Rubisco sequence-structure-function: coevolution and codon bias of *rbcL* gene

Animesh K. Agrawal

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University

July, 2012

Statement of originality

Except indicated below, the contents of this thesis are the result of my original research, which has been conducted under the principal supervision of Dr. Jill Gready (John Curtin School of Medical research, ANU).

The software described in Chapter 2 uses modules from the Biopython and BioSQL projects, and some of the design ideas behind it came from these projects.

In Chapters 3 and 4, I have used the results of an analysis of conserved and variable residues in plant Rubisco large subunit protein sequences compiled by Dr. Babu Kannappan in our group. Moreover, the structural analysis of residues identified in the coevolution analysis (section 3.3.1 and 3.3.3.1) in Chapter 3 has been done in collaboration with Dr. Kannappan.

I declare that the work presented in this thesis is to my belief original, except as acknowledged in the text and above, and that the material has not been submitted, either in whole or in part, for a degree at this or any other university.

A Agrawal

Animesh K. Agrawal

July, 2012

Acknowledgements

I would like to express my sincere gratitude to my research supervisor Dr. Jill Gready, for all the guidance, support and encouragement which she has given me during my PhD tenure. She gave me a free rein in planning and executing all my studies and was extremely patient and generous with her time, comments, suggestions, and corrections on every draft of my thesis.

I would like to thank Dr. Spencer Whitney and Dr. Peter Christen, advisors on my supervisory panel for their encouragement, advice and help. I would also like to thank Dr. Margaret Kahn for assisting me with installing programs on super computer facility at ANU. I would also like to thank Dr. J. Gregory Caporaso for advice on using coevolution module from Pycogent. My warm thanks to Dr. Lars Jermiin for discussions on phylogenetic analysis. I am grateful to Dr. Anna Cowan for her wise advice and support that helped me in completing this thesis.

I would like to especially thank Dr. Babu Kannappan in our group who gave me tremendous inputs during the entire course of my PhD studies. I also want to thank my wonderful colleagues from the Computational and Conceptual biology group: Tatiana Vassilieva, Peter Cummins, Peter Mathews, Mark Wallace and Hugo. My thanks also go to the friends I made here: Anjum, Abhishek, Anuj, Ajay, Debjani, Lokesh, Pooja, Kartik, Mishka and Julia for helping me feel at home in Australia, especially to Anuj for helping me out with python scripting in the course of my PhD studies.

My special gratitude is due to my second set of parents Ma and Baapi for their love and support. Thanks to my brothers Amit and Ankit for being there for me whenever I needed them.

My loving thanks to my wife Poulomi, without her unconditional love and support, I would not have been able to finish this work. I would also like to thank her for all her help in proof-reading my thesis.

This thesis is dedicated to my parents, to my father who has been the best teacher for me and to my mother for devoting herself to us.

Abstract

Ribulose-1, 5-bisphosphate carboxylase/oxygenase (Rubisco, EC 4.1.1.39) is the primary photosynthetic enzyme responsible for CO_2 fixation. Though Rubisco is critical to photosynthesis, it is a very inefficient enzyme with a low catalytic rate and competing oxygenase activity that initiates a wasteful photorespiratory pathway. This paradoxical relationship between the functional significance of Rubisco and its apparent inefficiency is puzzling and raises questions regarding the roles of evolution versus functional constraints in shaping Rubisco. This thesis examines the role of coevolution and codon-usage bias of the *rbcL* gene in Rubisco's fine-tuning at the molecular level.

The extent of information available on Rubisco is substantial. A local database was developed to archive Rubisco protein and nucleotide sequences in public databases, as well as to integrate the structural, kinetic and taxonomic data available on Rubisco. This database is based on BioSQL schema, employs MySQL as the relational database backend and uses the Biopython application programming interface (API). This local repository contains more than 11,000 unique Rubisco large-subunit (LSU) protein/*rbcL* nucleotide sequence entries from Angiosperms; kinetic data information from 40 species, including 11 species from flowering plants; and structural information from 50 PDB structures, including spinach, tobacco and rice from flowering plants.

Coevolution of Rubisco has been investigated in the intra-protein context using a large sequence dataset of Rubisco-LSU sequences, as well as in the inter-protein context through its interactions with Rubisco small subunit, chaperonins RbcX and Rubisco activase (RA). The intra-protein studies identified a novel cluster of coevolving sites spatially proximal to loop 6 and in the C- terminal tail, known regions of functional and structural importance in the Rubisco-LSU. The inter-protein analyses of the Rubisco-LSU and RA detected several new coevolving sites both in the Rubisco-LSU and in RA, in addition to predicting sites already identified by mutagenesis studies to be involved in Rubisco-LSU-RA interaction. In the Rubisco-LSU, these sites are located in the β-C-β-D loop which is known to be interacting with RA, along with a network of polar/charged residues in the C-terminal domain of the Rubisco-LSU.

The codon-usage bias of *rbcL* was analyzed using a large set of *rbcL* sequences and all available Angiosperm chloroplast nucleotide sequence data. Consistent with previous reports, studies on Angiopserm chloroplast genomes and their corresponding *rbcL* genes showed that both the *rbcL* genes and chloroplast genomes have obvious A+T bias. Based on evidence found in this study, a role for codon adaptation in *rbcL* is proposed, although it is limited to several two-fold and a three-fold codon-degenerate amino acids. Significantly, this study show that Rubisco's catalytic residues favor preferred codons (codons used more frequently in *rbcL* than other synonymous codons). Another important finding suggests translational accuracy selection in *rbcL*, based on statistically significant associations of preferred codons in *rbcL* with conserved and buried sites in the Rubisco-LSU.

Overall in this thesis, information available on Rubisco has been structured and integrated to develop a high-quality dataset for systematic studies, Rubisco's coevolution has been studied in both intra- and inter-protein contexts to identify coevolutionary constraints in its evolution, and codon preferences of *rbcL* has been investigated in order to understand the role of synonymous codons within the context of Rubisco's structure/function.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Abbreviations

Rubisco	Ribulose-1, 5-bisphosphate carboxylase/oxygenase
Rubisco-LSU	Rubisco large subunit
Rubisco-SSU	Rubisco small subunit
RA	Rubisco activase
L	Rubisco large subunit
S	Rubisco small subunit
rbcL	Rubisco large subunit gene
rbcS	Rubisco small subunit gene
NMI	Normalized mutual information
CO_2	Carbon dioxide
O_2	Oxygen
$S_{C/O}$	Specificity factor of Rubisco
tRNA	Transfer RNA
CABP	2-carboxyarabinitol-1,5-diphosphate

1 Introduction

1.1 Background

Rubisco is found in most autotrophic organisms, ranging from diverse prokaryotes, including photosynthetic and chemo-litho-autotrophic bacteria, cyanobacteria, and archaea, to eukaryotic algae and higher plants. In this chapter, I summarize the literature on the Rubisco superfamily relevant to this thesis. The amount of information available on Rubisco is enormous because of its biological importance at the core of photosynthesis. Extensive efforts since the 1990s have focused on engineering a Rubisco enzyme with an improved efficiency and/or specificity for CO_2 have also contributed to information available on Rubisco. However, despite detailed knowledge of Rubisco's molecular structure and catalytic mechanism now and numerous attempts to engineer a better Rubisco, there is no reported case in the literature of success. As the general goal of my thesis was to systematically analyze the variation in Rubisco sequences with *"in-silico"* techniques, I will focus in the chapter mainly on structural and functional aspects directly related to the computational studies conducted by me.

1.2 Perspectives and overview

Sustainability of life on earth is effectively reliant on transformation of solar energy to chemical energy, in the form of sugar molecules, by the process of photosynthesis. The initial step in this energy transformation reaction, i.e. photosynthetic fixation of CO_2, is catalysed by the enzyme Ribulose-1, 5-bisphosphate carboxylase/oxygenase as shown in Figure 1.1 (EC 4.1.1.39, Rubisco).

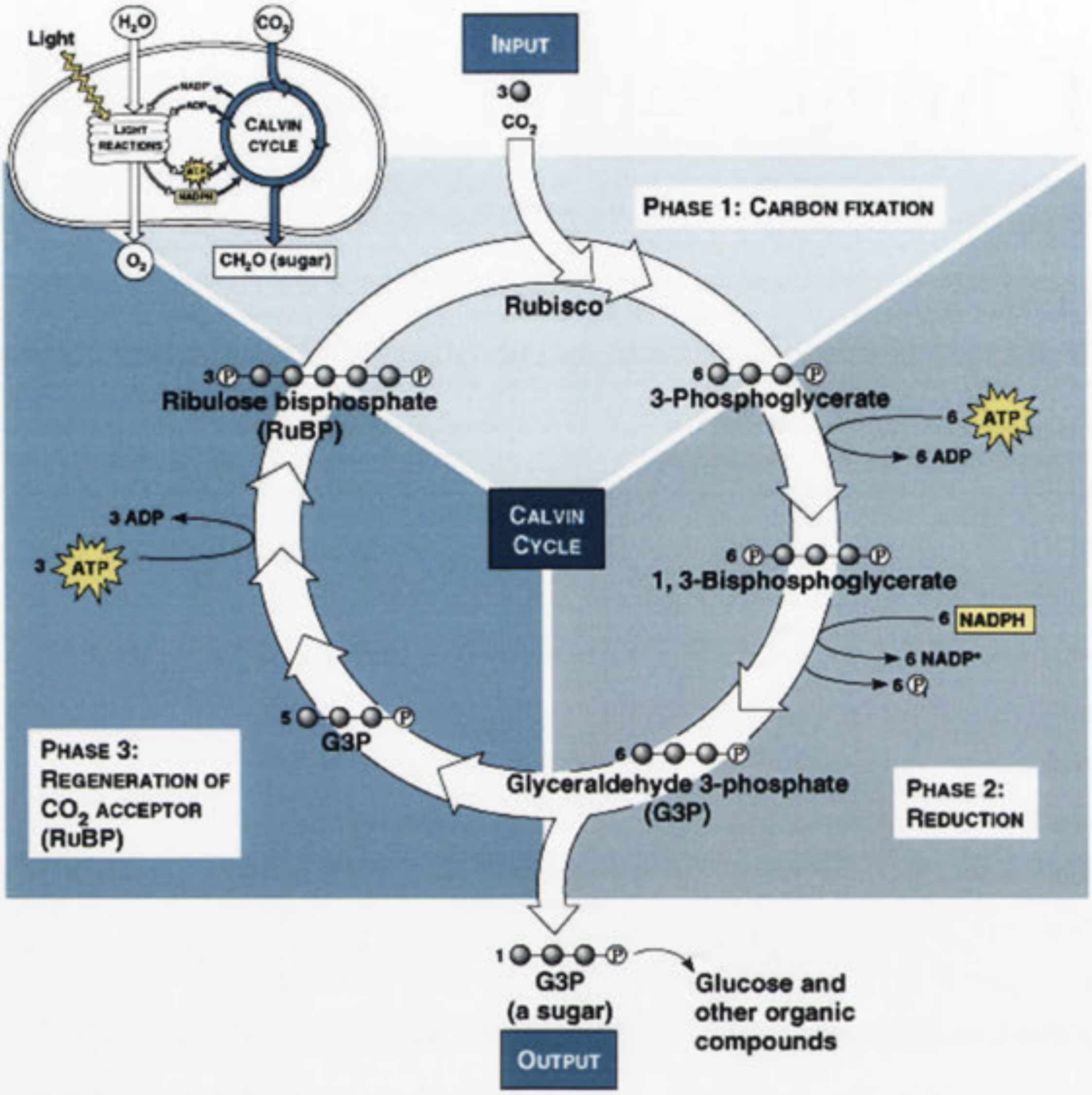


Figure 1.1 Overview of the Calvin cycle and carbon fixation

Rubisco is generally considered as the rate-limiting factor of photosynthesis in plants due to its inefficiency as a catalyst, compared with most enzymes. This is a compound of its slow catalytic rate of 3-5 s^{-1}, its use of O_2 as an alternative substrate and its low affinity for the desired substrate gaseous CO_2 from the atmosphere. To compensate for its inefficiency, Rubisco content in plants comprises up to 50% of total soluble leaf proteins, making it the most abundant protein on earth (Ellis, 1979).

For these reasons, Rubisco has been studied intensively as a prime target for genetic engineering to improve photosynthetic efficiency (Raines, 2006, Parry et al., 2007). As traditional crop-breeding methods now fail to deliver the productivity gains in food-grain production required to feed a growing global population, a "better Rubisco" is regarded as crucial to usher in a new "green revolution". Food, fiber, and fuel needs of the ever-increasing human population, shortages in the availability of water and land for agriculture, are challenges of the twenty-first century that would be impacted positively by successful manipulation of Rubisco in crop plants (Spreitzer and Salvucci, 2002).

Conceptual and technical advances in recent decades have increased our knowledge of the structure, function and regulation of Rubisco. Comprehensive studies on Rubisco have led to: (a) demonstration that a dimer of the Rubisco large subunit (L_2) is the basic catalytic unit of the Rubisco enzyme (Andrews, 1988, Gutteridge, 1991, Lee et al., 1991, Morell et al., 1997); (b) identification of the active-site residues and catalytic mechanism (Andersson et al., 1989, Knight et al., 1990, Hartman and Harpel, 1994, Cleland et al., 1998); (c) the finding that effector sites for Rubisco activase and RbcX, the Rubisco chaperones, are only found on large subunit (Larson et al., 1997, Ott et al., 2000, Saschenbrecker et al., 2007, Bracher et al., 2011); (d) understanding that the small subunit affects activity by influencing the conformation of the catalytic core of the large subunit (Andrews, 1988, Lee et al., 1991, Spreitzer, 2003); (e) determination of a variety of Rubisco atomic-level x-ray structures (Andersson and Taylor, 2003, Andersson and Backlund, 2008); (f) identification of new members of the Rubisco family from the green sulphur phototrophic bacterium *Chlorobium tepidum* and the heterotroph *Bacillus subtilis* (Hanson and Tabita, 2001, Tabita et al., 2008a, Tabita et al., 2008b); (g) discovery of post-translation modification of Rubisco (Houtz and Portis, 2003, Houtz et al., 2008), and (h) advanced understanding of Rubisco's catalytic chemistry via computational studies (Mauser et al., 2001, Kannappan and Gready, 2008, King et al., 1998). With the explosion of this new knowledge about Rubisco, there are reasons for confidence that improvement in Rubisco activity and crop productivity can be achieved.

1.3 Different forms of Rubisco

There are four forms of Rubisco found in nature (Table 1.1), each of which is placed in a separate category based on phylogenetic reconstructions (Tabita et al., 2008b). Forms I, II, and III catalyse the carboxylation and oxygenation of Ribulose 1,5-bisphosphate (RuBP), while form IV, also called the Rubisco-like protein (RLP), does not catalyse either of these reactions. Structurally, the RLPs lack key conserved active-site residues of Rubiscos and, therefore, do not bind RuBP. Of the four forms, the most abundant is form I; it is hexadecameric, consisting of eight large (L) and eight small (S) subunits (L_8S_8). The form I Rubiscos are further subdivided (phylogenetically) into a green branch, present in cyanobacteria, green algae and plants, and a red branch, present mainly in photosynthetic bacteria, red algae and phytoplankton (Tabita, 1999, Tabita et al., 2008a, Badger and Bek, 2008). A notable feature in plants and green algae is that the large subunit of Rubisco is encoded by a chloroplast gene (*rbcL*), whereas the small subunit is coded by a family of nuclear genes (*rbcS*).

