2.3 Phylogenetic Tree building

From the results of sequence alignments, phylogenetic trees are built using MEGA2 (Molecular Evolutionary Genetics Analysis) software [3], using neighbor-joining algorithm. Pairwise deletion is used for gap handling. Bootstrapping (500 times) is used as test of inferred phylogeny. The tree generated for all sequences depicts the same grouping as seen by trees for individual subunits. It is clear that the D1 - D2 cluster has an apparent ancient gene duplication

and then a very long edge followed by an explosion at the end, while the L - M data have quite a different edge length, despite the same overall topology.

2.4 Structural alignment

Alignment of multiple protein structures is the next and vital step for studying the similarity and homology between proteins. Although amino acid sequences of many proteins are available in databases, comparatively fewer protein structures are known. The coordinates for the three dimensional structures of the photosystem proteins are retrieved from Protein Data Bank (PDB) [4]. The PDB entry file names, PDB IDs, the protein chains studied and their resolution are enumerated below:

- α -proteobacteria:

(a) *Rhodobacter sphaeroides* (1AIJ), L,M chains, 2.20 Å [5] (b) *Rhodospseudomonas viridis* (1DXR), L, M chains, 2.00 Å [6]

- γ -proteobacteria :

Thermochromatium tepidum (1EYS), L,M chains, 2.20 Å [7]

- Cyanobacteria :

(a) *Thermosynechococcus elongatus* (1S5L), D1,D2 chains, 3.50 Å [8]

(b) *Synechococcus elongatus* (1JB0), A1, A2 chains, 2.50 Å [9]

Multiple structural alignments, based on conventional methods like CE [10] or DALI [11], use ‘master-slave’ pairwise alignments. These multiple alignments are not done by all-to-all

comparison of proteins but by ‘pile-up’ of structural neighbors. Other methods, like HOMSTRAD [12] and CAMPASS [13] provide multiple alignments not for user-selected chains, but only for predefined protein families. Hence to overcome the above problems, the CE-MC server is used to generate the overlays. The alignments are generated using combinatorial extension (CE) algorithm and are iteratively optimized using Monte-Carlo (MC) simulations [14]. Structures from same protein family are first aligned with each other. Further, protein families forming dimers in same photosystem, like pufL-pufM, D1-D2, or A1-A2, are aligned forming protein family pairs. Alignment of these pairs from photosystem II and reaction center II is done next. Finally, proteins of both systems are aligned to yield the overall homology picture. The sequence alignments for these proteins are also generated as shown in sequence alignment section.

Next, the structural alignments of entire dimer proteins were generated using MultiProt [15]. Figure 4 shows two alignments – one without photosystem I proteins and one with these proteins.

2.5 Computation of RMS distances

After aligning the structures, root mean square distances (RMSD) between the α -carbon atoms of the backbone chains of all possible protein pairs are calculated (Table 2). The RMSDs gradually increase along the breadth of rows, depicting good conformity between related families L-M or D1-D2 and modest degree of similarity between sequences from distant families.

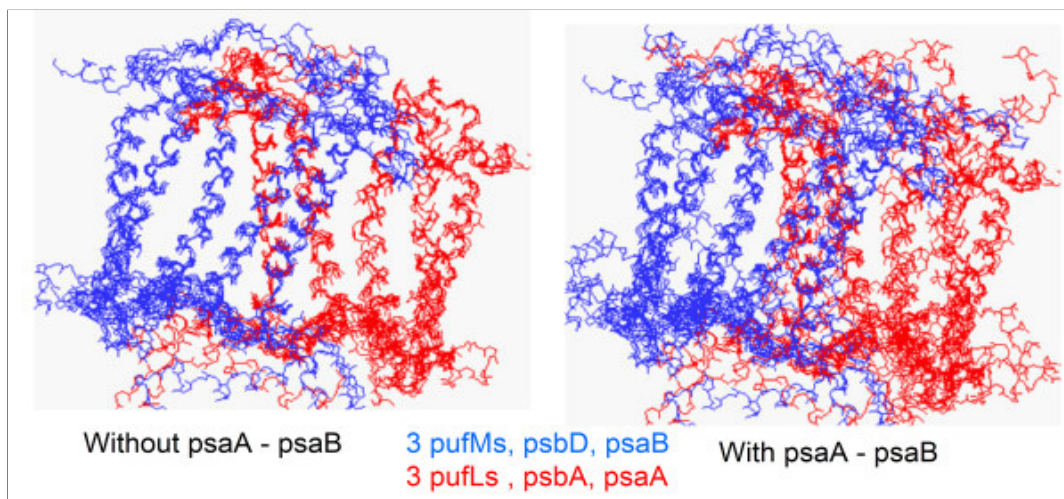


Figure 4. Structural alignments of entire protein dimers

	R. Sphaeroides L	R. Viridis L	T. Tepidum L	R. Sphaeroides M	R. Viridis M	T. Tepidum M	T. Elongatus D1	T. Elongatus D2	S. Elongatus A	S. Elongatus B
R. Sphaeroides L	--	0.74	0.65	1.8	1.8	1.52	2.59	3.46	3.86	3.55
R. Viridis L	0.5836	--	0.65	1.3	1.78	1.63	2.37	2.32	3.61	3.96
T. Tepidum L	0.6728	0.6654	--	1.32	1.79	1.31	2.42	2.4	3.97	3.89
R. Sphaeroides M	0.342	0.368	0.3502	--	1.24	1.2	2.26	2.19	3.9	3.44
R. Viridis M	0.3024	0.2702	0.2717	0.506	--	1.09	3.99	3.38	3.24	3.37
T. Tepidum M	0.3305	0.3348	0.318	0.6494	0.6151	--	3.61	3.42	4.11	3.28
T. Elongatus D1	0.2146	0.2017	0.195	0.2208	0.1872	0.2174	--	1.73	4	4.17
T. Elongatus D2	0.205	0.205	0.2024	0.2161	0.2083	0.217	0.3698	--	3.84	3.96
S. Elongatus A	0.0591	0.0699	0.0729	0.0677	0.0576	0.0838	0.0773	0.0674	--	1.71
S. Elongatus B	0.0806	0.086	0.0838	0.099	0.089	0.089	0.0674	0.1146	0.4242	--

Table 2. The table shows the root mean square distances (RMSD) between any pair of proteins in Angstrom units as upper triangular values. The lower triangular values are identities calculated from p-distance. The RMSDs are calculated between the alpha carbon atoms forming the protein backbone. The values show good conformity among proteins from related families like L-M or D1-D2 and gradually increase for distant families.