Table 1.1 Summary of the different forms of Rubisco [a]

Rubisco form	Quaternary structure	Type of Organisms	Enzymatic Function
I-A (green)	L_8S_8	α, β, γ-Proteobacteria, Cyanobacteria, Prochlorales	CBB cycle [b]
I-B (green)	L_8S_8	Cyanobacteria, Prochlorales, Eukaryotes-Viridiplantae (Streptophyta, Chlorophyta), Euglenozoa	CBB cycle
I-C (red)	L_8S_8	α, β-Proteobacteria, Chloroflexi	CBB cycle
I-D (red)	L_8S_8	α, β, γ-Proteobacteria, Eukaryotes-stramenopiles, Rhodophyta, Haptophyceae	CBB cycle
II	$(L_2)_n$	α, β, γ-Proteobacteria, Eukaryotes-Alveolata (Dinophyceae) dinoflagellates	CBB cycle
III	$(L_2)_n$	Methanogenic and thermophilic crenarchaeota, thermophilic and halophilic euryarchaeota	RuPP [c] pathway
IV	L_2	α, β, γ-Proteobacteria, Chloroflexi, Clostridia, Non-methanogenic euryarchaeota, Chlorobia, Firmicutes	Methionine salavage pathway

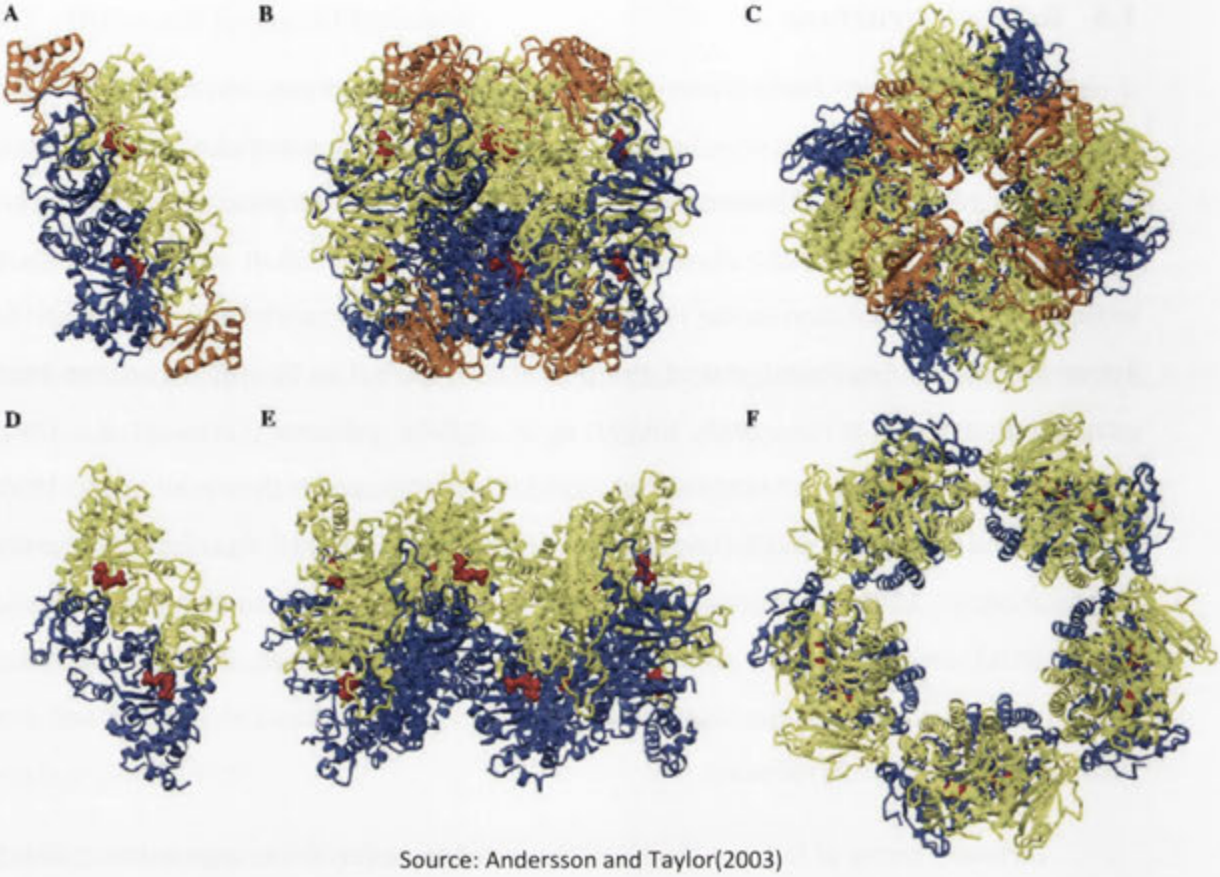
[a] Information presented in this table was obtained from various review resources including Tabita et al.(2008b) and Badger and Bek (2008). [b] CBB stands for Calvin-Benson-Bassham cycle. [c] RuPP stands for 5-phospho-D-ribose-1-pyrophosphate.

1.4 Rubisco structure

All four Rubisco holoenzyme forms are structurally unique, but in all forms the basic catalytic unit is the dimer (L_2) made of two large subunits encoded by the gene *rbcL*. The first x-ray structure of Rubisco was determined to 2.9 Å resolution (Schneider et al., 1986) from the recombinant dimeric enzyme from *Rhodospirillum rubrum*, a form II enzyme. The low-resolution model revealed an eight-stranded parallel α/β barrel with the active site at the C-terminal end of the β-strands (Figure 1.2, D). The structures from spinach (Andersson et al., 1989, Knight et al., 1990), tobacco (Curmi et al., 1992, Schreuder et al., 1993a, Schreuder et al., 1993b), *Synechococcus* (Newman et al., 1993, Newman and Gutteridge, 1993, Newman and Gutteridge, 1994), red alga, *Galdieria partita* (Sugawara et al., 1999), hyperthermophilic archaeon, *Thermococcus kodakaraensis* (Kitano et al., 2001), *Chlamydomonas* (Mizohata et al., 2002, Taylor et al., 2001), Rubisco-like protein from the green sulfur bacterium *Chlorobium tepidum* (Li et al., 2005b) and rice (Matsumura et al., 2012) followed.

Different forms of Rubisco have variable arrangements of the large-subunit dimer. Rubisco from higher plants, algae, and cyanobacteria is a hexadecamer of molecular mass 550 kDa composed of eight large (L: 50–55 kDa) subunits and eight small (S: 12–18 kDa) subunits (Figure 1.2, B, C). In form I Rubiscos, the 4-fold axis relates four L_2 dimers into a core of eight large subunits, $(L_2)_4$, with two groups of four small subunits capping the L_8 core to form an L_8S_8 molecule (Figure 1.2, B, C). Rubiscos from forms II and III lack small subunits, containing only L-subunits arranged into L_2 to $(L_2)_n$ complexes (Figure 1.2, D, E, F). Rubisco from some dinoflagellates and purple nonsulfur bacteria (e.g., *Rhodospirillum rubrum*) is a homodimer of two such L subunits related by a twofold rotational symmetry (Figure 1.2, D).

Rubisco-LSU sequences are highly conserved in Angiosperms, and any differences in length occur primarily at the N and C termini. Throughout this thesis, numbering of large subunit residues will be based on the sequence of the spinach large subunit.



Source: Andersson and Taylor(2003)

Figure 1.2 Different arrangements of the quaternary structure of Rubisco showing the molecular symmetry. (A) L_2S_2 unit of type I Rubisco from spinach viewed down the twofold symmetry axis. (B and C) Entire L_8S_8 hexadecamer viewed down the twofold and fourfold axes. (D) Dimeric type II Rubisco from *Rhodospirillum rubrum* showing the twofold symmetry; (E and F) L_{10} Rubisco from *Thermococcus kodakaraensis* viewed down the twofold and fivefold axes, respectively. Large subunits are blue and green, small subunits are orange, and substrate bound in the active-site are displayed as red spheres.

1.4.1 Rubisco large subunit

Despite apparent differences in amino acid sequence the overall fold of the large subunit is conserved in all forms of Rubisco: a smaller N-terminal domain of ~150 amino acids consisting of a five β-strands with two α-helices on one side of the sheet and a larger C-terminal domain (~320 amino acids) folded as a classic α/β-barrel (Figure 1.3). The C-terminal domain consists of eight consecutive βα-units arranged as an eight-stranded parallel α/β-barrel structure. The highly conserved "active-site" residues reside within the α/β-barrel domain, with a few residues supplied by the N-terminal domain of the adjacent

L-subunit. Loop 6, which is known to play important role in catalysis, is located between helix α6 and strand β6 in C-terminal α/β-barrel domain as shown in Figure 1.3.

All Rubisco x-ray crystal structures show very similar Cα backbone structures; the overall secondary structures of the Rubisco-LSUs from diverse sources including RLPs are conserved as well. In general, large subunits of forms I–IV display 25–30% sequence identity across different forms of Rubisco. Despite this relatively large divergence on the level of sequence, differences are localized to a few loops, specifically the loop between strands βC and βD (Tabita et al., 2007, Andersson and Backlund, 2008).

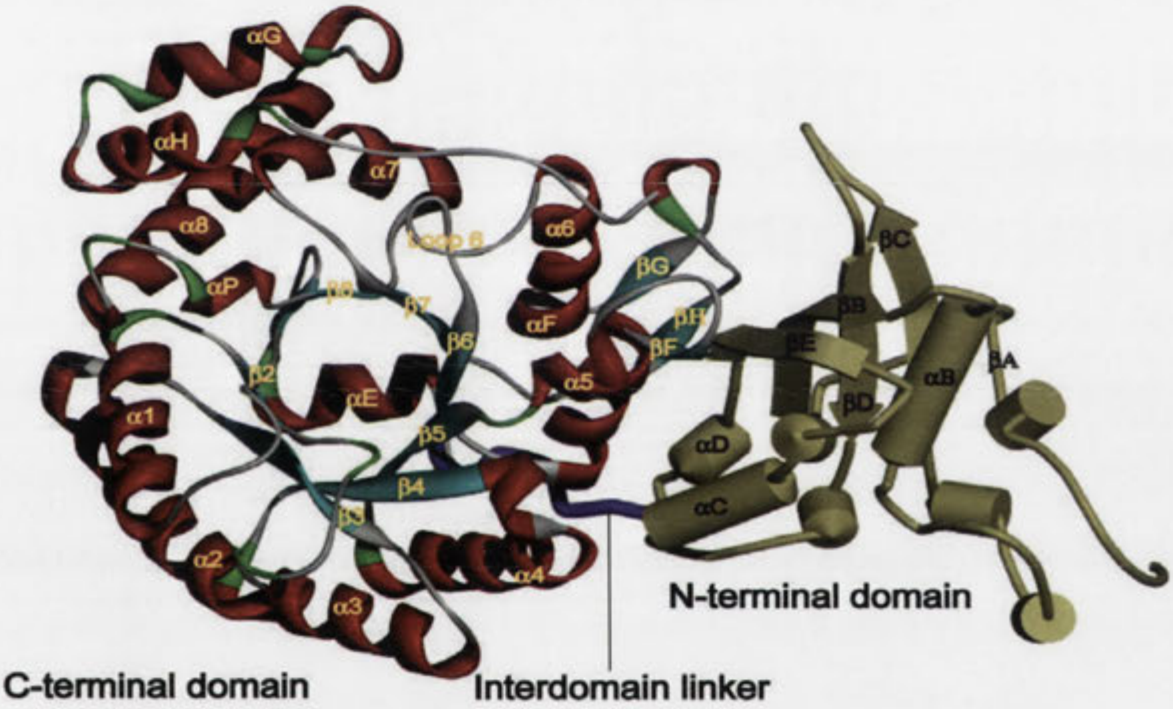


Figure 1.3 Monomer of Rubisco large subunit showing N-terminal, C-terminal domain and secondary structure (8RUC)**.** Numbering of helices and strand follows Knight et al. (1990). Strand β1 is not visible. N-terminal domain is yellow, interdomain linker is purple, C-terminal domain α-helices and β-strands are red and cyan, respectively. N-terminal domain is depicted as cylinder (α-helices) and arrows (β-strands); C-terminal domain is shown as ribbons. Important loop 6 between β6 and α6 is marked.

Thus, the functional unit of Rubisco is a L_2 dimer (Figure 1.4) with two active sites located at the L-L interface. The active site (Figure 1.5) is shaped like a funnel and is mainly formed by the eight loop regions that connect the eight β-strands with corresponding helices in the α/β barrel in the C-terminal domain (Andersson et al., 1989). The N-terminal domain of the second subunit covers part of the top of the active site. In particular two loop regions in this domain provide residues to active site.

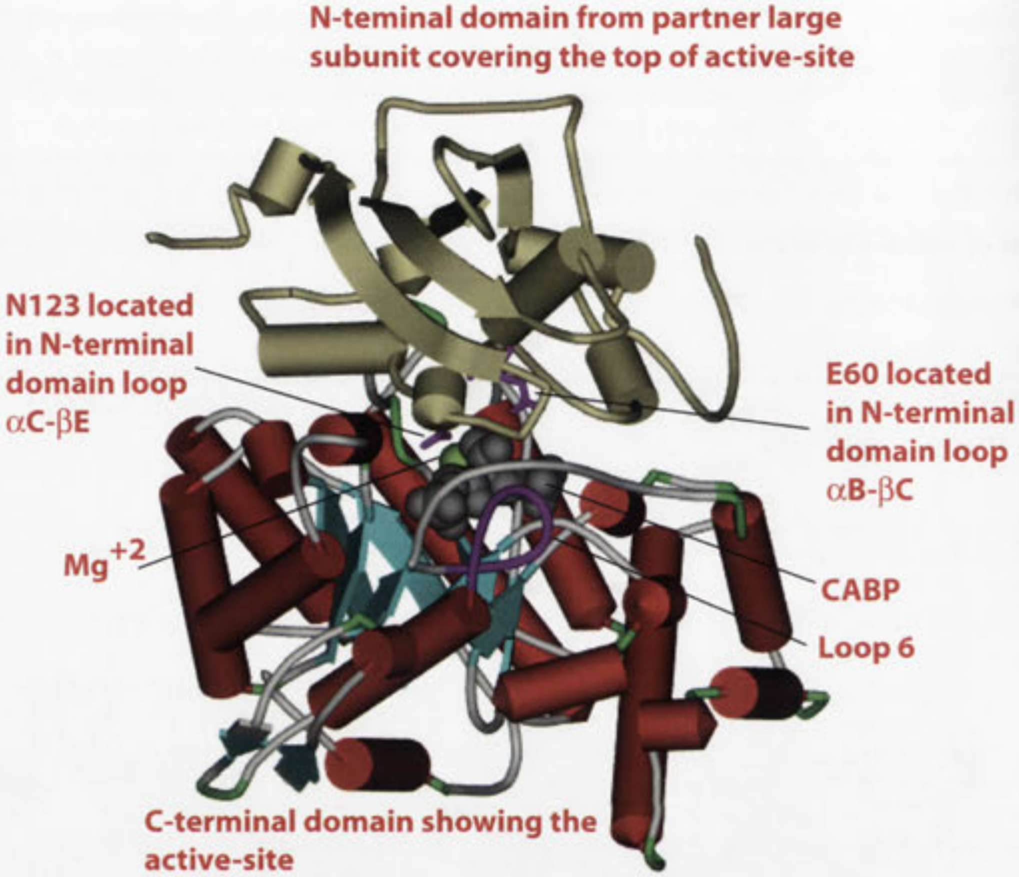

N-teminal domain from partner large subunit covering the top of active-site
N123 located in N-terminal domain loop αC-βE
E60 located in N-terminal domain loop αB-βC
Mg+2
CABP
Loop 6
C-terminal domain showing the active-site

Figure 1.4 L_2 dimer of Spinach Rubisco large subunit showing the active-site with CABP and Mg^{+2} (8RUC)**.** The N and C-terminal of partner subunit are omitted for clarity. CABP is shown in grey while Mg^{2+} is in green. The N-terminal domain is depicted in yellow. The C-terminal helices are red, β-strands are colored cyan. Loop 6 and N-terminal domain active-site residues are shown in purple. E60 and N123 are shown in stick representations. CABP stands for 2-carboxyarabinitol-1,5-diphosphate, an inhibitor of Rubisco's catalytic reaction. Mg stands for magnesium ion.