2.6 Hydropathy plots

Proteins consist of several amino acids held together by peptide linkages. These linkages bind the amino acids that can be hydrophobic or hydrophilic. The hydrophobicity of the amino acids determines where the amino acid will be located in the final structure of protein. Since the proteins studied here are all transmembrane proteins, it is anticipated that their intra-membrane α -helices show hydrophobicity. If all proteins are similar in structure, they are expected to show hydrophobicity at same domains in the aligned structures. The hydropathy plot, for the structural alignments generated with CE-MC, is calculated using Kyte-Doolittle algorithm [16] and it shows the expected trend. In these plots, the transmembrane regions are identified by peaks with hydropathy scores greater than 1.8. A window of size 19 is used in hydropathy algorithm. It is observed that there are five peaks above the threshold line of 1.8 (Figure 5). Comparison of these overshoots with structural overlays and sequence alignments reveal that these indeed correspond to the five aligned α -helices shown in Figure 5.

2.7 Identity and RMSD plots

From the CE-MC alignment results, variation of identity versus RMSD values was plotted as shown in (Figure 6). The plot was generated for all possible protein pairs. The identity values were calculated from p-distances evaluated using MEGA2 [3]. Another set of plots was generated showing the variation of amino acid statistics along the alignment length. Two examples are shown for alignments of *R. sphaeroides* pufM and pufL in (Figure 7) and *T. elongatus* D1 and D2 (Figure 8).

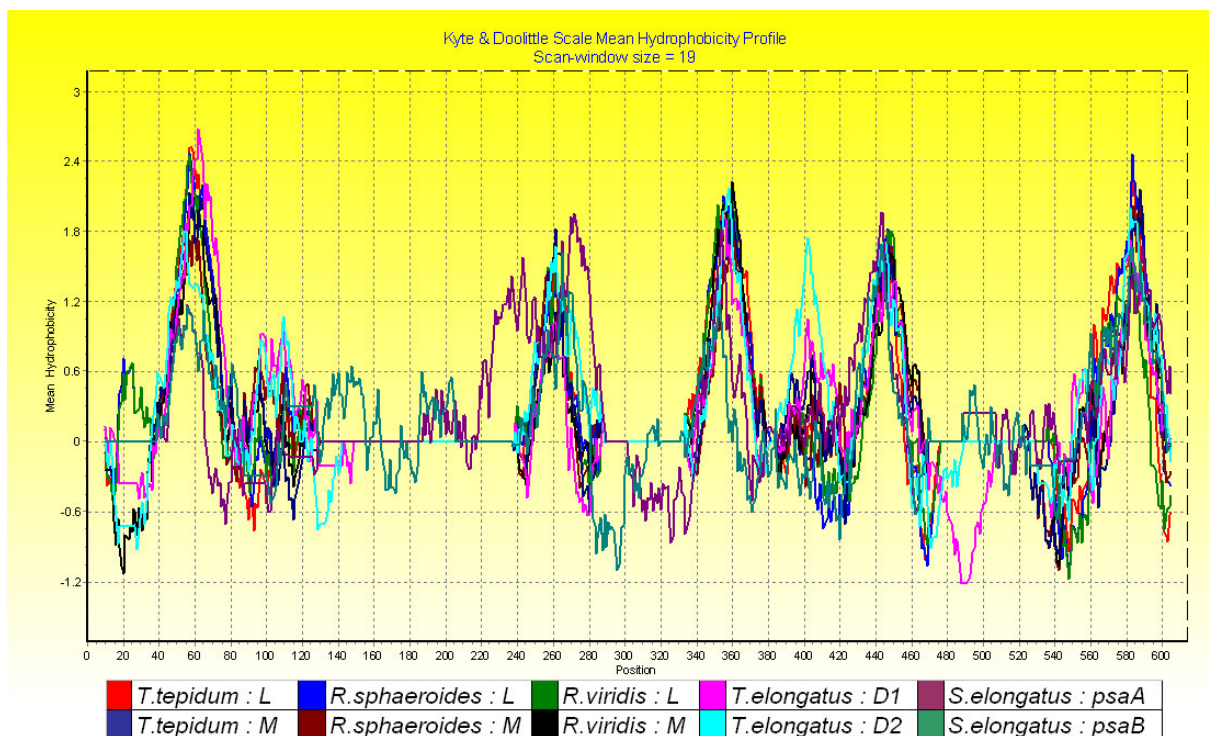


Figure 5. The figure shows Kyte and Doolittle's hydropathy plots. The overshoots above the threshold of 1.8 values are highly hydrophobic regions and correspond to the membrane spanning helices. These appear to be well conserved in all proteins.

These statistics were calculated using 3-D co-ordinates of alpha carbon atoms from CE-MC structural alignments and their sequences (fasta format) through Perl scripts and studied over a window of length 19. Over the entire length of alignment it was observed that as RMS between aligned alpha carbon atoms decreases, the identity increases. One may find the average identity from these plots lower than that portrayed by the plot of identity variation versus RMSD. This was because the latter one was calculated with pairwise deletion as well as deletion of spaces with gaps in one of the two aligned sequences. However, in the plots showing identity along the length, the calculations were done with pairwise deletion only. Positions with gaps in just

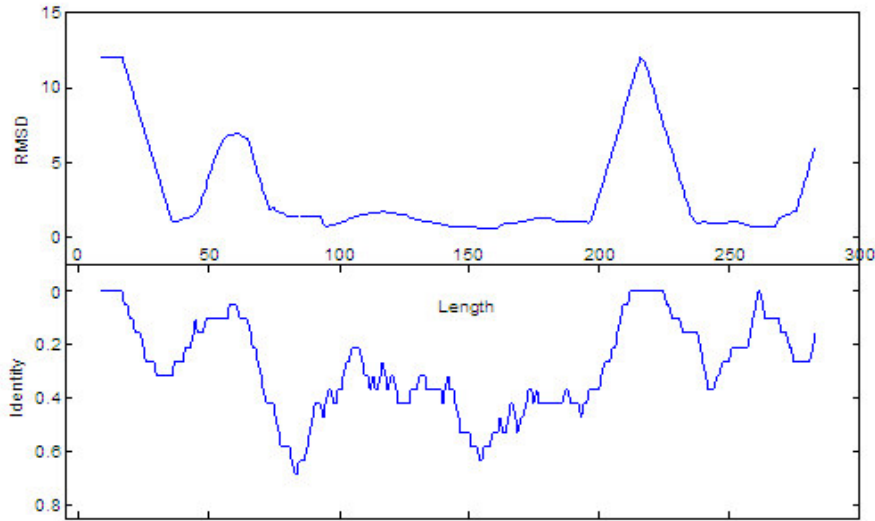


Figure 7. Variation of RMS and Identity between aligned carbon alpha atoms along the alignment length in *R. sphaeroides* pufL and pufM alignment

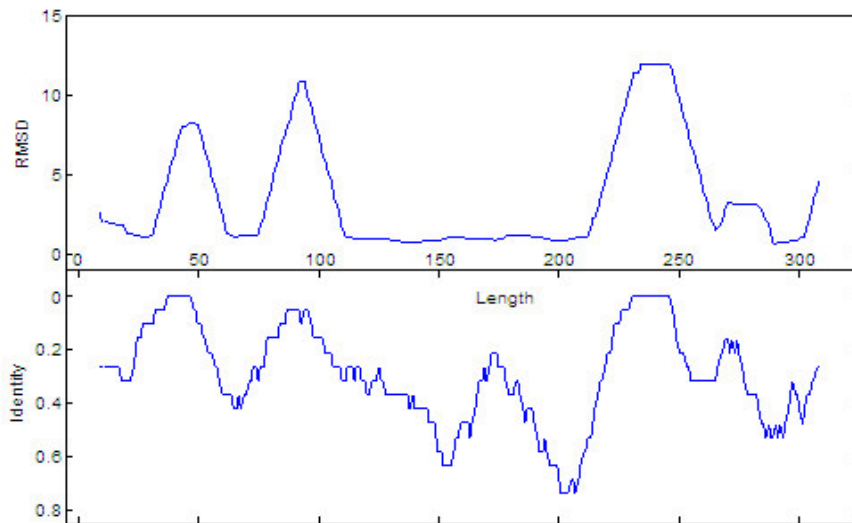


Figure 8. Variation of RMS and Identity between aligned carbon alpha atoms along the alignment length in *T. elongatus* psbA and psbD alignment

3. RESULTS

From pairwise alignments it was observed that D1-D2 proteins share around 15% identity with L-M proteins. Although this value is low, the conserved amino acids correspond largely to the binding sites of the photochemically active cofactors. The proteins *psaA* and *psaB* share very low identity with D1-D2 (6-7%) as well as with L-M (4-6%).

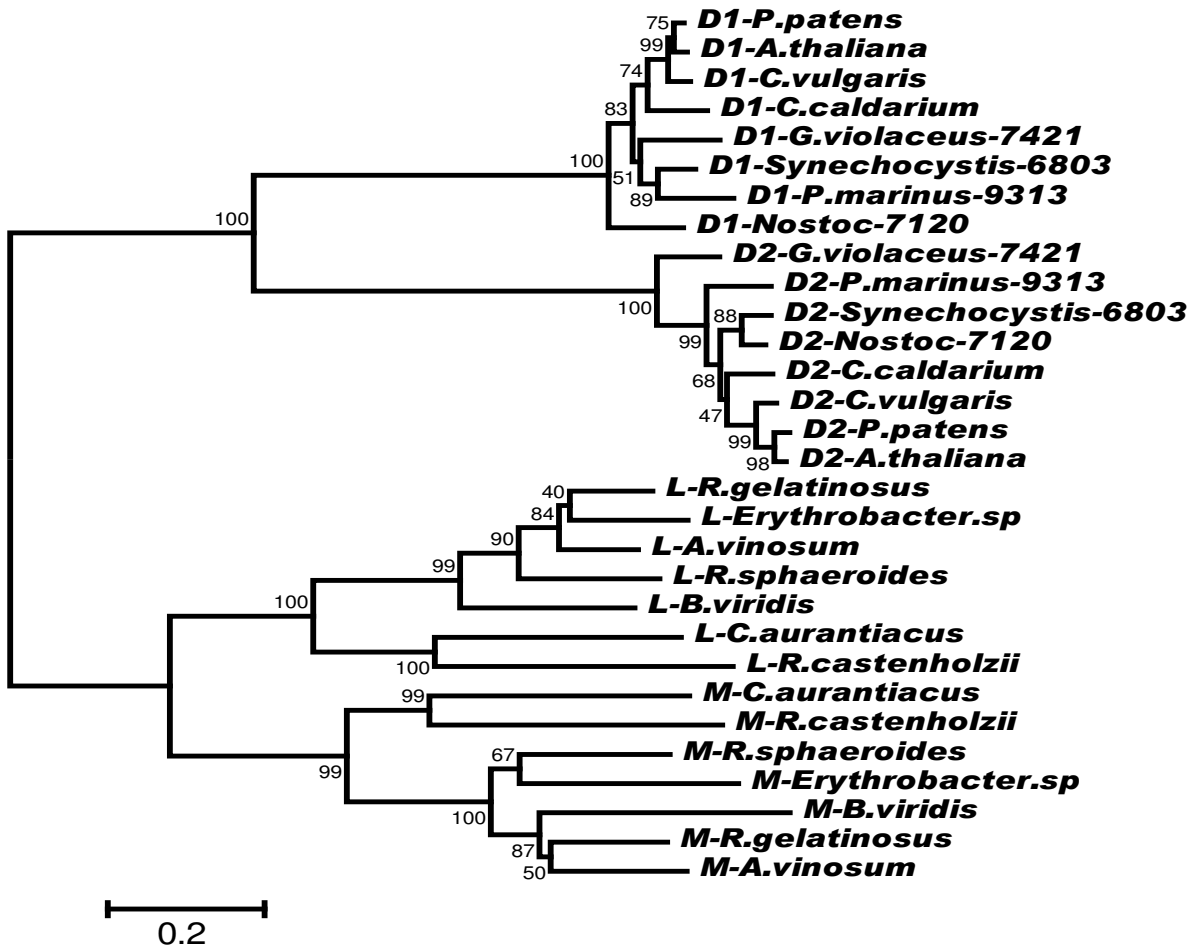


Figure 9. Neighbor-joining phylogenetic tree, with bootstrap values, depicting gene duplication between *pufL-pufM* and *psbA* (D1) - *psbD* (D2) proteins. The reaction center proteins L and M show gradual evolution as against the long branch of photosystem II exploding into D1-D2 at the end.