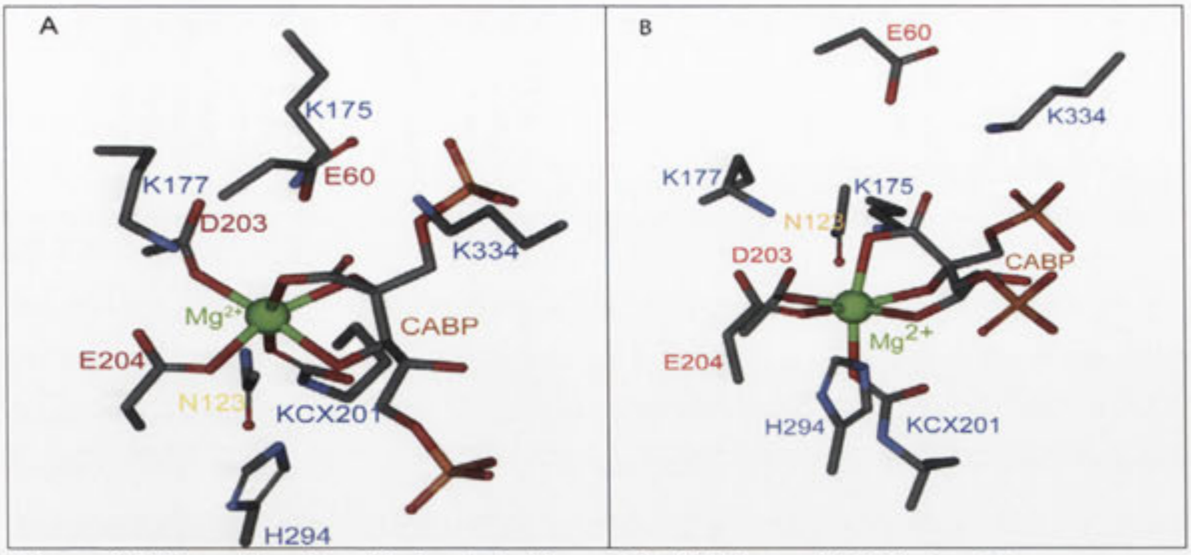


Figure 1.5 The active site of Spinach Rubisco with CABP bound (8RUC)**.** A and B, Cartoon representations of the two different orientations of the active site residues rotated at 180 degrees and CABP. Side chains of active site residues are shown in sticks. CABP is shown in orange. CABP stands for 2-carboxyarabinitol-1,5-diphosphate, an inhibitor of Rubisco's catalytic reaction. Mg stands for magnesium ion.

1.4.2 Rubisco small subunit

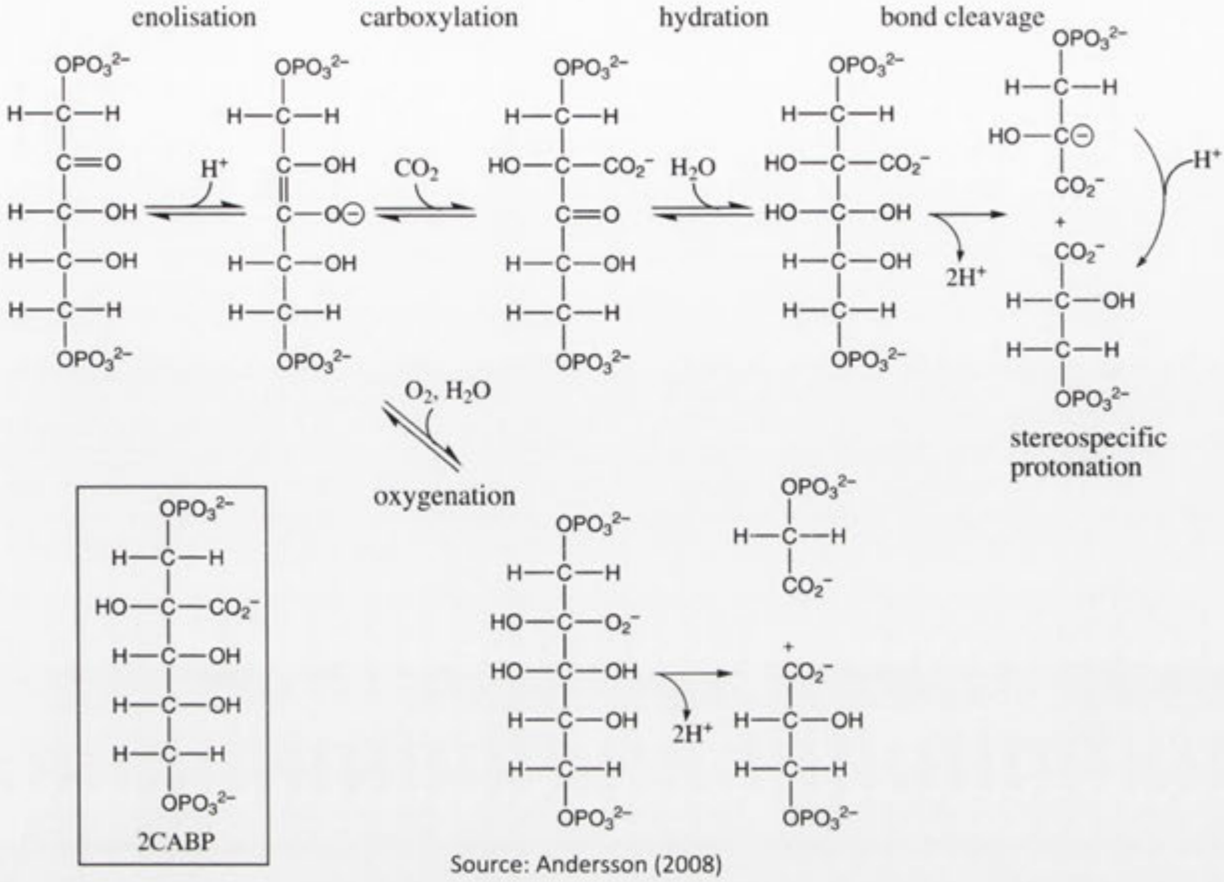
The common core structure of the small subunit consists of a four-stranded anti-parallel β-sheet covered on one side by two helices (Knight et al., 1990). The small subunit of Rubisco from cyanobacteria and non-green algae were found to be different from those of higher plants and green algae in two distinct locations i.e. the loop between β-strands A and B (βA-βB-loop) and the carboxy-terminus (Spreitzer, 2003). The βA-βB-loops of four small subunits line the openings of the solvent channel in the holoenzyme. Rubiscos from prokaryotes and non-green algae have only 10 residues in the βA-βB-loop, but Rubiscos from higher plants and green algae have 22 and 28 residues, respectively (Andersson and Backlund, 2008). To compensate for the short βA-βB-loop, the small subunits of non-green algae and some prokaryotes display carboxy-terminal extensions that form β-hairpin structures in the spaces that are normally occupied by the longer βA-βB-loops of the green algal and plant enzymes (Hansen et al., 1999, Sugawara et al., 1999). The arrangement of the small subunits on the L-subunit octamer suggests a structural function of the small subunit, i.e. in assembly of the large catalytic subunits. However, considering that Rubisco without small subunits have the lowest specificity values, they may contribute significantly to the differences in kinetic properties observed between form I and forms II/III Rubiscos (Andersson and Backlund, 2008).

1.5 Rubisco catalysis

The active-site residues in Rubiscos are totally conserved among forms I, II, and III. Accordingly, the basic steps of activation and the multi-step complex catalytic reaction are also similar in these forms of Rubiscos. Extensive structure-function relationship studies on Rubisco using x-ray crystallographic studies of Rubisco complexes, chemical modification, site-directed mutagenesis, molecular dynamics calculations, and quantum chemical analyses have resulted in definition of the roles of its active site residues and have provided insights into subtle alterations in its conformation at different stages of the reaction.

1.5.1 Rubisco catalyzes both carboxylation and oxygenation of RuBP

Rubisco requires activation, prior to catalysis by carbamylation of the active-site Lys201 (Lorimer and Miziorko, 1980) by a CO_2 molecule; this CO_2 molecule is distinct from the substrate-CO_2. The carbamylated Lys201 is stabilized by the binding of a magnesium ion to the carbamate. Subsequently, RuBP binds to Rubisco, and a complex five-step reaction adds a CO_2 and a water molecule to RuBP, followed by its cleavage and release of two 3-phosphoglycerate (3PGA) molecules (Figure 1.6). The electrostatic similarity between O_2 and CO_2 and high concentration of atmospheric O_2 (O_2 21% vs. CO_2 0.04%) in the present day atmosphere make it difficult for Rubisco to efficiently differentiate between them. Oxygenation of RuBP instead of carboxylation produces one molecule each of 3PGA and 2-phosphoglycolate. The 2-phosphoglycolate is recycled back to 3PGA via photorespiration, an energy-consuming process that releases fixed carbon as CO_2 (Peterhansel et al., 2008).



Source: Andersson (2008)

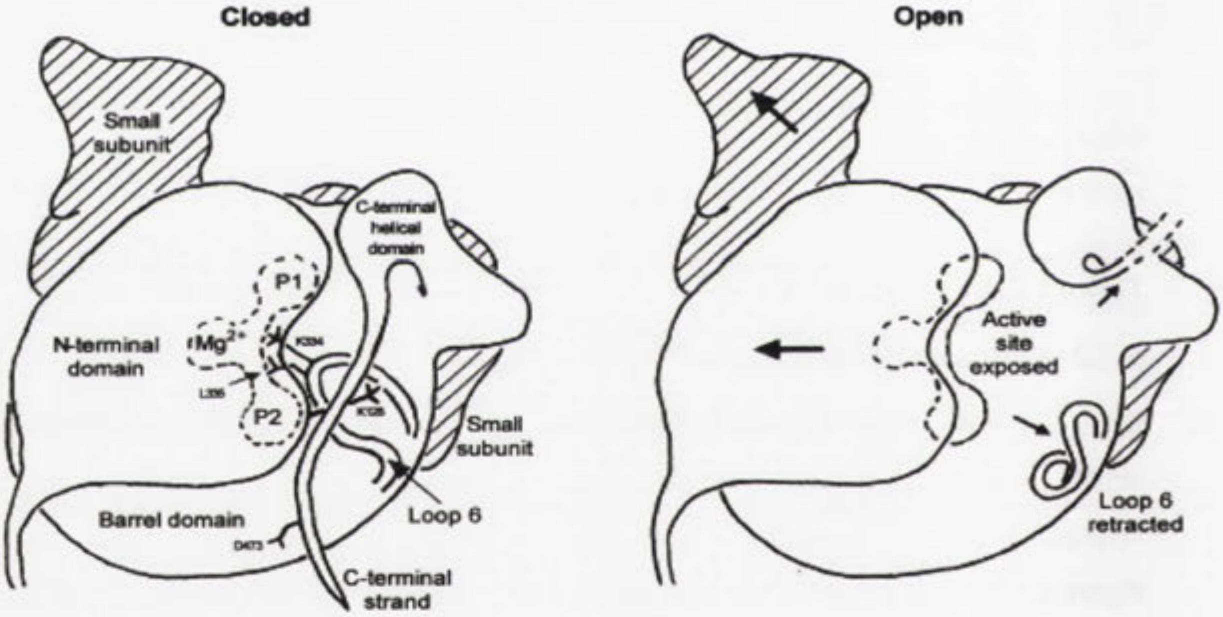
Figure 1.6 Main reactions catalysed by Rubisco, carboxylation and oxgenation of RuBP.

1.5.2 Conformational changes during catalysis

At some point during catalysis in Rubisco a conformational change closes the active site and inhibits water from entering the active site. The transition from "open" to "closed" form in Rubisco requires movements of loop 6 (residues 331–338) as shown in Figure 1.3, 1.4 and 1.7, the C-terminal tail (residues 463 to the C-terminal end) as depicted in Figure 1.7, and a loop from the N-terminal domain (residues 63–69) of the adjacent large subunit of the L_2 dimer(Schreuder et al., 1993a, Taylor and Andersson, 1996, Duff et al., 2000) The importance of loop 6 for catalysis and specificity has been demonstrated by genetic selection and site-directed mutagenesis (Chen and Spreitzer, 1989, Chen et al., 1991).

The Rubisco active site is either "open" or "closed" (Duff et al., 2000): a) open with the active site occupied by loosely bound substrates or with no ligand (Figure 1.7, Open),

b) closed with substrates or inhibitors tightly bound with no solvent access (Figure 1.7, Closed). Other than the movement of loop 6 to cover the opening of the α/β-barrel, the transition between open and closed forms also entails a rigid-body movement to bring together the N- and C-terminal domains of adjacent subunits. The packing of the C-terminal tail against loop 6 completes the closure (Schreuder et al., 1993a, Taylor and Andersson, 1996). As shown by site-directed mutagenesis, the carboxy-terminus is not absolutely required for catalysis, but is needed for maximal activity and stability (Morell et al., 1990, Ranty et al., 1990, Gutteridge et al., 1993, Esquivel et al., 2002). Residue Asp473 has been proposed as a latch responsible for placing the large-subunit carboxy-terminus tail over loop 6 and stabilizing the closed conformation required for catalysis (Duff et al., 2000).



Source: Duff et al.(2000)

Figure 1.7 Schematic representation of the closed and open conformation of an active site of Rubisco. The N-terminal domain from the adjacent large subunit covers the top of the α/β barrel. In the open state, the N-terminal domain has moved left and correspondingly one small subunit moves up and to the left. On opening, Loop 6 of the barrel domain retracts to extend helix 6 in a stable configuration and the C-terminal tail pulls away from the active site and the α/β barrel domain and is usually disordered in the open state crystal structures. In the closed state, there is no solvent access to the substrate and substrate binding can be divided into three distinct zones; the P1-binding site, the P2-binding site and the metal site.

1.5.3 The natural catalytic diversity of Rubisco

The efficiency with which CO_2 is able to compete with O_2 is quantified by the CO_2/O_2 specificity factor ($S_{C/O}$) and is defined as V_cK_o/V_oK_c, where V_c and V_o are the maximal velocities of carboxylation and oxygenation, respectively, and K_c and K_o are the Michaelis constants for CO_2 and O_2, respectively (Laing et al., 1974). $S_{C/O}$ is the ratio of the carboxylase to oxygenase rate when CO_2 and O_2 are present at equal concentrations. Thus, the relative rates for carboxylation and oxygenation are defined by the product of the specificity factor and the ratio of CO_2 to O_2 concentrations at the active site.

$S_{C/O}$ values differ substantially among Rubisco enzymes from divergent species (Table 1.2). Form II enzymes of α-proteobacteria have the lowest $S_{C/O}$ values (~6-8), whereas Form I enzymes from Rhodophyta (red algae) have the highest (100-160). The $S_{C/O}$ values of different higher plant enzymes are very similar i.e. ranging from 80-100, whereas Rubiscos from cyanobacteria and green algae have lower values i.e. ranging from 40-60. However the increase in the specificity in the higher plant enzyme has been at the expense of the catalytic turnover rate of the carboxylation, e.g. cyanobacteria displaying low specificity values and high turnover rates whereas higher plants have high specificity values coupled to low turnover rates. In addition, cyanobacteria, green algae, and C_4 plants have CO_2-concentrating mechanisms (Kaplan and Reinhold, 1999, Matsuoka et al., 2001), reducing the importance of $S_{C/O}$ in vivo. It also appears that environmental factors such as temperature have important roles in influencing evolution of Rubisco's carboxylation rate. For example, plants from cooler habitats have a Rubisco with lower specificity and higher turnover rate than those of warm and dry habitats (Sage, 2002, Galmes et al., 2005). Unfortunately, comprehensive catalytic analysis have only been done for relatively few Rubiscos, limiting our capacity to fully appreciate the connections between catalytic and sequence diversity and the influence of temperature on the activity of evolutionarily diverse Rubiscos (Whitney et al., 2011a).