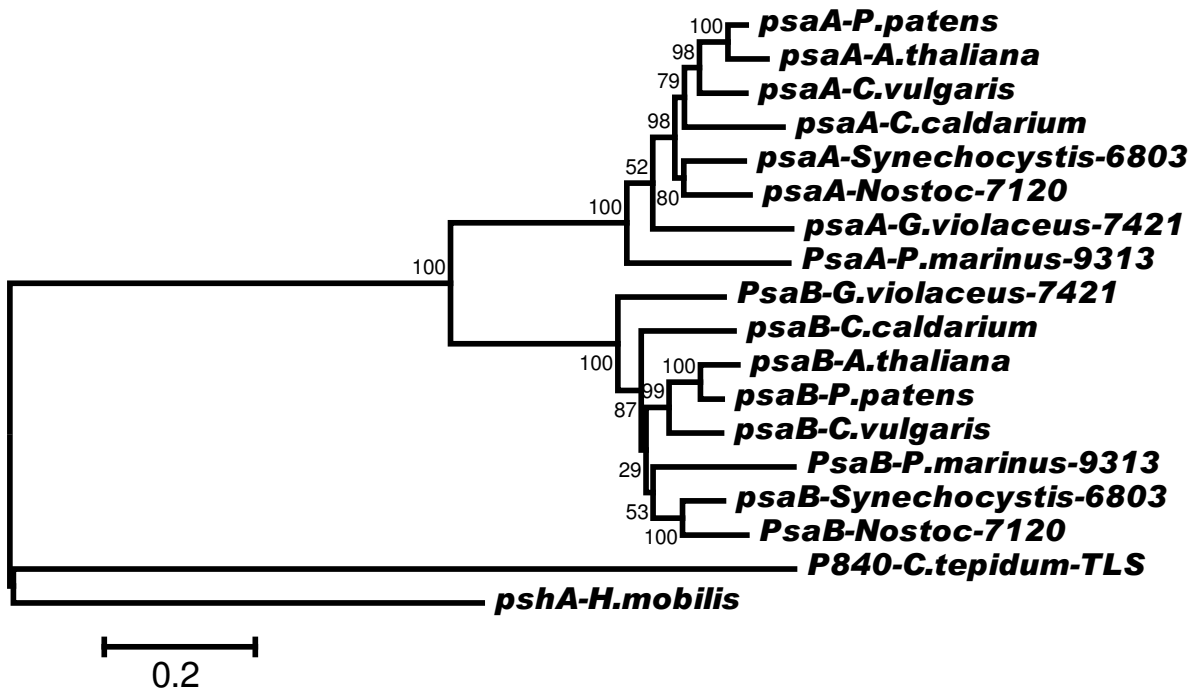


Figure 10. Phylogenetic tree, with bootstrap values, based on sequence alignments of photosystem-I and Reaction Center-I.

From the multiple sequence alignments, it was observed that the two molecules of the heterodimeric reaction center of photosystem I i.e. *psaA* and *psaB*, are more closely related to each other than either is to any other protein family. In parallel, the high level of identity in amino acids between L and M or between D1 and D2 suggests that these pairs are results of two separate gene duplications. The proteins L, M, D1, and D2 show high degree of conserved regions, implying that these proteins of photosystem II and reaction center II had a common ancestor (Figure 9). Also, the similarity between A1, A2 and reaction center I homodimer proteins, indicate them having a single precursor (Figure 10).

Figure 11. Overview of the unrooted neighbor-joining phylogenetic tree created with MEGA2 software. The tree comprises of reaction center proteins from photosystem-I, photosystem-II and the bacterial photosystem -I and photosystem-II like reaction centers. Corresponding accession numbers are given in the text

The structural overlays portray likewise evolutionary drift. The five α helices align beautifully depicting good structural alignments and thereby functional similarity (Figure 12). These helices comprise of around 25-30 amino acids, mostly isoleucine, leucine, valine and phenylalanine. On careful observations it was seen that the fourth alpha helices were not as well aligned as other helices. Only the fourth alpha helices from pufLs, pufMs, D1 and D2 proteins were very well aligned forming an almost straight helical structure while the fourth helices from A1 and A2 wound around this structure like creepers. This may be due to possible indels or pigment/cofactor arrangement in this region. Another reason can be protein-protein interactions constituting a selection pressure. This high similarity in the eight proteins from reaction centers and photosystem II was also evident in the dimer alignments (Figure 4). The alignments became less clear on addition of photosystem I proteins. Nonetheless, the overall protein structure fold was preserved over all ten dimer helices in all ten proteins. The sequence alignments of these proteins show high degree of similarity in the hydrophobic domains forming the intra-bilipid α -helices, which are the main functional regions. See sequence alignment section. The unaligned portions of certain proteins, corresponding to gaps in rest of the sequences, are large overhangs joining these helices. The trees generated from these alignments agree with the topology of those generated for the exhaustive sequence alignments

mentioned above (Figure Figure 13), showing early division into two branches, each diversifying into two offshoots of D1-D2 and L-M.

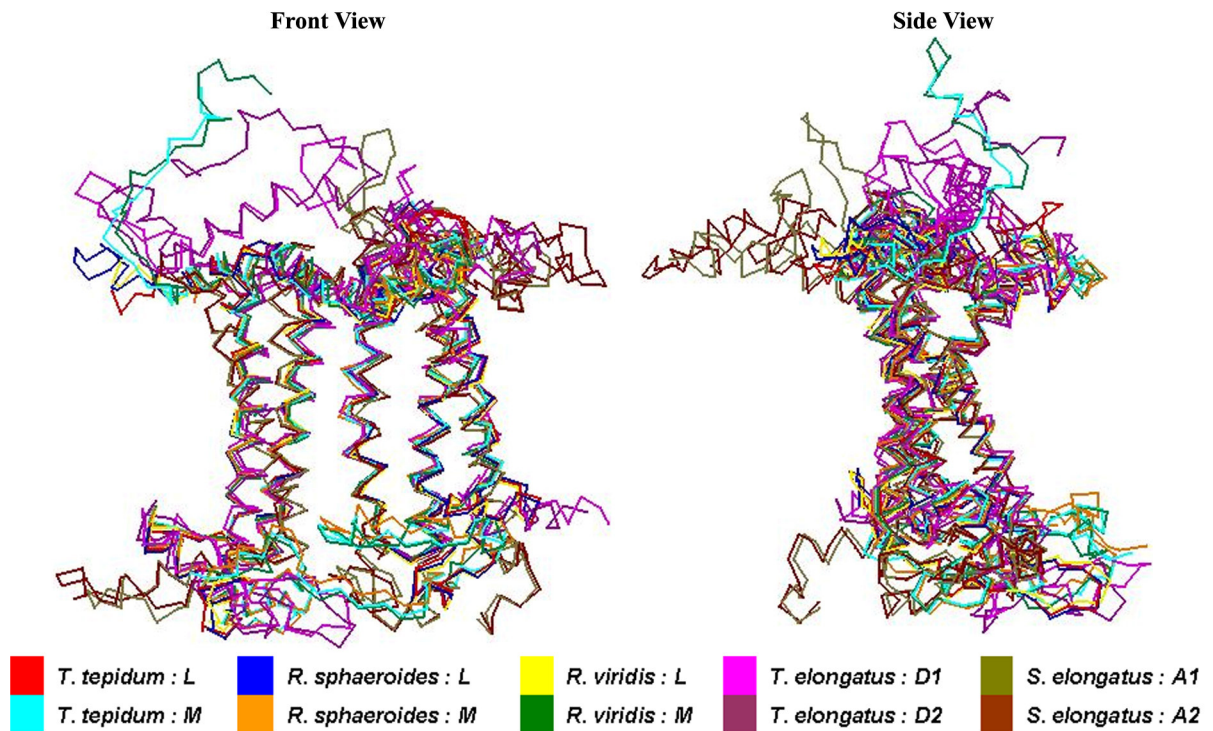


Figure 12. The figure shows structural alignments of all proteins: alpha-proteobacteria: (a)*Rhodobacter sphaeroides* (1AIJ), L,M chains, 2.20, (b)*Rhodospseudomonas viridis* (1DXR), L, M chains, 2.00 Å ; gamma-proteobacteria : *Thermochromatium tepidum* (1EYS), L,M chains, 2.20 Å ; Cyanobacteria : (a)*Thermosynechococcus elongatus* (1S5L), D1,D2 chains, 3.50 Å, (b) *Synechococcus elongatus* (1JB0), A1, A2 chains, 2.50 Å. The proteins show highly conserved five transmembrane alpha-helices. The unaligned thread-like portions on the top and the bottom are the loops outside the membranes, joining these helices. The left figure shows the front-view with five aligned alpha helices while the right figure shows the side-view.

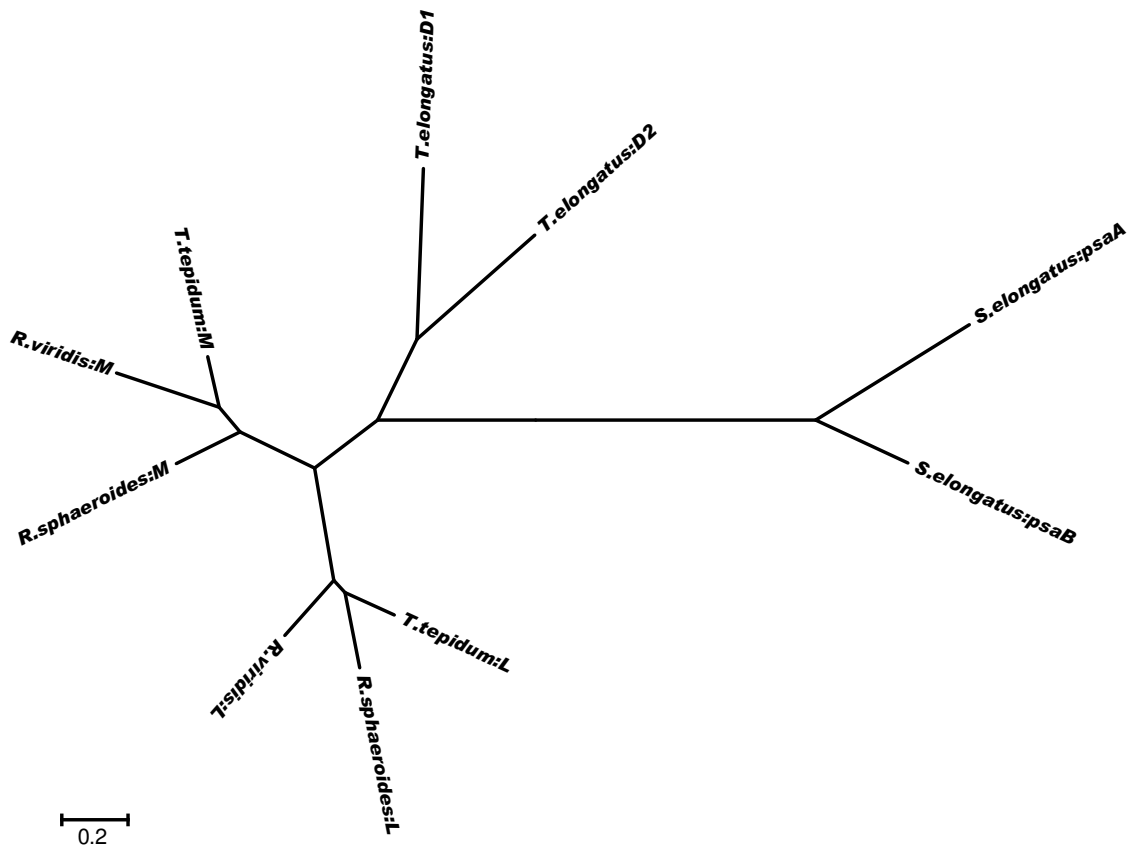


Figure 13. Phylogenetic tree based on the structural alignment results.

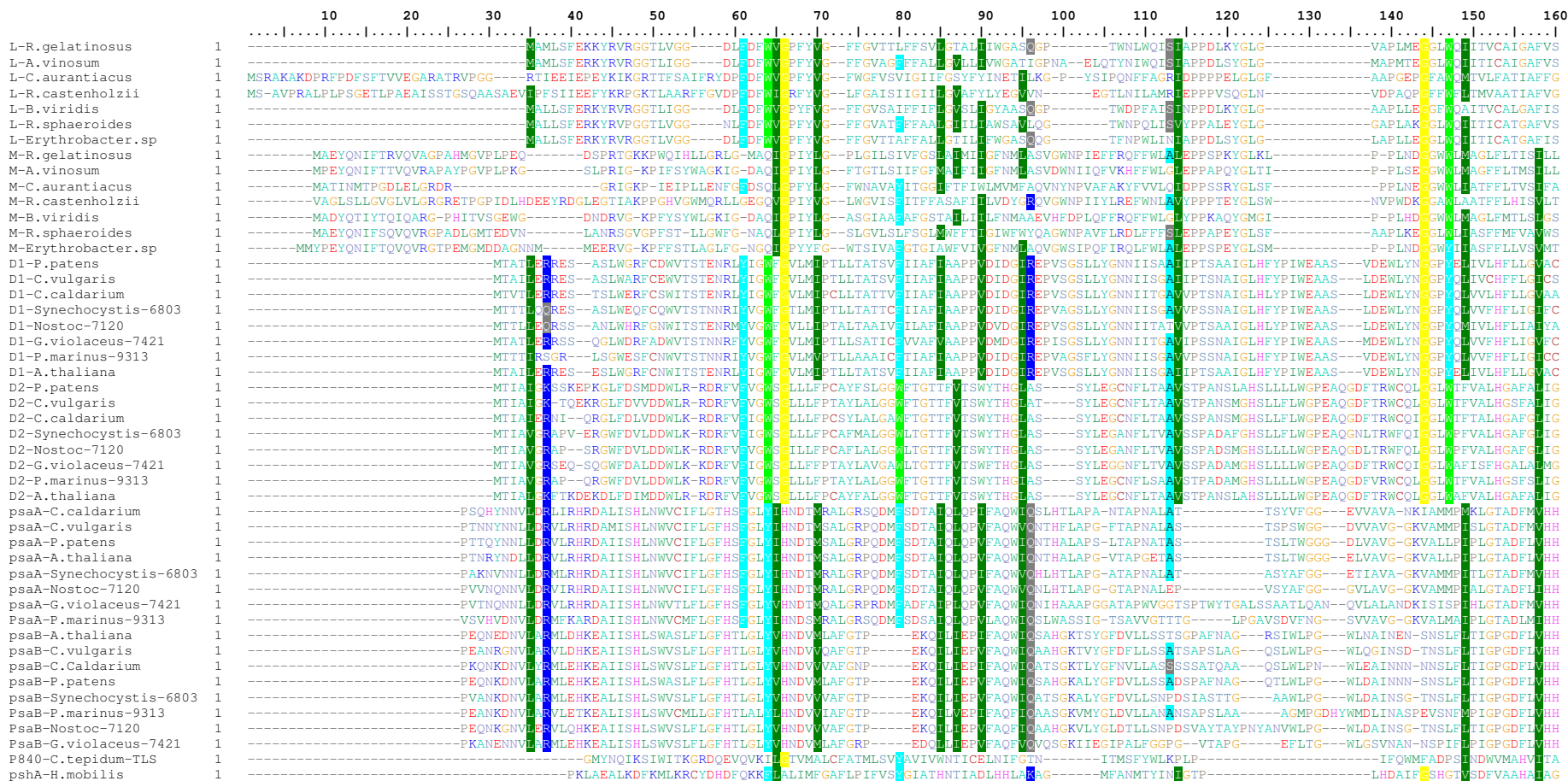
The plot of identity variation versus RMSD (Figure Figure 6) shows inverse relation between these attributes. As the RMS distance increases, the identity decreases, portraying increase in dissimilar amino acids, as expected. The points - red diamond and red triangle corresponding to RMS of $\sim 2.2 \text{ \AA}$ depict the values for R. sphaeroides pufM protein and T. elongatus D1 and D2 respectively. These stand apart from points - red diamonds and red triangles at RMS of $\sim 3.3\text{-}4 \text{ \AA}$, corresponding to values for R. viridis pufM and T. tepidum pufM proteins with respect to T. elongatus D1 and D2. A possible reason for this discrepancy can be presence of certain cofactors in R. viridis pufM and T. tepidum pufM proteins thereby increasing the RMS