Table 1.2 Catalytic properties for different Rubisco forms determined at 25°C [a]

Organism	CCM	Form	v_{CO2} (s^{-1})	$K_m^{CO_2}$ (μM)	$S_{C/O}$	Reference
Cyanobacteria						
Anabaena variabilis	Present	$\mathrm{I}_{(green)}$	n.m.	n.m.	43	Badger (1980)
Synechococcus 7002	Present	$\mathrm{I}_{(green)}$	13.4	246	52	Andrews and Lorimer,(1985)
Synechococcus 6301	Present	$\mathrm{I}_{(green)}$	11.8	200	42	Mueller-Cajar and Whitney (2008)
Green algae						
Chlamydomonas reinhardtii	Present	$\mathrm{I}_{(green)}$	2.1	31	61	Genkov et al.(2010)
Euglena gracilis	Present	$\mathrm{I}_{(green)}$	n.m.	n.m.	54	Jordan and Ogren (1981)
C_4 higher plants						
Amaranthus edulis	Present	$\mathrm{I}_{(green)}$	4.1	18.2	77	Kubien et al. (2008)
Amaranthus hybridus	Present	$\mathrm{I}_{(green)}$	3.8	16.0	82	Jordan and Ogren (1981)
Flaveria australasica	Present	$\mathrm{I}_{(green)}$	3.8	22.0	77	Kubien et al. (2008)
Flaveria bidentis	Present	$\mathrm{I}_{(green)}$	4.2	20.2	76	Kubien et al. (2008)
Flaveria kochiana	Present	$\mathrm{I}_{(green)}$	3.7	22.7	77	Kubien et al. (2008)
Flaveria trinervia	Present	$\mathrm{I}_{(green)}$	4.4	17.9	77	Kubien et al. (2008)
Sorghum bicolor	Present	$\mathrm{I}_{(green)}$	5.4	30.0	70	Sage and Seemann (1993)
Zea mays	Present	$\mathrm{I}_{(green)}$	4.1	21.2	75	Kubien et al. (2008)
C_3 higher plants						
Atriplex glabriuscula	Absent	$\mathrm{I}_{(green)}$	n.m.	n.m.	87	Badger and Collatz (1977)
Chenopodium alba	Absent	$\mathrm{I}_{(green)}$	2.9	11.2	79	Kubien et al. (2008)
Flaveria cronquistii	Absent	$\mathrm{I}_{(green)}$	3.1	10.8	81	Kubien et al. (2008)
Flaveria pringlei	Absent	$\mathrm{I}_{(green)}$	3.1	12.0	81	Kubien et al. (2008)
Helianthus annuus	Absent	$\mathrm{I}_{(green)}$	2.9	n.m.	84	Sharwood et al. (2008)
Nicotiana tabacum	Absent	$\mathrm{I}_{(green)}$	3.4	11.0	82	Whitney et al. (1999)
Oryza sativa	Absent	$\mathrm{I}_{(green)}$	n.m.	n.m.	85	Kane et al.(1994)
Spinacia oleracea	Absent	$\mathrm{I}_{(green)}$	3.2	12.1	80	Kubien et al. (2008)
Triticum aestivum	Absent	$\mathrm{I}_{(green)}$	2.5±0.2	14±3	98±4	Zhu et al. (1998)
Non-green algae						
Cylindrotheca fusiformis	? [b]	$\mathrm{I}_{(red)}$	2.0	36.0	111	Read and Tabita (1994)
Cylindrotheca N1	?	$\mathrm{I}_{(red)}$	0.8	31.0	106	Read and Tabita (1994)
Galdieria sulfuraria	?	$\mathrm{I}_{(red)}$	1.2	3.3	166	Whitney et al. (2001)
Griffithsia monilis	?	$\mathrm{I}_{(red)}$	2.6	9.3	167	Whitney et al. (2001)
Olisthodiscus	?	$\mathrm{I}_{(red)}$	0.8	59.0	100	Read and Tabita (1994)
Phaeodactylum tricornutum	Present	$\mathrm{I}_{(red)}$	3.4	28.0	113	Whitney et al. (2001)
Porphyridium	?	$\mathrm{I}_{(red)}$	1.6	22.0	129	Read and Tabita (1994)
Bacteria						
Chromatium vinosum	Anaerobic	II	6.7	37	41	Jordan and Chollet (1985)
Rhodospirillum rubrum	Anaerobic	II	7.3	67	12	Morell et al. (1990)
Riftia pachyptila symbiont	Anaerobic	II	1	240	9	Robinson et al. (2003)
Archaea						
Methanocococcus burtonii	?	III	2	130	1.2	Alonso et al. (2009)
Methanocococcus Jannaschii	?	III	n.m.	n.m.	0.5	Watson et al. (1999)
Thermococcus kodakaraensis	?	III	0.3	52	11	Yoshida et al. (2007)

[a] Information presented in this table was obtained from various review resources including Whitney et al.(2011a) and Tcherkez et al. (2006). [b] we don't know yet if these organism have CCM (Carbon concentrating mechanism)

Considering that Rubisco active-site residues are totally conserved and its fold is also well conserved, it is intriguing why Rubiscos from different sources often show vastly different catalytic properties and distinct kinetic behaviour. There is evidence that the catalytic diversity of Rubiscos originates from residues distant from the active-site, as mutagenesis of such residues in *R. rubrum*, *Synechococcus*, *Chlamydomonas*, or tobacco Rubisco has shown they are required for maximal rates of catalysis (reviewed in Spreitzer and Salvucci (2002)).

1.6 Rubisco large subunit and its interactions

Rubisco interacts with many proteins at different stages in its life cycle: RbcX helps in assembly (Saschenbrecker et al., 2007), Rubisco-SSU is part of the holoenzyme and Rubisco activase (Portis, 2003) assists in the re-activation by releasing the inhibitors. In the last two decades detailed studies on Rubisco interactions have led to many novel insights on their molecular basis. As noted before, a significant finding of all these studies is that all the effector sites are located on the Rubisco-LSU. The small subunit has almost no role to play in these interactions. In this section, I summarize our current knowledge of Rubisco's interactions with RbcX, the Rubisco-SSU and Rubisco activase.

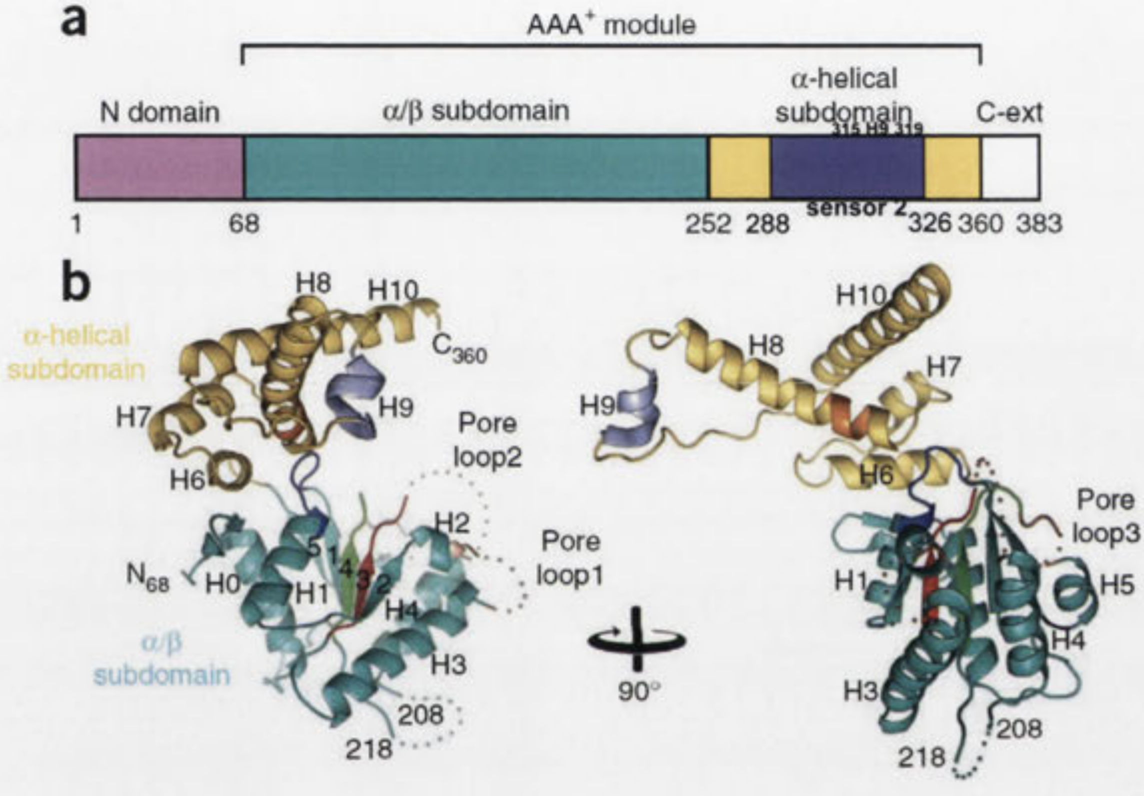
1.6.1 Rubisco activase

As noted in the previous section, lysine 201 must be carbamylated and bound to an Mg^{2+} ion for Rubisco to become catalytically active. Binding of RuBP to uncarbamylated Rubisco, or of 2-carboxy-D-arabinitol 1-phosphate (CA1P) to the carbamylated enzyme at night, results in the trapping of the sugar phosphate and inhibition of the enzyme (Portis, 1992). To ensure efficient photosynthesis, the inhibitory sugar must be removed by the protein Rubisco activase, an ATPase of the class of AAA+ proteins associated with various cellular functions that require ATP for their energy-dependent reactions (Portis, 2003).

The first evidence for a physical association between Rubisco and Rubisco activase came from the observation that the activase from two *Solanaceae* species (tobacco and petunia) did not activate Rubisco from several non-*Solanaceae* species (e.g., spinach, barley, *Chlamydomonas*) and vice versa (Wang et al., 1992). A comparison of the Rubisco large subunit sequences of these two groups (*Solanaceae* and non-*Solanaceae*) revealed that a small subset of residues clustered on the surface of the large subunit is substantially different in the two groups (Portis 1995). Site-directed mutagenesis studies (Larson et al., 1997, Ott et al., 2000) in *Chlamydomonas* found that Pro-89 to Arg, Pro-89 to Ala and Asp-94 to Lys substitutions resulted in Rubiscos that could no longer be activated by spinach activase, but the mutant enzymes could now be activated by tobacco activase. Therefore, the loop (residue 90-96) between β-strands C (residue 83-89) and D (residue 97-103) in the N-terminal domain of the large subunit was identified as an activase-recognition region.

Rubisco activase belongs to a superfamily of proteins possessing the "Sensor 2" domain. This domain is involved in substrate recognition (Wickner and Maurizi, 1999, Smith et al., 1999). Li et al.(2005a) have demonstrated that the "Sensor 2" domain (Figure 1.8a) in the C-terminal region of Rubisco activase is responsible for differences in Rubisco substrate recognition and identified two amino acid residues i.e. residue 311 and 314 in this region as key for differential recognition between activases from Solanaceous and non-Solanaceous plant species.

In the absence of structural information for Rubisco activase, the mechanistic details of Rubisco activation by Rubisco activase remained elusive for two decades. However in 2011, Stotz et al. (2011) reported resolution of the 2.95-Å crystal structure of a "short form" of Rubisco activase A from tobacco which lacks 67 amino acids from the N-terminal and 23 amino acids from the C-terminal of the 383-residue long protein. They found that Rubisco activase is composed of a 67-residue N-terminal domain, a classical AAA+ module consisting of an N-terminal nucleotide-binding α/β subdomain and an α-helical subdomain (Hanson and Whiteheart, 2005), followed by a 23-residue C-terminal extension (Figure 1.8a). Moreover, they suggested that Rubisco activase functions as a hexameric AAA+ enzyme with a central pore that mediates Rubisco remodeling. Most importantly, Stotz et al. (2011) propose that helix H9 (Figure 1.8b) of the α-helical subdomain with the N-terminal domain of Rubisco activase, makes up the substrate recognition motif for Rubisco. In addition they speculate that Rubisco activase may engage an exposed loop segment (the βC-βD-loop in Rubisco-LSU) of green-type Rubisco because the central pore in the Rubisco activase hexamer is substantially wider (~36 Å) than found in red-type activase CbbX (25 Å) (Mueller-Cajar et al., 2011). The structure of N- and C-terminally truncated Rubisco activase provides a basic structural framework for future detailed mechanistic analysis of Rubisco activation in plants; further research will be needed to assess the validity of their conclusions.



Source: Stotz et al. (2011)

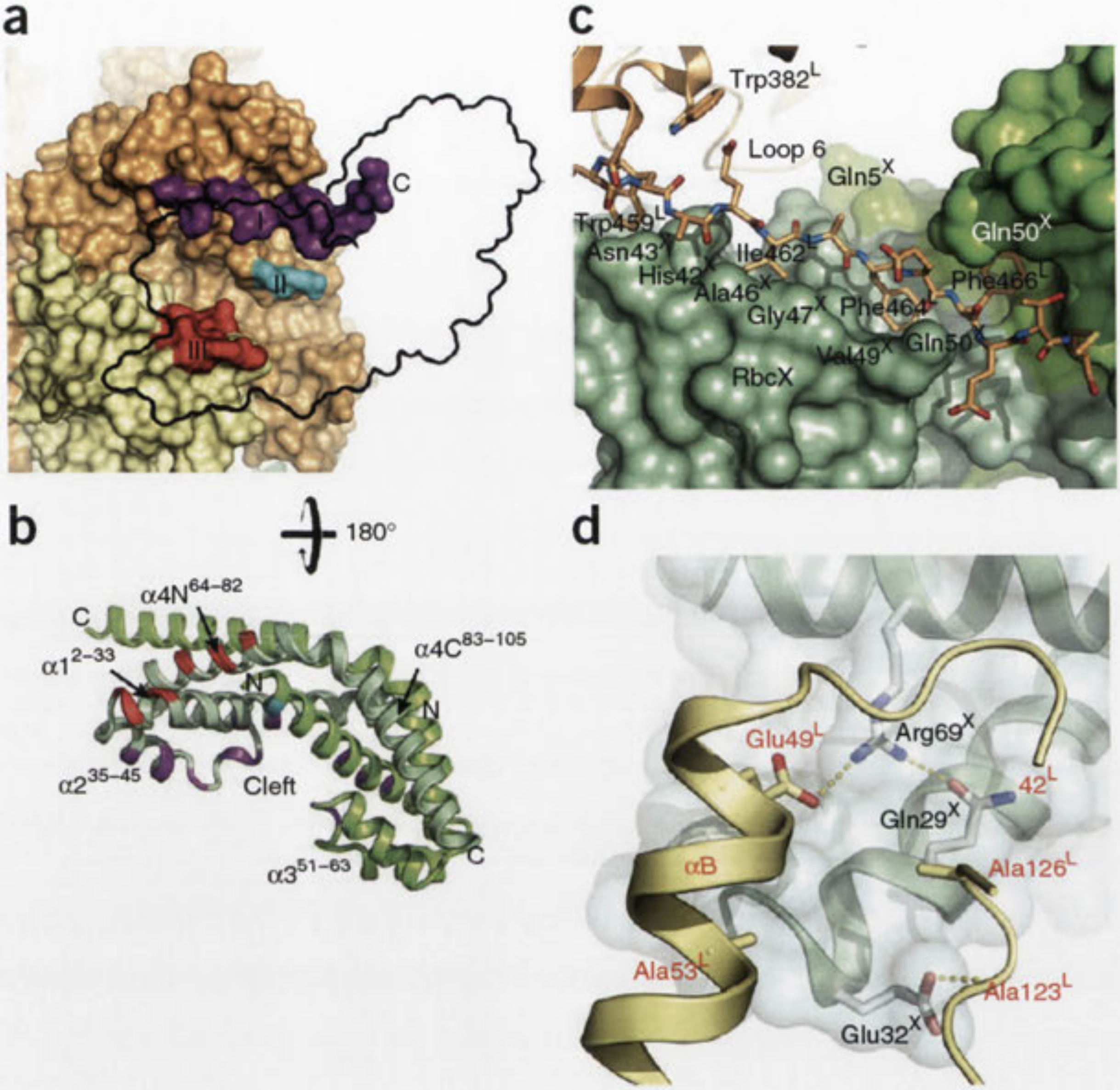
Figure 1.8 Structural and functional analysis of Rubisco activase. (a) Schematic representation of the domain structure of Rubisco activase. C-ext, C-terminal extension. Location of sensor 2 domain (position 288 to 326) and helix H9 (position 315-319) is marked. (b) Ribbon representation of the crystal structure of Rubisco activase from Tobacco. Disordered loops are indicated by dotted lines. Two views related by 90° are shown. The α/β and the α-helical subdomains are indicated in teal and gold, respectively. The canonical AAA+ structural motifs are indicated as follows: Walker A (dark blue), Walker B (red), sensor I (green) and sensor II (orange). The disordered pore loops are indicated by dots colored with attached secondary structure. The specificity helix (H9) is shown in violet. Secondary structure elements, pore loops, and chain termini are indicated.

1.6.2 Rubisco large subunit interactions with RbcX

Studies with cyanobacteria L_8S_8 Rubisco have shown that chaperonin-folded L-subunits interact with RbcX, a Rubisco-specific chaperone whose gene (rbcX) is often located between *rbcL* and *rbcS* in cyanobacteria (Saschenbrecker et al., 2007, Liu et al., 2010). RbcX dimers facilitate the assembly of L-subunits into $(L_2)_4$ complexes and are then displaced by the stable binding of S-subunits to produce the native L_8S_8 enzyme (Bracher et al., 2011).

RbcX, an ~15-kDa protein, is an arc-shaped homo-dimer ($RbcX_2$) and functions downstream of the folding of Rubisco large subunits (Saschenbrecker et al., 2007, Tanaka et al., 2007, Tarnawski et al., 2008). It binds the flexible C-terminal tail sequence of the Rubisco large subunit —EIKFEFD—in a central hydrophobic cleft. This sequence motif is conserved in all form I large subunits.