values because of structure change. The identities for all three pufMs with respect to D1 and D2 are almost same i.e. ~20%. The other set of plots - Figure 7 and Figure 8 renders a similar drift in identity with respect to RMS along the length. However, on careful observation, it was seen that region 230-270 in Figure 7 and region 150-200 in Figure 8 do not agree with the expected trend. Although RMS is small in these regions, the identity is low. A possible explanation can be positive selection in these areas. Another reason can be presence of certain cofactors. The study of these issues is beyond the scope of this project.

Although the proteins belong to the twilight zone [17] of sequence alignments, other heuristic approaches yield quite good information about the evolution of reaction center proteins. There is no definitive structural data available for proteins of the reaction centers of two important classes of bacteria, the heliobacteria and the green sulfur bacteria. It would be interesting to incorporate this data, when it becomes available, to extend our study to these bacterial phyla. It is a common understanding that all photosynthetic reaction centers can be divided into two groups, one with pheophytin and quinone as electron acceptors and other with iron-sulfur center as electron acceptor. The above results provide support to the theory that a single primordial reaction center, probably a homodimer, produced these two separate systems through subsequent divergence and multiple gene duplications [18]. Further gene duplication and divergence resulted in heterodimeric complexes, thus developing the bacterial and plant-type reaction centers.

4. SEQUENCE ALIGNMENTS

Sequence Alignments 1



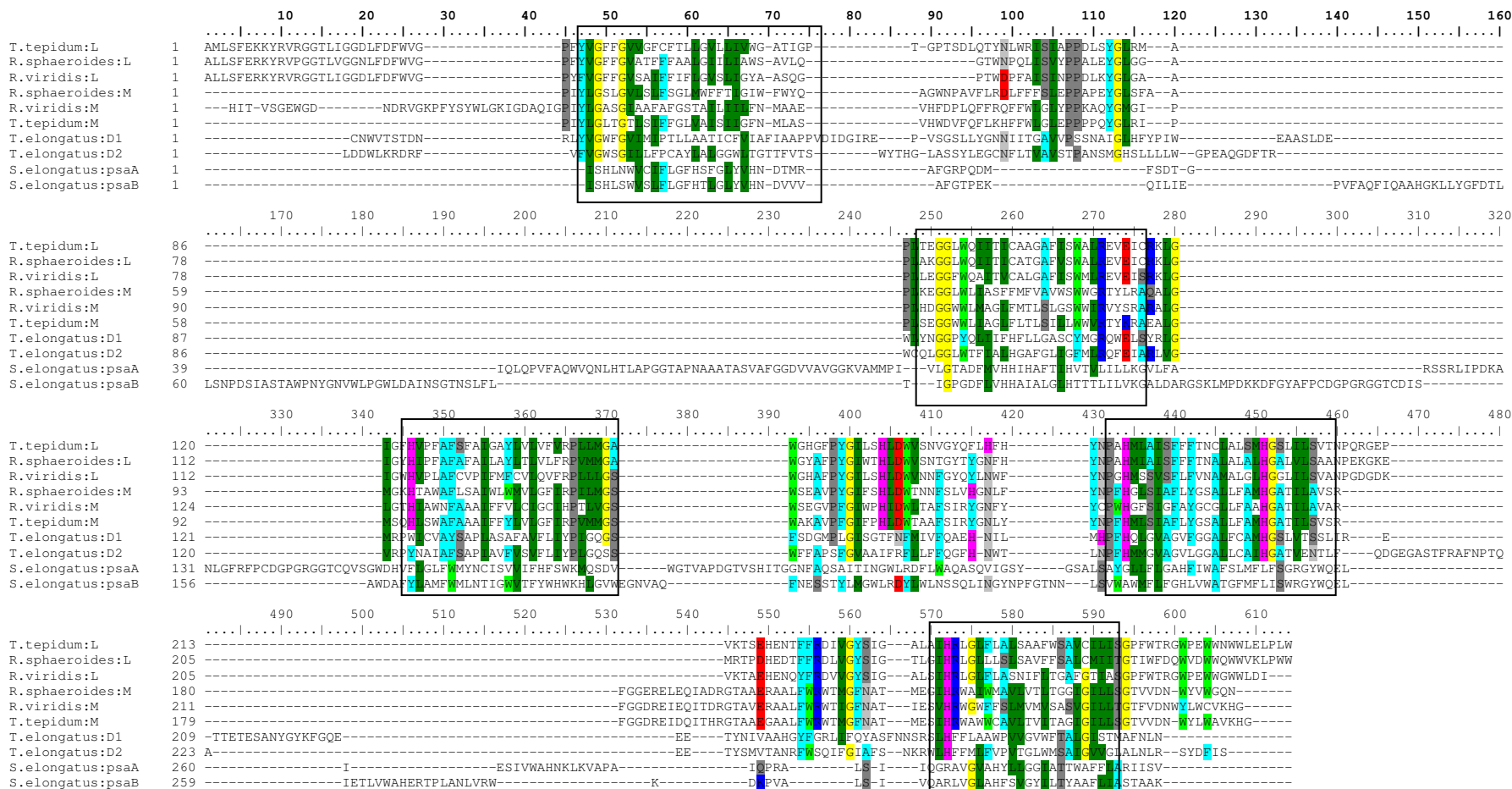
Multiple sequence alignment of pufL, pufM, psbA, psbD, psaA, psaB and RC-I protein. The alignments are generated through Clustal-W software using Gonnnet series matrix.