Recently, Bracher et al. (2011) resolved the 3.2-Å crystal structure of an assembly intermediate of cyanobacterial Rubisco, consisting of eight Rubisco large subunits and eight $RbcX_2$ molecules. This showed three contact areas between $RbcX_2$ and the dimer of large subunits (L_2) (Figure 1.9a). The largest interface, area I comprises the C-terminal peptide of the large subunit, residues 458L-468L (LWKEIKFEFET) (Figure 1.9c). A second interface area is formed by residues Leu332L and Glu333L and the N terminus of one $RbcX_2$ chain. These residues belong to "loop 6" of the Rubisco-LSU, which is involved in regulating substrate access to the active site and is stabilized in the open conformation by helix α1 of $RbcX_2$. The third interface of $RbcX_2$ with the adjacent large subunit of the large subunit dimer (Figure 1.9d) comprises large subunit residues 42L–46L, 49L and 53L in helix αB as well as the preceding loop, and residues 123L-126L in the loop connecting helix αC and β-strand βE. All these residues are highly conserved in sequences of form I large subunits.



Source: Bracher et al. (2011)

Figure 1.9 The interaction of the RbcX$_2$ and Rubisco in molecular detail. (a) Close-up view showing surfaces on the anti-parallel dimer of the large subunit that interact with RbcX$_2$. The outline of the bound RbcX$_2$ is shown for orientation. The interaction surfaces area I (purple) and area II (cyan) are located on one large subunit, whereas area III (red) is located on the adjacent subunit of the large subunit dimer. (b) Ribbon diagram of the RbcX$_2$ dimer showing the contact regions with Rubisco large subunit, colored as in (a). RbcX$_2$ is rotated 180° relative to the view shown in (a). (c) Close-up view of the RbcX$_2$ interface with the C-terminal peptide of the large subunit (area I). RbcX$_2$ is shown in surface representation, whereas the C-terminal peptide of the large subunit is shown in stick model. In the background, the area II contact between loop 6 of the large subunit and residue Gln5 of one RbcX chain (green) is visible. (d) Cutaway view of the RbcX$_2$ interface with the opposing large subunit (area III). The surface of RbcX$_2$ is shown as a transparent skin. Crucial contact residues in RbcX and the large subunit are shown in stick representation.

1.6.3 Interface between Rubisco large and small subunits

The small subunits are located in crevices formed between the ends of adjacent L_2 dimers. Each small subunit is in contact with three different large subunits from two different L_2 dimers as well as with two neighbouring small subunits (Knight et al., 1990). In the following, the description of the subunit interactions between small and large subunits is limited to consideration only of the interactions of one small subunit, S, with the three large subunits B, C and D, which are in contact with S.

As shown in Figure 1.10, the small subunit S, situated between the AB and the CD dimers, makes contact with large subunits B, C and D. The total area buried in the S-L interfaces covers about 3000 $Å^2$ for each small subunit, with the S-B interface contributing 1800 $Å^2$ and the S-CD interface contributing 1200 $Å^2$ (Knight et al., 1990). The interface areas between small and large subunits show some interesting general features. Although the contact area of the small subunit shows the normal distribution between non-polar, polar and charged atoms (Janin et al., 1988), the corresponding areas from the large subunits are enriched in charged and polar atoms.

1.6.3.1 Interactions between small subunit S and large subunit B

The small subunit S packs against the bottom of the α/β-barrel of the large subunit B of the AB dimer (Figure 1.10, large subunit B). Residues from S that are involved in the contact area are mainly from the N-terminal arm and the hairpin loop between strands βA and βB, but also from helix αA and strand βD. These parts of the small subunit make contact with residues in the C-terminal domain of the B subunit, mainly located in helices αE, α2, and α8 as well as in loop regions on the N-terminal side of the α/β-barrel. Helix α8 of the α/β-barrel interacts extensively throughout its whole length with the N-terminal arm of the small subunit.

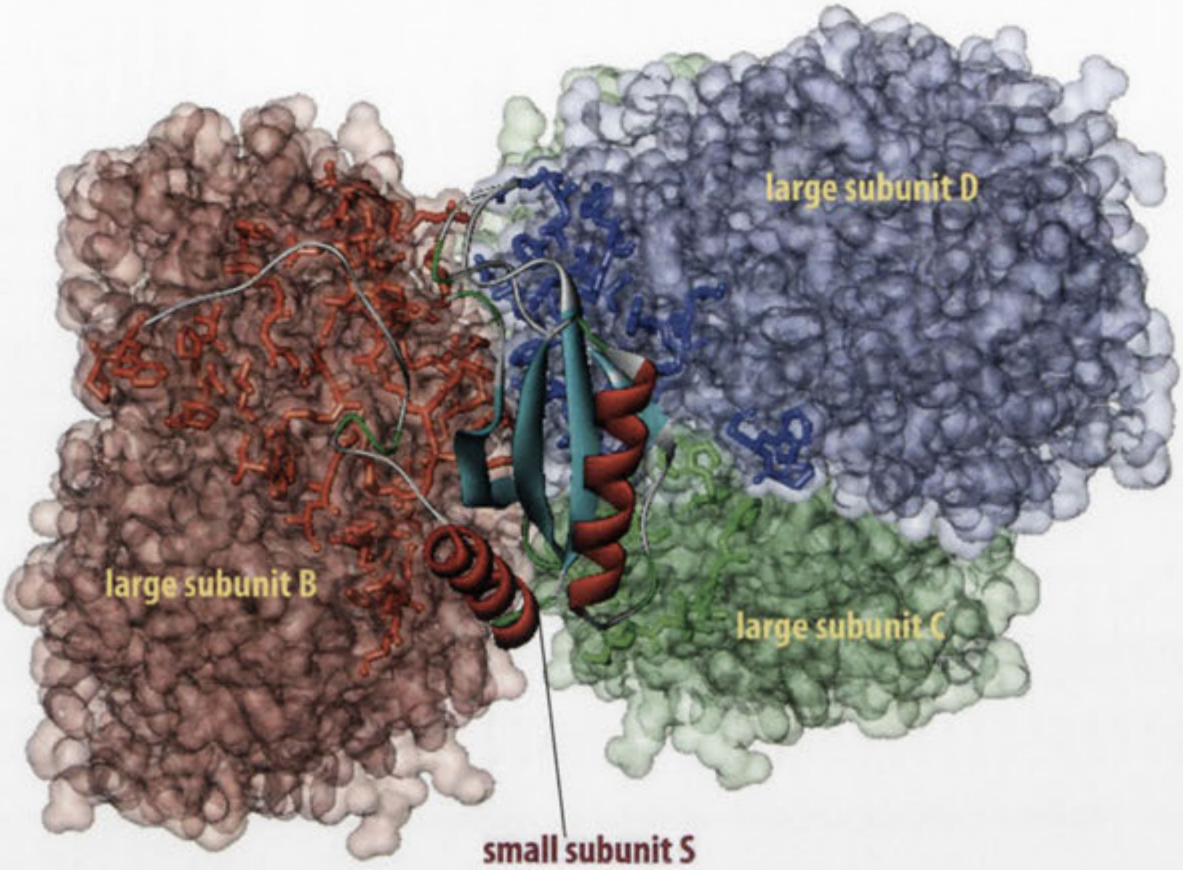


Figure 1.10 Interaction interfaces between Rubsico-LSU and Rubisco-SSU. Each small subunit interacts with 3 different large subunits in the L_8S_8 molecule with the S-B interface contributing 1800Å^2 and the S-CD interface contributing 1200Å^2. Residues at small subunit interaction interface of the Rubisco-LSU are shown in stick representation.

1.6.3.2 Interactions between small subunit S and large subunit dimer CD

While the small subunit S packs against the bottom of the α/β-barrel in the B subunit, the interactions with the large subunit D involve mostly residues from helices α1, α2 and α3 from one side of the barrel (Figure 1.10, large subunit D). There are also interactions with loop 8 at the C-terminal end of the α/β-barrel in the D subunit as well as with residues from the loop between αB and βC in the N-terminal domain of the C subunit. Most of the residues from the small subunit that make contacts to the CD dimer are within the hairpin loop, strand βB and the loop between βC and βD.

1.7 Wealth of Rubisco structure and sequences

More than 95000 *rbcL* and 3000 *rbcS* sequences are presently available from public databases, generated primarily for phylogenetic reconstructions of photosynthetic lineages (Table 3). Numerous crystal structures of Rubisco from diverse origins, including site-directed mutants, have been determined and form a basis for attempts to understand functional aspects. More than 50 Rubisco X-ray crystal structures now exist within the Protein Data Bank (Berman et al., 2000). These range from the homodimeric holoenzyme of *Rhodospirillum rubrum* (Schneider et al., 1986) that provided the baseline information to resolve the high-resolution structures of spinach Rubisco with bound substrate, product, and transition-state analogs to Rubisco-like protein from the green sulfur bacterium *Chlorobium tepidum*. Despite the wealth of information that exists about Rubisco's sequence and structure, systematic analysis of Rubisco's sequences in conjunction with structure to understand functional details is still in initial stages.

Table 3 Rubisco large subunit sequence data in public databases as of June, 2012

Species	No. of Sequences
Green plants	71660
Red algae	5941
Green algae	3723
Brown algae	2481
Diatoms	1078
Yellow green algae	312
Other Eukaryotes	3139
Bacteria	4839
Archaea	160

1.8 Assessment and aims

There is a vast repertoire of information available regarding the structure of Rubisco, its catalytic mechanism, and about the interactions with its substrates. On the other hand, little is known about how sequence variation and specific residues contribute to kinetic profile of Rubiscos from varied sources, enzyme assembly and enzyme activation.

Rubisco is a major limitation to photosynthetic CO_2 assimilation in C_3 plants. Improvements in Rubisco's kinetic properties in C_3 plants can be achieved by identifying and introducing a Rubisco with either higher catalytic rate or that with better specificity for CO_2, or both. There is a significant natural variation in Rubisco's kinetic properties (as shown in Table 2) among Rubiscos from diverse sources that could be exploited for engineering a better Rubisco. However, a number of major technical hurdles must still be overcome before these approaches can be utilized for Rubisco improvement in C_3 plants. For example, the attractive kinetic properties of Rubiscos from non-green algae cannot be exploited in higher plants because Rubiscos from Red algae do not assemble in higher plants (Whitney et al., 2001). Rubisco activation by Rubisco activase adds another dimension to the Rubisco puzzle, and current knowledge raises more questions than answers. Mutagenesis studies on Rubisco have had little success in identifying sequence changes accountable for distinct kinetic profiles between land plants, green algae, non-green algae, cyanobacteria or even among groups of land plants. These hurdles in re-engineering Rubisco could be due to coevolutionary constraints on Rubisco evolution because of its complex structure, and its dependence on other proteins for assembly and activation. In the case of mutations, it is an absolute must that complementarity of interactions (for instance residue charge or size complementarity between two or more sites) is maintained both within the Rubisco holoenzyme and between its interactions with other proteins. A sound understanding of complex coevolutionary processes is essential for fine tuning Rubisco's performance.

Another interesting aspect of Rubisco's evolution is codon usage bias in the *rbcL* gene, which has very high A+T content, and has been shown to be affected by genome compositional biases and genetic drift. Evidence of weak selection in codon usage bias of *rbcL* has also been observed (Wall and Herbeck, 2003). There is a range of structural, biochemical, biophysical, and computational evidence that support the critical role of synonymous codons within the context of protein structure/function (Gu et al., 2003, Kimchi-Sarfaty et al., 2007, Komar, 2007, Zhou et al., 2009). Previous studies on codon usage bias of *rbcL* dealt with the questions of mutational dynamics, drift, and selection on the evolution of codon choice in *rbcL* (Albert et al., 1994, Morton, 1994, Morton and

Levin, 1997) but few studies has been conducted to decipher the role of synonymous codons within the context of Rubisco's structure/function.

Thus, there is considerable scope to explore these aspects in Rubisco, which makes Rubisco a highly attractive target for computational studies. The systematic analysis of natural variation in Rubisco-LSU sequences both at the protein and nucleotide level and in reference to structural data including enzymes responsible for its assembly such as RbcX, and activation i.e. Rubisco Activase, could provide new leads for rational re-engineering of Rubisco in plants. Acquisition of new knowledge about these mechanisms could shed light towards engineering a "better" Rubisco, which has a huge potential to impact crop productivity.

The specific project aims of this thesis can be summarized as:

- To accumulate, integrate and annotate information on Rubisco, to provide a high-quality dataset for studies of its structure, function and evolution.
- To study the coevolution both within the Rubisco holoenzyme, and between Rubisco subunits and its interacting partners, to gain a better understanding of fine tuning of Rubisco's function at molecular level.
- To study codon preferences of *rbcL* in order to understand the role of synonymous codons within the context of Rubisco's structure/function.

2 Database of Rubisco sequences

2.1 Background

The large subunit of Ribulose-1, 5-biphosphate carboxylase/oxygenase (Rubisco, EC 4.1.1.39) is arguably the most sequenced gene with > 95,000 sequences available in public databases from all three forms of life, eukaryotes (plants and algae), archaea and prokaryotes (autotrophic bacteria). As described in Chapter 1, Rubisco is central to photosynthesis and has been studied intensively. Furthermore, the gene *rbcL* that encodes Rubisco large subunit has been widely used as a marker gene for phylogenetic analysis. For these reasons, the large subunit of Rubisco (*rbcL*) gene has been sequenced extensively.

The literature on structural and functional information of Rubisco is also substantial. A PubMed search with a Rubisco-specific text query returns more than 3500 matches, and there are 52 Protein Data Bank entries containing experimentally determined Rubisco structures. As my objectives were to perform a range of studies on Rubisco sequences, such as comparative sequence analysis and sequence-structure-function relationships analysis, a systematic analysis of the complete, non-redundant and annotated collections of Rubisco sequences, both at protein and nucleotide level was required. A customized local repository of Rubisco sequences/structure / kinetic data was required for retrieving, storing, annotating and accessing this information. To meet this need, I created a relational database for storing the Rubisco sequences/structure / kinetic data and a set of Python modules for accessing and retrieving this data collection.

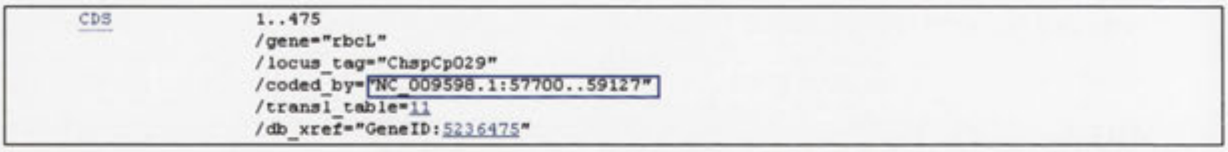
2.2 Requirements

The local Rubisco database has wide ranging requirements. First and foremost I needed a local repository to store Rubisco protein/nucleotide sequence. In addition there was a need to find a way to integrate other information available about Rubisco such as taxonomy, kinetic data and data mapping Rubisco structure to sequence. I also needed to remove redundancy from Rubisco sequences sourced from public databases to create unique sequence datasets for further analyses.

2.2.1 A local repository of Rubisco protein/nucleotide sequences

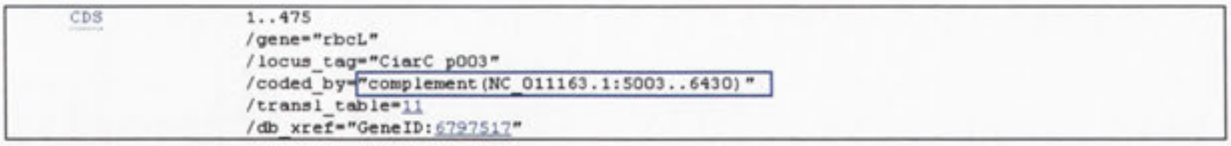
The basic task was to assemble and maintain a non-redundant local collection of Rubisco-LSU protein/nucleotide sequences. As noted in Background, non-redundant collections of Rubisco sequences selected by various criteria were required to perform a range of planned studies. Hence, I needed a system that makes it easy to extract a sequence dataset based on ad-hoc decision making.

The sequences were primarily sourced from National Center for Biotechnology Information Protein/Nucleotide databases (http://www.ncbi.nlm.nih.gov/). My initial attempts to download coding sequences (CDs) for Rubisco-LSU protein sequences made it clear that retrieval of *rbcL* coding sequence corresponding to a given Rubisco-LSU is not trivial. Many of the Rubisco-LSU GenPept records were derived from sequence coordinates from the corresponding chloroplast genome record. Some of these coding sequences could be obtained by extracting the nucleotide sequence from its chloroplast genome such as *rbcL* sequence for *Chloranthus spicatus* (chloroplast genome id NC_009598.1) was extracted from sequence coordinates "57700 to 59127" as shown in Figure 2.1. Then again in chickpea, the sequence between genomic coordinates "5003 to 6430" within it chloroplast genome (NC_011163.1) was extracted and then reverse complemented to get the right coding sequence for *rbcL* gene (Figure 2.2). For ease of the data collection, these tasks needed to be automated.



```
     CDS             1..475
                     /gene="rbcL"
                     /locus_tag="ChspCp029"
                     /coded_by="NC_009598.1:57700..59127"
                     /transl_table=11
                     /db_xref="GeneID:5236475"
```

Figure 2.1 Part of GenPept record YP_001294106 i.e. R-LSU for *Chloranthus spicatus*. It shows that the record was derived from translation of the nucleotide sequence extracted from its chloroplast genome based on location coordinates (highlighted in the figure).



```
     CDS              1..475
                      /gene="rbcL"
                      /locus_tag="CiarC_p003"
                      /coded_by="complement(NC_011163.1:5003..6430)"
                      /transl_table=11
                      /db_xref="GeneID:6797517"
```

Figure 2.2 Part of GenPept record YP_002149717 i.e. R-LSU for *Cicer arietinum* (chickpea). It shows that the protein sequence was derived from translation of the complement of the nucleotide sequence extracted from its chloroplast genome based on location coordinates (highlighted in the figure).

An accurate nucleotide-to-protein sequence correspondence was essential for planned codon-usage studies on Rubisco, as one of my foremost questions was how codon-usage bias of *rbcL* is linked to the Rubisco-LSU's structural/ biochemical/biophysical properties. To analyze the relationship of codon preferences of *rbcL* with Rubisco-LSU's physicochemical properties i.e. secondary structure, solvent accessibility, evolutionary conservation and structural constraints, one-to-one mapping between a given codon in the *rbcL* coding sequence with the corresponding residue is essential for each position in a given Rubisco-LSU protein sequence (As illustrated in Figure 2.3).

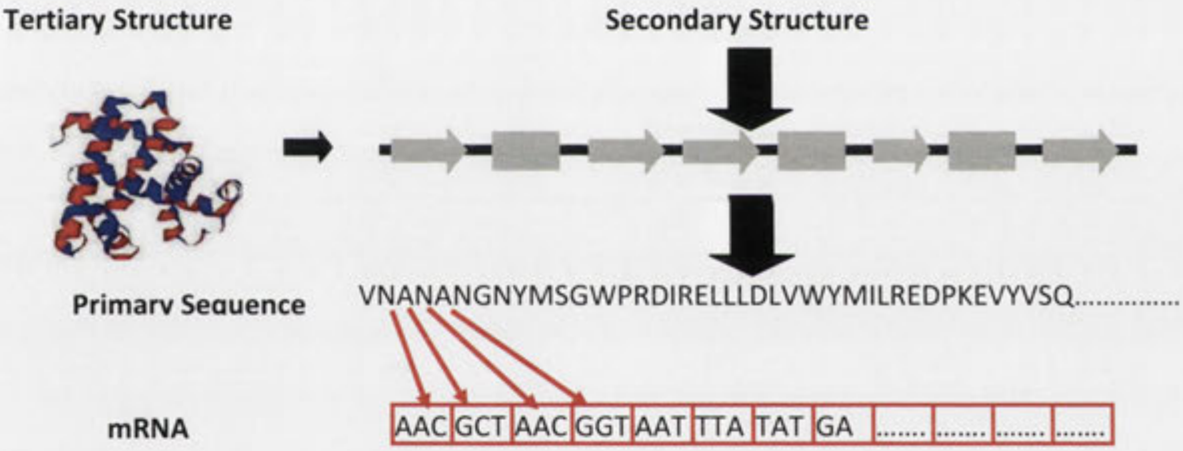

Tertiary Structure
Secondary Structure
Primary Sequence
VNANANGNYMSGWPRDIRELLLDLVWYMILREDPKEVYVSQ...............
mRNA
AAC GCT AAC GGT AAT TTA TAT GA ....

Figure 2.3 One-to-one mapping between each amino acid in a protein sequence with its corresponding codons in the mRNA sequence. Study of the relationship between synonymous codon-usage and protein structure requires a precise mapping between codons in the mRNA and amino acids in the solved protein secondary and tertiary structure. One-to-one mapping between each amino acid in a protein sequence with its corresponding codons in the mRNA sequence is highlighted in red (Figure 2.3 adapted from Deanne and Saunders (2011)).

2.2.2 Data integration

In order to answer complex biological questions, the data from two or more biological databases needs to be accessed from within one computational framework. Making this technically feasible is called data integration. Data integration becomes more difficult the more different types of data have to be included. The necessary effort grows closer to exponentially than linearly with the different types of source data. The reasons for this are of semantical and technical nature (Stein, 2002). In the context of my study, integration of taxonomic, kinetic and structural data with sequence data was necessary to facilitate the planned complex comparative sequence analyses and sequence-structure-function relationships analyses.

2.2.2.1 Indexing the sequence collection by taxonomy

The next major task was to index the sequence collection by taxonomy. I intended to sample sequence space for the Rubisco-LSU within a given taxon in order to identify specific sequence signatures in the Rubisco-LSU for specific taxa. For example, some unique sequence signatures have been identified in the *Solanaceae* (Portis, 1995) and *Poaceae* (Terachi et al., 1987) that are linked to functional/kinetic properties of the Rubisco. For this analysis, complete taxonomic lineages of the Rubisco-LSU sequences were required; linking taxonomy with the sequence collection facilitates sequence analysis based on selected taxonomic Rank.

2.2.2.2 Kinetic and structural data

To conduct the sequence-structure-function relationship studies it was necessary to add functional information to Rubisco sequences in the form of kinetic and structural data. There is a vast amount of kinetic data for different native and mutant Rubiscos available in the literature and one of the major objectives of my study was to link such data with specific sequence positions. As noted in Background, there are 52 Rubisco structures available in the PDB (Berman et al., 2000) as-of-now; analysis of structural data in the context of sequence and kinetic data forms the basis of the planned sequence-structure-function relationship analysis.

2.2.3 Curation

The main goal of curation was to provide accurate and full-length non-redundant sequence data of Rubisco sequences. The NCBI Protein/Nucleotide databases contain a high level of redundancy. The user has to find a way to remove redundant entries from datasets retrieved from these databases. This became increasingly difficult with the volume of the data that needed to be processed for my study. For example, rice (*Oryza sativa*) has more than 5 Rubisco-LSU sequences in the NCBI Protein database with sequence length ranging from 234 to 477 residues. While preparing the sequence dataset for analysis, I wanted to include only one sequence from rice with the full-length sequence data.

2.3 Existing solutions

There are three major public databases that provide information related to nucleotide/protein sequence, summarized in Table 2.1. Among biologists, GenBank (Benson et al., 2011) is probably the most popular database from which to retrieve sequences (and other data); it is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI). GenBank participates with the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) (Kulikova et al., 2007) and the DNA Databank of Japan (DDBJ) (Sugawara et al., 2008) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC); this exchanges data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide.

Table 2.1 Major public databases for biological information

Name of the public database	Web link
National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/
European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL)	http://www.embl.org/
DNA Databank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp/

2.3.1 NCBI GenBank/Nucleotide database

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS) and other high-throughput data from sequencing centers. Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references and a table of features listing areas of biological significance, such as coding regions and their protein translations, transcription units, repeat regions and sites of mutations or modifications.

2.3.2 NCBI Protein/GenPept database

The protein sequences in the NCBI Protein database come from several different sources. There are GenPept translations for each of the coding sequences within the GenBank Nucleotide database. This means that there can be more than one protein sequence associated with a corresponding Nucleotide sequence record.

2.3.3 NCBI Entrez query interface

Entrez developed by Schuler et al. (1996), is an extremely useful tool for data retrieval from NCBI as it integrates data from 35 diverse databases including Genbank and GenPept. Entrez also periodically incorporates new databases as and when they are introduced to the public domain. Text searching in Entrez is supported by typing simple Boolean queries in the search box and data can be downloaded either singly or in batches in multiple formats such as fasta or XML. The accessed records are in turn linked between the various databases for ease of cross-referencing, for instance, a protein sequence is cross-referenced with its coding sequence, its 3D structure and the reference where the sequence was first reported (Sayers et al., 2012).

2.3.4 NCBI taxonomy database

The biological databases in Entrez are organized on the basis of NCBI taxonomy database which provides links to data for each taxonomic node right from super-kingdoms to subspecies (Federhen, 2012). The sequences in the NCBI databases are classified with the assistance of external advisers and curators that can then be conveniently queried using the taxonomy browser (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome

.html/). This browser can be used to view the taxonomic position of a particular organism or group as well as retrieve the data from any of the Entrez databases.

2.4 Limitations of public databases

Although public databases are a rich source of biological information that provide ready access to biological data with a large collection of interactive tools and intuitive web interfaces, they still struggle to cope with the wide-ranging requirements of integration, customization and redundancy of biological information.

2.4.1 Difficulty in processing complex queries

In a project such as mine, where complex queries on several databases are the routine rather than the exception, manual tracking of the hyperlinks for desired features across web pages and databases is tedious. In most cases, I would end up mining the content in a clickathon of cut-and-paste and screen-scraping. For the large datasets required for my studies on Rubisco, this procedure is absolutely unsuitable. Also in many cases, users like to perform complex queries that exceed web servers' capabilities.

2.4.2 Restrictive form-based query interfaces

Most of the public repositories of biological data have form-based query interfaces that can be restrictive and usually have very little scope for customization of a query. To overcome this problem they provide an "advanced" form that provides a user with more choices than the "basic" form. For instance, a keyword search for Rubisco at NCBI returns over 95000 entries and if one wants to restrict the search by sequence length, it can be done by their advanced search form. However, if the user wants to exclude sequences which have missing residues (represented by character "X" in a sequence) within the sequence, there is no way of doing it. Therefore, the above drawback remains in spite of the additional form complexity. Another drawback of form-based query interface is that they provide users with few options to customize the format and content of the query's result. To illustrate, consider the Entrez search engine of NCBI. It typically displays result at extremes: either by displaying a result overview that contains multiple results per page but limited information about each, or by displaying one result per page (showing all the available information about that result) with links to the other results.

2.4.3 Redundancy in public databases

Moreover most public databases contain a high level of redundancy. The user is confronted with the problem of sorting redundant entries out of datasets he/she is retrieving from a database, and the difficulty of this certainly rises rapidly with the volume and complexity of the data.

2.4.4 A local data repository solves most of the problems

To overcome these problems, biological data in public databases need to be stored locally and accessed via programming libraries. Most of the public databases release their data in a structured, machine-readable format. To use this data with minimal effort, various programming libraries have been developed. They contain parsers, interfaces to web databases and bioinformatical algorithms, thus enabling a user to do almost anything with the data. The Bio* projects (Stajich, 2007, Cock et al., 2009, Mangalam, 2002, Holland et al., 2008, Goto et al., 2010) are a well-known example for this category.

Hence, although public resources with their intuitive web interfaces are easy to use and can be very useful for users interested in single entries or small subsets of a database, their flexibility is too limited for use in studies such as mine.

2.5 Implementation

My implementation of the Rubisco database includes a relational database backend and a Python application programming interface (API) that allows accessing sequence and annotation objects programmatically. The relational database backend has two databases, the sequence database and the annotation database (Figure 2.4). The schema of the sequence database is based on the BioSQL project (http://www.biosql.org/wiki/Main_Page) as shown in Figure 2.5. The software components of the system are written in Python by using modules from the Biopython project (Cock et al., 2009).

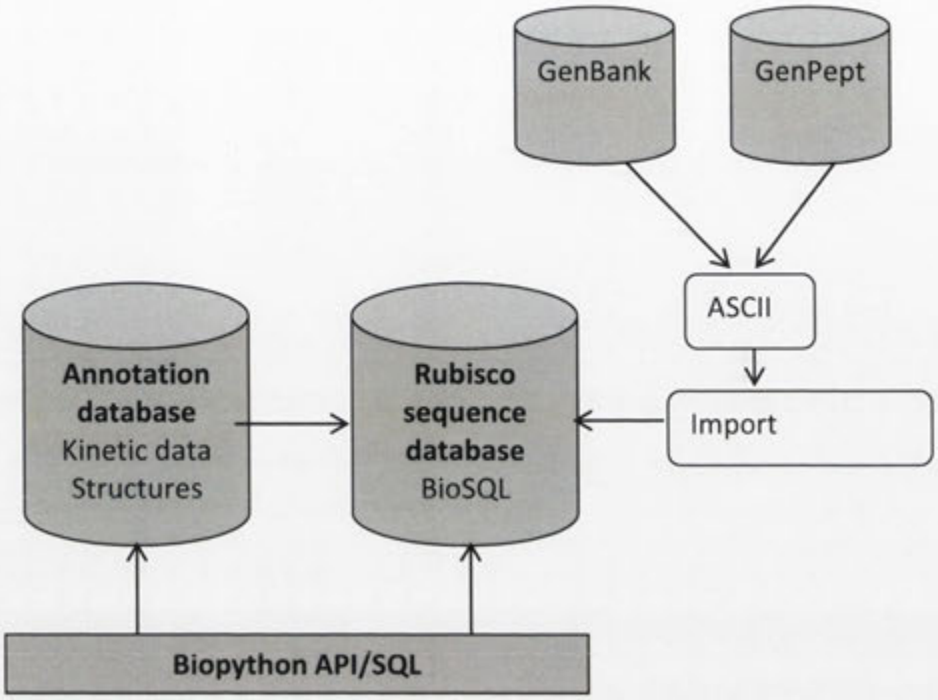


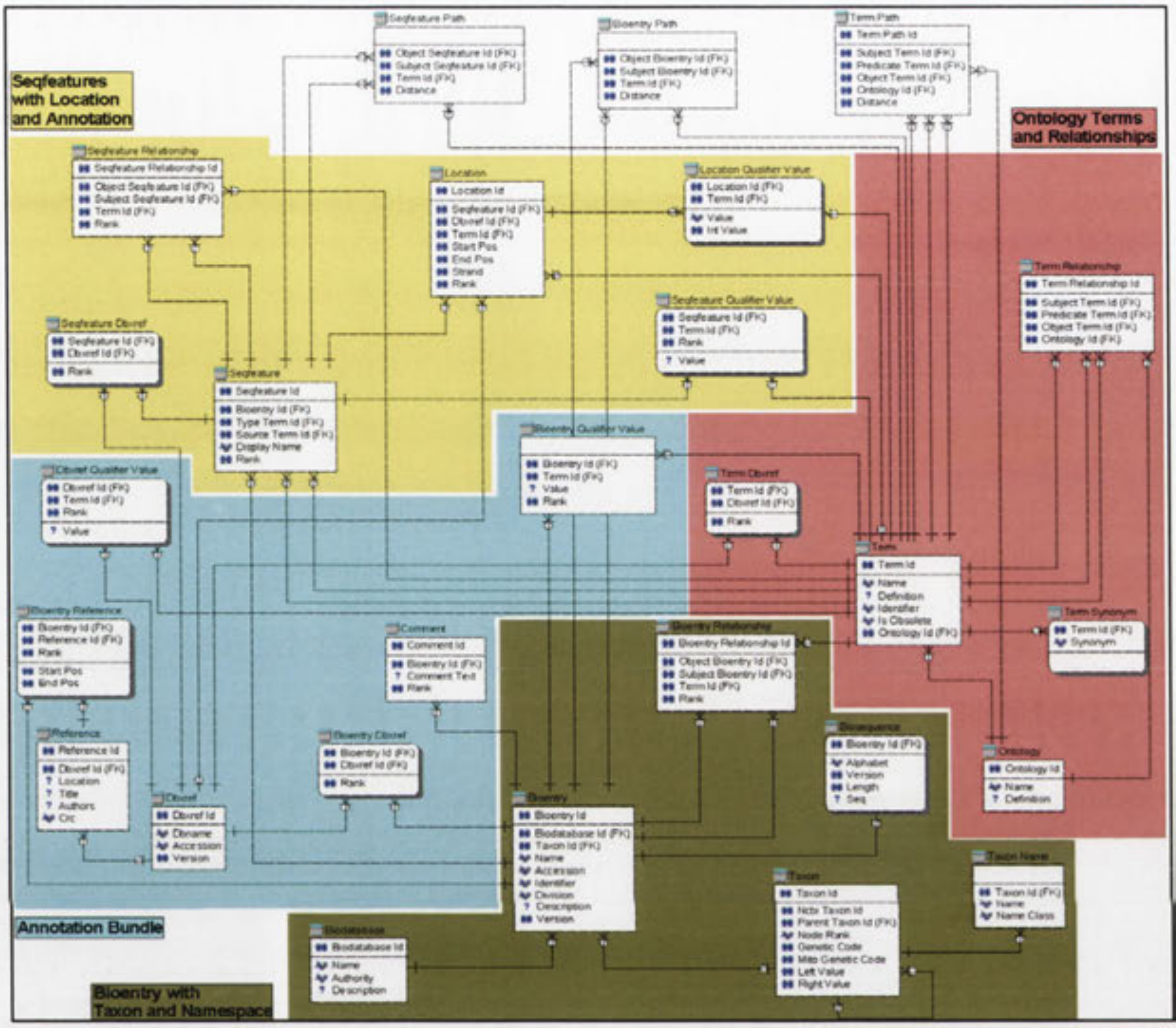
Figure 2.4 Organization of the Rubisco sequence database system. The relational database backend consists of two databases, the sequence database and the annotation database. The sequence database is based on the BioSQL schema and serves two purposes. First, it locally stores the sequence information imported from the public databases. Second, the results of local annotation that can logically be associated with a location within the sequence (rather than being a property of the sequence as a whole or a group of sequences) are stored in the sequence database as sequence features. The sequence database is accessed via the Biopython modules.