		170	180	190	200	210	220	230	240	250	260	270	280	290	300	310	320																																																					
L-R.gelatinosus	101	WALREVE	CRKLS	MOYHVP	IAES	FAFLAY	VLVVIRE	ILMGAM	HGFPY	IFSHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	AMSMH	GLLS	LSAAN	PKKG	----	EPMK	ITD	HED	ITFR	DAV	GS																																										
L-A.vinosum	106	WALREVE	CRKLS	EPH	PFAPA	FAL	DAYL	VVVR	ILMGAM	HGFPY	IFSHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	ALSMH	GLLS	LSVNP	PKKG	----	EBVK	SEH	ENT	FFRD	IVG	YS																																									
L-C.aurantiacus	140	WMMRQV	DSMK	DMGYH	VP	IAES	GVAFSA	NLV	QVIRE	IALGM	HGFPY	IFSHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LLACH	GLLS	LSAAQY	----	RGEG	GGD	IE	NVFF	RDV	QYS																																									
L-R.castenholzii	139	WLLRQID	ISLRK	DMMEV	PIAS	CAV	VSSN	ITL	QNLRE	IAMGAM	HGFPY	IFSHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	FLHMR	SAW	LSLSE	EAK	----	RNIS	QNI	H	VFN	RIL	GS																																								
L-B.viridis	101	WMLREVE	SRKLS	SHVHVP	LA	CVF	PMFCV	QVFR	LLGSM	HGFPY	IFSHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	ALGLR	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																								
L-R.sphaeroides	101	WALREVE	CRKLS	EPH	PFAPA	FAL	DAYL	VVVR	ILMGAM	HGFPY	IFSHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	ALGLR	GLLS	LSAAN	PKKG	----	KEMR	IPD	HED	ITFR	DAV	GS																																									
L-Erythrobacter.sp	101	WALREVE	CRKLS	MOYHVP	IAES	FAFLAY	VLVVIRE	ILMGAM	HGFPY	IFSHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	ALGLR	GLLS	LSAAN	PKKG	----	EBVK	SEH	ENT	FFRD	IVG	YS																																										
M-R.gelatinosus	130	WVWRYT	TRARAD	SMGTH	VAWAF	AAAT	WLV	LV	GFIRE	VLMGSM	SEAVP	YSG	IFPHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LFAMH	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																						
M-A.vinosum	129	WVWRYT	KRAEAL	MSQHL	SWAF	AAAT	WLV	LV	GFIRE	VLMGSM	SEAVP	YSG	IFPHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LFAMH	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																						
M-C.aurantiacus	119	WYMHY	YTRAKAL	SPY	LAY	CGT	GAT	ALY	LV	GFIRE	VLMGSM	SEAVP	YSG	IFPHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LFAMH	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																					
M-R.castenholzii	139	WVWRYT	TRAKAT	NYT	QLAWG	SA	SLY	V	LV	GFIRE	VLMGSM	SEAVP	YSG	IFPHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LFAMH	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																					
M-B.viridis	128	WVWRYT	SRARAL	EG	THI	AWN	AAAT	WLV	LV	GFIRE	VLMGSM	SEAVP	YSG	IFPHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LFAMH	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																					
M-R.sphaeroides	130	WVWRYT	LRQAAL	SMG	KHTAWA	LSA	WLV	LV	GFIRE	VLMGSM	SEAVP	YSG	IFPHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LFAMH	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																						
M-Erythrobacter.sp	133	WLLRAYL	IAEQH	KMKH	IFWGA	AAA	WLV	LV	GFIRE	VLMGSM	SEAVP	YSG	IFPHLD	WNSN	VGQYL	HFHYNP	PA	MLA	ITFFFT	TTIT	LFAMH	GLLS	LSVAN	PGDG	----	DKVK	TA	EH	ENQY	FRD	VVG	GS																																						
D1-P.patens	126	WMCREWEL	SYRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D1-C.vulgaris	126	WMCREWEL	SFRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D1-C.caldarium	126	WMCREWEL	SYRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D1-Synechocystis-6803	126	WMCREWEL	SYRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D1-Nostoc-7120	126	WMCREWEL	SYRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D1-G.violaceus-7421	126	WMCREWEL	SYRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D1-P.marinus-9313	126	WMCREWEL	SYRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D1-A.thaliana	126	WMCREWEL	SFRDL	MRPW	I	AVAS	SAP	VAAAT	AV	FLI	NP	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																																		
D2-P.patens	126	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
D2-C.vulgaris	125	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
D2-C.caldarium	124	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
D2-Synechocystis-6803	125	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
D2-Nostoc-7120	124	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
D2-G.violaceus-7421	125	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
D2-P.marinus-9313	124	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
D2-A.thaliana	126	WMLRQPE	LARS	V	LR	YPNA	IA	ES	GP	IAV	F	V	S	FLI	N	P	I	QGG	S	SDG	MPL	LS	SGT	FN	M	V	QAE	EHN	----	ILMH	PF	M	L	G	V	A	G	V	F	G	S																													
psaA-C.caldarium	120	IHAFTI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
psaA-C.vulgaris	120	IHAFTI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
psaA-P.patens	121	IHAFTI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
psaA-A.thaliana	121	IHAFTI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
psaA-Synechocystis-6803	120	IHAFTI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
psaA-Nostoc-7120	120	IHAFTI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
psaA-G.violaceus-7421	133	IFALCI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
PsaA-P.marinus-9313	125	IHAFTI	HV	V	TL	ILK	GV	L	F	AR	S	R	L	FD	K	AN	L	FR	FP	CD	GP	GR	GGT	CQ	V	S	W	D	H	V	F	L	G	F	R	M	Y	N	S	I	S	V	I	F	H	S	W	K	M	Q	S	D	V	W	M	C	I	V	S	----	Q	N	S	L	V	S	H	V	V	G
psaB-A.thaliana	123	AIALGL																																																																				