2.5.1 Sequence database

I have used the generic relational BioSQL model (http://www.biosql.org/wiki/Main_Page) to support and develop a shared database schema for storing sequence data. BioSQL is a generic relational model covering sequences, features, sequence and feature annotation, a reference taxonomy, and ontologies as shown in Figure 2.5. It was originally conceived by Ewan Birney in 2001 as a local relational store for GenBank. The project has since become a collaboration between the BioPerl, BioPython, BioJava, and BioRuby projects. Its schema (see Figure 2.4) allows for continuous non-transient storage of sequences, features, and annotation in a way that is interoperable between the Bio* projects. Each Bio* project has a language binding (object-relational mapping, ORM) to BioSQL.

I have used MySQL as a supported Relational Database Management System (RDBMS), together with the associated python library. GenBank/GenPept files were used to supply and maintain the information necessary for the database. The sequences, features, and annotations were introduced into the database using modules of the Biopython project (Cock et al., 2009).

The database presented here consists of more than 11,000 unique Rubisco-LSU protein /*rbcL* nucleotide sequence entries from Angiosperms. Rubisco-LSU sequences in the database belong to 47 orders and 396 families, providing exhaustive coverage of the most taxon-rich lineage of phototrophs (80% of flowering plant orders and 96% of families' *sensu* Angiosperm Phylogeny Group III; Bremer et al. (2009)). The taxonomy tables were downloaded from NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/) and introduced into the database using a perl script provided with the BioSQL package. Rubisco sequences were retrieved from NCBI Protein/Nucleotide databases by using Perl scripts and semi-automatically curated to remove redundancy before introducing to the database.



Source: http://www.biosql.org/wiki/Main_Page)

Figure 2.5 BioSQL 1.0 Entity relationship diagram

2.5.2 Python API

The implementation and data acquisition functions of the Rubisco database are based on the Biopython project and, therefore, allow for the use of routines available in the public domain. Five packages from the Biopython project i.e. BioSQL, Bio.Seq, Bio.SeqRecord, Bio.SeqIO and Bio.Entrez have been utilized in this implementation.

2.5.2.1 BioSQL

The BioSQL package contains three sub modules:

1. **BioSeqDatabase:** This provides interfaces for loading biological objects from a relational database, and is compatible with the BioSQL standards. Its basic task is to connect with a BioSQL database and load Biopython-like objects from it.

2. **BioSeq:** This allows retrieval of items stored in a BioSQL database using a Biopython-like SeqRecord and Seq interface.
3. **Loader:** This loads Biopython objects into a BioSQL database for continuous storage. Loader makes it possible to store Biopython objects in a relational database and then retrieve them.

2.5.2.2 Bio.Seq

Bio.Seq provides basic methods to manipulate proteins, DNA and RNA sequences, but additionally provides the ability to extend and customize the sequence manipulation requirements.

2.5.2.3 Bio.SeqRecord

The SeqRecord (Sequence Record) class allows higher-level features such as identifiers and features to be associated with a sequence; this is the basic data type for the Bio.SeqIO sequence input/output interface.

2.5.2.4 Bio.SeqIO

Bio.SeqIO provides a simple uniform interface to input and output assorted sequence file formats. The workhorse function Bio.SeqIO.parse() is used to read in sequence data as SeqRecord objects. This function expects two arguments: 1. a handle to read the data, which can be a filename or data downloaded from the internet, and 2. sequence format.

2.5.2.5 Bio.Entrez

The Bio.Entrez module makes use of the Entrez Programming Utilities (also known as EUtils), consisting of eight tools that are described in detail on NCBI's page at http://www.ncbi.nlm.nih.gov/entrez/utils/. Each of these tools corresponds to one Python function in the Bio.Entrez module (see Table 2.2). This module ensures that the correct URL is used for the queries, and that not more than one request is made every three seconds, as required by NCBI.

Table 2.2 Entrez EUtils functions

Function	Description
efetch	Retrieves records in the requested format from a list of one or more primary IDs or from the user's environment.
epost	Posts a file containing a list of primary IDs for future use in the user's environment to use with subsequent search strategies.
esearch	Searches and retrieves primary IDs (for use in EFetch, ELink, and ESummary) and term translations and optionally retains results for future use in the user's environment.
elink	Checks for the existence of an external or Related Articles link from a list of one or more primary IDs. Retrieves primary IDs and relevancy scores for links to Entrez databases or Related Articles; creates a hyperlink to the primary LinkOut provider for a specific ID and database, or lists LinkOut URLs and Attributes for multiple IDs.
einfo	Provides field index term counts, last update, and available links for each database.
esummary	Retrieves document summaries from a list of primary IDs or from the user's environment.
egquery	Provides Entrez database counts in XML for a single search using Global Query.
espell	Retrieves spelling suggestions.
read	Parses the XML results returned by any of the above functions.

2.5.3 Annotation database

The annotation database is based on custom schema. The annotation database has two tables: One for storing Rubisco kinetic data and another for storing information about available Rubisco PDB structures.

2.5.3.1 Kinetic data

The kinetic data table contains the manually compiled information based on the available literature on the kinetic properties of Rubisco. Each record contains the data from only one organism. Only the reported values have been recorded - no attempt has been made to calculate missing values. The kinetic properties of each Rubisco have been condensed to one row. Where different values of the same kinetic property have been reported in the literature, the range of reported values is listed. If the original paper could not be located, the reference in which the original data was cited has been given as the source reference. The current table includes kinetic values from 40 species, including 11 species from flowering plants. As shown in Table 2.3, kinetic data table stores form of Rubisco, taxonomic rank, name of the organism, kinetic data and the reference that published/cited original data.

Table 2.3 Fields in kinetic data table in annotation database

Fields	Description
Form	Form of Rubisco characterized (IA/IB/IC/ID/II/III/IV)
Org class	Taxonomic class of organism
Name	Organism name (according to current NCBI taxonomy database)
Specificity	Ratio of carboxylation to oxygenation
Binding Constant for CO_2	Michaelis-Menten constant for carboxylation
Carboxylation rate	Rate of carboxylation
Catalytic rate	Catalytic turnover rate for carboxylation per site per second
Ref	The reference in which the original data was published/cited.

2.5.3.2 Structure data

In the structure table, only the description from the PDB header, but not the structure itself is included. For each entry the following information is retrieved by using the Biopython Bio.PDB.Header.parser method. As shown in Table 2.4, structure data table hold information on method of structure determination, resolution of the structure, name of source organism, deposition and release date and the reference in which the original data was published.

Table 2.4 Fields in structure data table in annotation database

Fields	Description
Structure_method	Method of structure determination i.e. X-ray diffraction/ NMR/Electron microscopy
Head	Classification of enzyme
Journal	The reference in which the original data was published/cited.
Journal_reference	More details about the reference
Compound	Chemical name of the molecule and chain details
Keywords	Keywords to search the structure in a database
Name	Common name of source organism
Author	Author of the structure
Deposition_date	Date of deposition
Release_date	Date of release in PDB
Source	Source organism
Resolution	Resolution of the structure
Structure_reference	Other references related to the structure

Rubisco PDB structures have been downloaded and dumped locally to facilitate structural analyses. Currently it holds 31 PDB structures including spinach, tobacco and rice from flowering plants.

2.6 Extracting and inserting data into Rubisco database

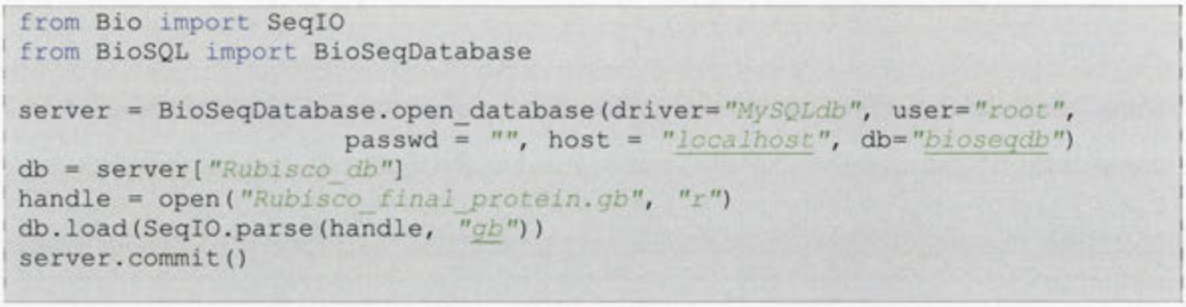
2.6.1 Insertion of data

As mentioned before, primary sequence data for Rubisco-LSU protein and *rbcL* nucleotide sequences are obtained from primary databases such as GenBank/GenPept using Perl/Python scripts and semi-automatically curated. As noted in section 2.2.1, in some cases downloading the coding sequences of Rubisco-LSU was not straight-forward. To solve this problem, scripts to download coding sequences of a given GenPept sequences were developed. The sequences, features and annotations are inserted into the database using Python scripts. All the scripts used in this chapter are available in the RUBISCO_DB directory of the supplementary compact disk. For instance, the following bit of code (Example 1) was used to download protein sequences.

```
from Bio import Entrez
Entrez.email = "animesh.agrawal@anu.edu.au"

fh=open("Flowering_plant_id_list.txt", "r")
myfile =fh.read()

temp_id_list=myfile.split('\n')
temp_id_list = temp_id_list[0:-1]

search_results = Entrez.read(Entrez.epost("protein",
id=",".join(temp_id_list)))
webenv = search_results["WebEnv"]
query_key = search_results["QueryKey"]

count = 11452
batch_size = 500
out_handle = open("Rubisco_final_protein.gb", "w")

for start in range(0,count,batch_size):
    end = min(count, start+batch_size)
    print "Going to download record %i to %i" % (start+1, end)
    fetch_handle = Entrez.efetch(db="protein", rettype="gb",
retmode="text",
                                  retstart=start, retmax=batch_size,
                                  webenv=webenv, query_key=query_key)
    data = fetch_handle.read()
    fetch_handle.close()
    out_handle.write(data)
out_handle.close()
```

Example 1 - Shows how to download a large sequence dataset (containing 11452 R-LSU sequences) given a list of unique ids (gi/accession) from NCBI. Here the sequence ids are first posted to NCBI, then sequences are downloaded in batches of 500 sequences using the NCBI search history. Code adapted from Biopython cookbook.

Likewise, the following bit of code (Example 2) was used to insert sequences into the database.



```
from Bio import SeqIO
from BioSQL import BioSeqDatabase

server = BioSeqDatabase.open_database(driver="MySQLdb", user="root",
                                   passwd = "", host = "localhost", db="bioseqdb")
db = server["Rubisco_db"]
handle = open("Rubisco_final_protein.gb", "r")
db.load(SeqIO.parse(handle, "gb"))
server.commit()
```

Example 2 - Shows how to insert sequences into sequence database. Here sequences downloaded in the previous example are inserted into the database by opening a connection to the database and then parsing the downloaded sequence file using the SeqIO module. Code adapted from Biopython cookbook.

2.6.2 Extraction of data

In the Rubisco database, datasets can be generated and made available starting from the content of the local database. There are many ways information can be processed. The following illustrates some of these using cases of SQL queries for the datasets created in the course of my studies.

2.6.2.1 Dataset of Rubisco-LSU protein/*rbcL* nucleotide sequences based on threshold length

Both the coevolution and codon-usage studies of Rubisco required a non-redundant dataset selected by sequence-length criteria. Many sequences were incomplete, lacking residues at the N-terminal and/or C-terminal end; to create final dataset sequences < 450 residues in length these were excluded from analysis. Likewise many *rbcL* sequences lacked bases at the 5' and/or 3' end; sequences < 450 codons (1350 bases) in length were excluded from analysis. For example, the *rbcL* nucleotide sequence dataset with sequences >1350 bases can be created by the following SQL query (Example 3).

```
SELECT biosequence.bioentry_id, taxon_name.name, biosequence.length, biosequence.alphabet,
biosequence.seq
FROM biosequence JOIN bioentry USING (bioentry_id) JOIN taxon USING (taxon_id) JOIN
taxon_name USING (taxon_id)
WHERE taxon_name.taxon_id = bioentry.taxon_id
AND taxon_name.name_class = 'scientific name'
AND biosequence.alphabet='dna'
AND biosequence.length > '1350'
```

Example 3 - Shows an instance of SQL query to select sequences with sequence-length criteria.

Similarly to create a Rubisco-LSU protein sequence dataset with sequences >450 residues, all that is required is to change the field “biosequence.alphabet='dna'” to “biosequence .alphabet ='protein'” and “biosequence.length > '1350'” to “biosequence.length > '450'”.

2.6.2.2 Dataset of Rubisco-LSU protein/*rbcL* nucleotide sequences belonging to a particular taxon.

During the coevolution studies on Rubisco-LSU the coevolution analysis of four orders, i.e. *Solanales, Gentianales, Poales* and *Caryophyllales* was performed. These orders were chosen due to the presence of unique sequence signatures in Rubisco-LSU protein as found in the literature or observed in the course of this study. To conduct these studies I needed to create datasets which belonged to particular taxa. To create the Rubisco-LSU dataset for *Solanales* the following SQL query (Example 4) was executed.

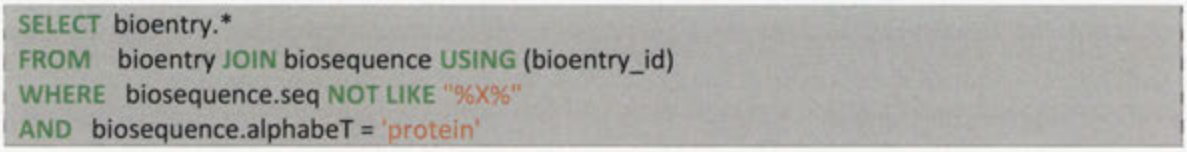
```
SELECT biosequence.bioentry_id, taxon_name.name, biosequence.length, biosequence.alphabet,
biosequence.seq
FROM biosequence JOIN bioentry USING (bioentry_id) JOIN taxon USING (taxon_id) JOIN
taxon_name USING (taxon_id)
WHERE taxon_name.taxon_id = bioentry.taxon_id
AND taxon_name.name_class = 'scientific name'
AND taxon.left_value > (SELECT taxon.left_value FROM taxon JOIN  taxon_name USING (taxon_id)
WHERE taxon_name.name = 'Solanales')
AND  taxon.right_value < (SELECT taxon.right_value FROM taxon JOIN taxon_name USING
(taxon_id)
WHERE taxon_name.name = 'Solanales')
AND biosequence.alphabet='protein'
AND biosequence.length > '450'
```

Example 4 - Shows an instance of an SQL query to select sequences belonging to a particular taxon.

Likewise to create the R-LSU protein sequence dataset for another taxon, the field "taxon_name.name = 'Solanales'" has to be changed to the name of desired taxon.