		330	340	350	360	370	380	390	400	410	420	430	
L-R.gelatinosus	224	IGSLG	IRLGLFLALSAA	WNSAV	CVIVISGP			FWTRGWP	EWGWWL	EP	NSQWPLN		
L-A.vinosum	229	IGALAT	IRLGLFLALSAA	WNSAV	CVIVISGP			FWTRGWP	EWNNWLE	EP	W		
L-C.aurantiacus	260	VGESGV	IRLGYIFAIG	ILSAD	CIILSG			WPVQD	WVSW	WNEWNN	EP	WNSGV	
L-R.castenholzii	257	IGEIC	IRVAFWTC	GAASV	LSNICH	ELSG		TFVKD	WNA	WGFWD	KKPF	WNGVQ	GAL
L-B.viridis	224	IGALG	IRLGLFLAS	NIFL	TAFG	IASGP		FWTRGWP	EWGWWL	EP	WFS		
L-R.sphaeroides	224	IGTLG	IRLGLLLLS	LSAV	FSAP	CMITGT		IWFQ	WVD	WQWV	WV	W	W
L-Erythrobacter.sp	224	VGTLG	IRLGYLLAINA	GLWSA	CIIVISGP			VWTAG	WPE	EWNNWLD	EP	W	W
M-R.gelatinosus	260	ATTES	IRMAWVAVL	CLP	CGG	ETILSG		TVVD	NWYL	RAV	KG	V	W
M-A.vinosum	259	ASME	IRMAWVAVL	TVITAG	ETILSG			TVVEN	WYL	RAIK	H	G	V
M-C.aurantiacus	249	ANAY	IRMAWVAVL	CGITG	AVFFSMP			DFV	NWFO	Q	IE	A	G
M-R.castenholzii	269	ANSY	IRMAWVAVL	TAITG	AVFFSMP			TLVP	D	WYA	GE	T	A
M-B.viridis	258	ATLES	IRMAWVAVL	TVITG	ETILSG			TFVD	NWYL	RAV	KG	V	W
M-R.sphaeroides	260	ATME	IRMAWVAVL	TVITG	ETILSG			TVVD	NWYL	RAV	KG	V	W
M-Erythrobacter.sp	263	ATME	IRMAWVAVL	TVITG	ETILSG			TVVD	NWFL	RAQ	E	H	F
D1-P.patens	260	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D1-C.vulgaris	260	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D1-C.caldarium	260	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D1-Synechocystis-6803	260	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D1-Nostoc-7120	260	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D1-G.violaceus-7421	260	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D1-P.marinus-9313	259	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D1-A.thaliana	260	FQYAS	FN	SR	SL	DF	LA	AW	PV	IC	W	PT	A
D2-P.patens	259	G--VAF	S	K	R	W	D	E	F	M	L	F	V
D2-C.vulgaris	258	G--VAF	S	K	R	W	D	E	F	M	L	F	V
D2-C.caldarium	257	G--VAF	S	K	R	W	D	E	F	M	L	F	V
D2-Synechocystis-6803	258	G--IAP	S	K	R	W	D	E	F	M	L	F	V
D2-Nostoc-7120	257	G--IAP	S	K	R	W	D	E	F	M	L	F	V
D2-G.violaceus-7421	258	G--IAP	S	K	R	W	D	E	F	M	L	F	V
D2-P.marinus-9313	257	G--IAP	S	K	R	W	D	E	F	M	L	F	V
D2-A.thaliana	259	G--VAF	S	K	R	W	D	E	F	M	L	F	V
psaA-C.caldarium	257	ABF	V	W	A	F	S	L	M	L	F	S	G
psaA-C.vulgaris	256	ABF	V	W	A	F	S	L	M	L	F	S	G
psaA-P.patens	258	ABF	V	W	A	F	S	L	M	L	F	S	G
psaA-A.thaliana	258	ABF	V	W	A	F	S	L	M	L	F	S	G
psaA-Synechocystis-6803	257	ABF	V	W	A	F	S	L	M	L	F	S	G
psaA-Nostoc-7120	257	ABF	V	W	A	F	S	L	M	L	F	S	G
psaA-G.violaceus-7421	284	ABF	V	W	A	F	S	L	M	L	F	S	G
PsaA-P.marinus-9313	262	GBF	V	W	A	F	S	L	M	L	F	S	G
psaB-A.thaliana	253	GHL	V	W	A	T	G	M	F	L	I	S	W
psaB-C.vulgaris	253	GHL	V	W	A	T	G	M	F	L	I	S	W
psaB-C.Caldarium	253	GHL	V	W	A	T	G	M	F	L	I	S	W
psaB-P.patens	253	GHL	V	W	A	T	G	M	F	L	I	S	W
psaB-Synechocystis-6803	252	GHL	V	W	A	T	G	M	F	L	I	S	W
PsaB-P.marinus-9313	255	GHL	V	W	A	T	G	M	F	L	I	S	W
PsaB-Nostoc-7120	256	GHL	V	W	A	T	G	M	F	L	I	S	W
PsaB-G.violaceus-7421	250	GHL	V	W	A	T	G	M	F	L	I	S	W
P840-C.tepidum-TLS	219	HLV	W	F	I	S	AV	W	E	D	R	C	S
psaH-H.mobilis	230	VQ	A	L	L	G	A	F	I	W	A	F	T

Continued multiple sequence alignment of pufL, pufM, psbA, psbD, psaA, psaB and RC-I protein. The alignments are generated through Clustal-W software using Gonnet series matrix.

Sequence Alignments 2



Sequence alignments of proteins that are structurally aligned. The large gaps correspond to the overhangs and β -sheets of the transmembrane photosynthetic proteins, connecting the intra bilipid layer α -helices. The boxes enclose the structurally aligned helices.

5. REFERENCES

- [1] Inoue, H., et al., Unique constitution of photosystem I with a novel subunit in the cyanobacterium *Gloeobacter violaceus* PCC 7421. *FEBS Letters*, 2004. 578(3): p. 275-279.
- [2] Thompson, J.D., D.G. Higgins, and T.J. Gibson, Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research*, 1994. 22(22): p. 4673-4680.
- [3] Nei, M. and S. Kumar, *Molecular evolution and phylogenetics*. 2000, Oxford; New York: Oxford University Press. xiv, 333.
- [4] Berman, H.M., et al., The Protein Data Bank. *Nucleic Acids Research*, 2000. 28(1): p. 235-242.
- [5] Stowell, M.H.B., et al., Light-induced structural changes in photosynthetic reaction center: Implications for mechanism of electron-proton transfer. *Science*, 1997. 276(5313): p. 812-816.
- [6] Lancaster, C.R.D., et al., Structural Basis of the Drastically Increased Initial Electron Transfer Rate in the Reaction Center from a *Rhodospseudomonas viridis* Mutant Described at 2.00-Å Resolution. *Journal of Biological Chemistry*, 2000. 275(50): p. 39364-39368.
- [7] Nogi, T., et al., Crystal structures of photosynthetic reaction center and high-potential iron-sulfur protein from *Thermochromatium tepidum*: Thermostability and electron transfer.

Proceedings of the National Academy of Sciences of the United States of America, 2000. 97(25): p. 13561-13566.

[8] Ferreira, K.N., et al., Architecture of the photosynthetic oxygen-evolving center. *Science*, 2004. 303(5665): p. 1831-1838.

[9] Jordan, P., et al., Three-dimensional structure of cyanobacterial photosystem I at 2.5 angstrom resolution. *Nature*, 2001. 411(6840): p. 909-917.

[10] Shindyalov, I.N. and P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 1998. 11(9): p. 739-747.

[11] Holm, L. and C. Sander, Protein-Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology*, 1993. 233(1): p. 123-138.

[12] Mizuguchi, K., et al., HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science*, 1998. 7(11): p. 2469-2471.

[13] Sowdhamini, R., et al., CAMPASS: a database of structurally aligned protein superfamilies. *Structure*, 1998. 6(9): p. 1087-1094.

[14] Guda, C., et al., CE-MC: a multiple protein structure alignment server. *Nucleic Acids Research*, 2004. 32: p. W100-W103.

[15] Shatsky, M., R. Nussinov, and H.J. Wolfson, MultiProt - a Multiple Protein Structural Alignment Algorithm. *Lecture Notes in Computer Science*, 2002. 2452: p. 235--250.

[16] Kyte, J. and R.F. Doolittle, A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 1982. 157(1): p. 105-132.

[17] Rost, B., Twilight zone of protein sequence alignments. *Protein Engineering*, 1999. 12(2): p. 85-94.

[18] Blankenship, R.E., Origin and Early Evolution of Photosynthesis. *Photosynthesis Research*, 1992. 33(2): p. 91-111.