2.6.2.3 To find a pattern in Rubisco-LSU protein sequence

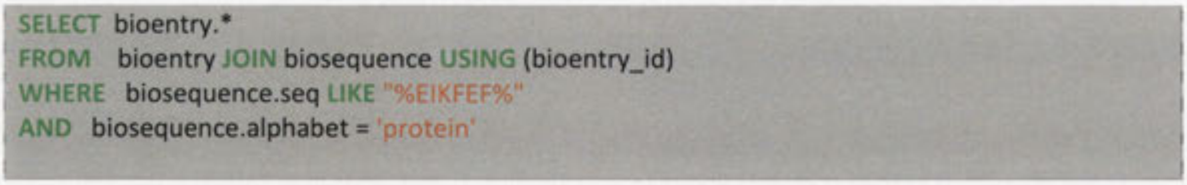
Before using an alignment in a coevolutionary analysis, there can be no ambiguous residue/base codes (e.g., B/Z/X in protein alignments); although some of the algorithms can tolerate them (e.g. Mutual Information), others, which rely on information such as background residue frequencies (e.g. Statistical Coupling Analysis), cannot handle them. The best strategy is to exclude ambiguous codes altogether. The following SQL query (Example 5) can exclude all such sequences.



```
SELECT bioentry.*
FROM   bioentry JOIN biosequence USING (bioentry_id)
WHERE  biosequence.seq NOT LIKE "%X%"
AND  biosequence.alphabeT = 'protein'
```

Example 5 - Shows an instance of SQL query to select all protein sequences which don't contain ambiguous character "X".

Similarly, the following query (Example 6) can search for a sequence signature "EIKFEF" and returns only those sequences which contain the queried pattern.



```
SELECT bioentry.*
FROM   bioentry JOIN biosequence USING (bioentry_id)
WHERE  biosequence.seq LIKE "%EIKFEF%"
AND  biosequence.alphabet = 'protein'
```

Example 6 - Shows an instance of SQL query to select all protein sequences which contains pattern "EIKFEF".

2.7 Conclusions and future development

The Rubisco database has proven to be very useful for my studies. It provides a much more flexible way to access the sequence collection and the annotations and I have employed it extensively for preparing sequence datasets for my studies (Chapter 3 and 4) on Rubisco sequences.

Development of the database is a work in progress. Curation of Rubisco sequences has been automated to an extent, but manual interventions are required frequently. For instance, although sequence length and ambiguous residue/base codes can be checked automatically before inserting the data into the database, redundancy has to be sorted out manually in most of the cases. Likewise compilation of kinetic and structure data for Rubiscos has to be done manually. Also the annotation database has very primitive schema for utilitarian purposes; efforts are ongoing to develop more robust schema for annotation database. Regardless of these issues, the system offers an adaptable interface to retrieve Rubisco sequence datasets and their annotation.

The next step is development of a web user interface for the Rubisco database to provide access to the system to other researchers. The most important advantage of a curated database such as my Rubisco database is that, due to the curatorial effort, the information content is vastly superior to that of public databases. Making the Rubisco database publicly accessible would be useful for the Rubisco research community. I have zeroed in on GBrowse, i.e. Genome Browser or Generic Genome Browser as the tool of choice for the web interface. The Generic Genome Browser developed by Stein et.al (2002), is a web-based application for displaying genomic annotations and other features. It's readily available open source components, simple installation, flexible configuration, and easy integration with BioSQL schema makes it a tool of choice for Rubisco database.

3 Coevolution analysis of Rubisco

3.1 Background

To gain new insights in Rubisco function and structure, I performed a range of computational studies to investigate sequence-structure-function relationships. These studies, aimed to take advantage of the large number of Rubisco-LSU sequences available in public databases, both at protein and nucleotide level. In this chapter, I present the outcomes of coevolution analyses performed on protein sequences of the large subunit of Rubisco in both intra/inter-molecular contexts.

3.1.1 What is coevolution

The original ideas on the mutual influence of species on their evolution were formulated in Darwin's (1862) studies on orchids, where he explored the intricacies of how the petals of a flower guided specific bees or moths for successful pollination. But the term "coevolution" was first used by Ehrlich and Raven (1964) in their studies on reciprocal evolutionary changes between butterflies and plants. Thompson (1994) defined coevolution to describe the correlated evolution of two populations in response to selection imposed by one on the other in a reciprocal manner.

Many examples of coevolution of morphological traits from paired species have been discovered over the last century. Most of these instances can be ascribed to biological interactions such as host- parasite relationships, predator and prey relationships, symbiotic relationships (Moya et al., 2008) and inter-specific competition for resources. It has been observed that at times these interacting species show similar phylogenies, for example the taxonomy of parasites and their hosts (Stone, 1985, Hafner and Nadler, 1988). Although the resemblance of phylogenies indicates analogous evolutionary processes, it cannot be taken as conclusive proof of mutual influence on evolution.

Evolution of a species is the outcome of complex interactions with its environs. Often it is difficult to pinpoint the mechanism of coevolution between a given pair of species; as it involves all the other species in the environs of the species in question, explicit instances of coevolution between individual species cannot be distinguished. This process is referred to as "diffuse coevolution" (Thompson, 1994, Futuyma, 1997). The so called "continual improvement" in the fitness of species is the function of "diffuse coevolution". This phenomenon forms the basis of the famous "Red Queen Hypothesis" proposed by Van Valen (1977), also known as the "evolutionary arms race between competing species."

3.1.2 Molecular basis of coevolution

At molecular level, the term coevolution signifies the evolutionary processes by which a heritable change in the features of one entity exerts selective pressure for a change in another entity. These entities can span many different levels of complexity as long these levels are heritable and under selection from nucleotides to amino acids to proteins (Fares et al., 2011). A case in point is protein-protein interactions, where complementary structural conformations are critical to maintain the interactions between the proteins. In general, these interactions between proteins are mediated through specific set of residues, so that mutations in one of the proteins at interacting sites can disrupt these complementary structural conformations. This may necessitate compensatory mutations at the interacting sites of the other protein to restore the structural complementarity; this process constitutes the coevolutionary dynamics. Although the concept is straightforward to state the reality is not always so simple; coevolutionary dynamics could also be generated among amino acid sites that do not interact due to shared ancestry or to stochastic processes (Fares et al., 2011).

Within a protein, coevolution processes can be accounted for by restating the covarion hypothesis, put forth by Fitch and Markowitz (1970). This postulates that at any time point during the evolution of a protein only a small fraction of possible mutations are admissible, but as one site changes it can alter the selective forces associated with other sites, thus altering the set of mutations that are selectively admissible at those sites.

This form of coevolutionary process could be recognized within a protein as residue pairs and these interactions are called covariation/correlated interactions.

Both within and between proteins, many forms of coevolutionary processes involving compensatory/complementary dynamics can be expected (Fukami-Kobayashi et al., 2002). For example, a "big-for-small" replacement at position *i* might be compensated by a corresponding "small-for-big" replacement at position *j*, to conserve the overall size in the packed core of the fold, and therefore conserve a functional behavior related to packing (the stability of the fold). Alternatively, a "positive-for-negative" charge replacement at position *i* might be compensated by a "negative-for-positive" charge replacement at position *j*, to conserve overall charge and, therefore, conserve a functional behavior related to net charge. In the language of the neutral theory of evolution (Kimura, 1983), we would say that the first replacement was selectively disadvantageous, the second was positively selected (in the context of the first), and both together lead to a result that is at least neutral or may be somewhat better for overall function of the protein.

3.1.3 Model for Covariation/Correlated interactions

Atchley et al. (2000) formalized a simple linear model to explain Covariation /Correlation (C) between two sites in a sequence alignment.

$$C= C_{structure} + C_{function} + C_{phylogeny} + C_{interaction} + C_{stochastic}$$

$C_{phylogeny}$ is correlation due to phylogenetic relationships between homologous sequences that are related by a tree-like evolutionary structure and, therefore, cannot be considered to be statistically independent observations. Thus, we expect that the outcome of compensatory substitutions that occurred in a sequence ancestral to a group of sequences under consideration will be manifest in the descendent sequences and that simple pairwise comparisons between sequences will not be sufficient to provide an accurate account of evolutionary events.

$C_{structure}$ and $C_{function}$ signify correlation due to structural and functional constraints, effectively the signal that covariation analyses attempt to uncover. However, these sources of correlation may not be independent from one another or indeed from phylogenetic correlation. $C_{interaction}$ describes interactions between the aforementioned sources of correlation. Finally, random effects from uneven or incomplete sequence sampling, casual co-variation and other stochastic factors are represented by $C_{stochastic}$.

In reality, it is difficult to distinguish between structural and functional correlations; hence, most methods employed to uncover correlated interactions endeavor to eliminate stochastic and, potentially phylogenetic noise. This is a major challenge; as demonstrated by Noivirt et al. (2005) the strength of correlations due to phylogenetic factors are often of the same order of magnitude as those due to structure and function.

3.1.4 Methods to detect coevolution at residue level

Multiple sequence alignments (MSAs) are extensively exploited to examine correlated interactions in proteins. In MSA of homologous proteins, "corresponding residues" are placed in the same column. However, it can be difficult to define "corresponding residues" without structural comparisons if there are several insertions-deletions in the homologous sequences. For conserved proteins, for a given position in the alignment, MSAs are reasonably accurate representations of the amino-acid substitutions tolerated in the course of evolution. As functional and structural constraints lead to restrictions on these substitutions, MSAs provide a robust framework to study coevolutionary processes in the context of protein structure-function relationships. All the coevolution detection algorithms use MSA as a starting point of the analysis.

Most coevolution algorithms published so far can be broadly classified in two categories: tree-based and tree-ignorant methods (Caporaso et al., 2008). Tree-based methods attempt to control for phylogeny by accounting for explicit phylogenies in the coevolution statistic, whereas tree-independent methods have implicitly assumed a star phylogeny.

3.1.4.1 Tree-independent methods

Tree-independent methods have become very popular over the last decade due to short compute times and the fact that they does not require phylogenies, thus not being subject to model misspecification. Some algorithms that have received significant attention are those: based on observed and expected patterns of data distribution (Larson et al., 2000, Kass and Horovitz, 2002, Noivirt et al., 2005), use a correlation coefficient (Gobel et al., 1994, Olmea and Valencia, 1997, Afonnikov et al., 2001, Vicatos et al., 2005) or the Information theoretic "Mutual Information (MI)" statistic (Martin et al., 2005, Gloor et al., 2005, Dunn et al., 2008), or are based on alignment perturbation i.e. "Statistical coupling analysis (SCA)" (Lockless and Ranganathan, 1999, Suel et al., 2003). Table 3.1 gives a comparison of tree-independent methods, categorized according to their strong and weak points.

Table 3.1 Comparison of tree-independent methods

Method	Short Description	Strong points	Weak points
1. Chi-square	Several studies (Larson et al., 2000, Kass and Horovitz, 2002, Noivirt et al., 2005) have exploited the observed patterns of amino acid occurrence at two positions and compared them to the frequencies of data expected under an independent sites model, generating a chi-squared observed minus expected squared (OMES) statistic.	Low compute time	Coevolution statistic does not consider shared ancestry
2. Correlation	The early correlation coefficient-based methods (Gobel et al., 1994) do not consider substitutions but correlations between physicochemical properties of amino acids found in pairs of sites in individual sequences. Later implementations (Olmea and Valencia, 1997, Afonnikov et al., 2001, Vicatos et al., 2005) included correlations between the weights for substitutions that are inferred to have occurred at each of two sites during the course of evolution.	Low compute time	Coevolution statistic does not consider shared ancestry
3. Mutual Information (MI)	Mutual information (MI) is a measure of the mutual dependence of two variables. In the case of two positions *i* and *j* in a sequence alignment it can be considered a measure of how much information about site *j* in a given sequence is given by knowledge of site *i* $\mathrm{MI}\ (\mathrm{i},\mathrm{j}) = \sum_{x=1}^{n}\sum_{y=1}^{m} p_{xy} \log_2 \frac{p_{xy}}{p_x p_y}$ Where MI(*i*, *j*) is the mutual information between sites *i* and *j*, the indices refer to the 20 possible amino acid states and p_{xy} reflects the probability of observing amino acid *x* at site *i* and amino acid *y* at site *j*, p_x and p_y are the respective independent probabilities of these events (Martin et al., 2005).	Low compute time	Coevolution statistic does not consider shared ancestry

Method	Short Description	Strong points	Weak points
4. Normalized mutual information (NMI)	Normalized mutual information (NMI) is computed by dividing the MI score for each pair of alignment positions by the joint entropy of that pair of positions. This normalization makes an effort to moderate the lessening effect of higher sequence conservation on mutual information. In a given alignment, if two pairs of columns have perfectly correlated substitutions, but one pair is more highly conserved than the other, the more highly conserved pair will have a lower MI. Normalizing by joint entropy combats this: the pairs will have equal NMI (Martin et al., 2005).	Low compute time. Null hypothesis of independence accounts for shared ancestry by controlling the false discovery rate. It uses "z-scores or standardization", which computes the departure of a given value from the mean of all observations, in standard deviation units.	Coevolution statistic does not use explicit phylogenies
5. Resample mutual information (RMI)	The resampling approach corrects MI by creating permutations of the data for the pair of aligned columns (Easton, 2006). The modified data sets are identical to the observed data except for a specific residue whose observed state is replaced by one of the alternate states present in other sequences at that position. Accordingly, the modified data set has close to identical shared ancestry. The frequency with which such permuted data sets result in a MI less than that from the observed data is taken as the probability of observing a permutation with less dependence. Thus, RMI explicitly adjusts for shared ancestry and computes probabilities of coevolution between residue pairs.	Low compute time. Null hypothesis accounts for shared ancestry by utilizing "Non-parametric bootstrap" methods to build replicate datasets by sampling from the original dataset.	Coevolution statistic does not use explicit phylogenies
6. Corrected mutual information (MIp)	Corrected mutual information (MIp) is a MI-based approach (Dunn et al., 2008). It attempts to control for 'background MI', or MI arising from random variation or shared ancestry. Initially raw MI is calculated and the MIp score is then computed by subtracting the product of the mean MIs for the two positions divided by the overall mean MI score from the MI score for the pair of positions.	Low compute time. Coevolution statistic accounts for shared ancestry by accounting for background MI. Also Null hypothesis incorporates "z-scores or standardization" to compute the departure of a given value	Coevolution statistic does not use explicit phylogenies

Method	Short Description	Strong points	Weak points
		from the mean of all observations, in standard deviation units.	
7. Statistical coupling analysis (SCA)	Statistical coupling analysis (SCA) measures the change in distribution of residues at one position associated with a change in the distribution of residues at another position (Lockless and Ranganathan, 1999, Suel et al., 2003). If a correlated change in the distribution of residues exists between a pair of positions, that pair is said to be statistically coupled and potentially coevolving. Statistical coupling of a pair of positions *i* and *j* is calculated by selecting a subset of the MSA (the sub-alignment) where the residue at position *j* is fixed. The difference in the distribution of amino acid residues occurring at position *i* is calculated between the full MSA and the sub-alignment as the difference in their Euclidean norms.	Low to moderate compute time. Null hypothesis accounts for shared ancestry by controlling the false discovery rate. It uses a measure similar to *z*-scores.	Coevolution statistic does not use explicit phylogenies

3.1.4.2 Tree-based methods

The biggest drawback with tree-independent methods is lower specificity due to confounding of correlations arising from selective pressure with correlations arising from shared ancestry represented by the phylogeny (Atchley et al., 2000, Pollock and Taylor, 1997, Dutheil et al., 2005). All tree-based methods require both a MSA of the given protein sequences and a corresponding phylogenetic tree as input for the further analysis.

Several algorithms have been developed in this category by harnessing the vast resources of the statistical framework routinely used by phylogeneticists: Ancestral states (Shindyalov et al., 1994, Tuff and Darlu, 2000)), CoMap algorithm (Dutheil et al., 2005), Generalized Continuous-Time Markov Process Coevolutionary Algorithm (GCTMPCA) (Yeang et al., 2007, Yeang and Haussler, 2007), and LnLCorr (Pollock et al., 1999, Wang and Pollock, 2007). Table 3.2 gives a comparison of tree-based methods, categorized according to their strong and weak points.

3.1.4.3 Tree-independent *vs* Tree-based methods

Generally, tree-based methods have performed extremely well with simulated data, but have been scarcely utilized by biologists, mainly due to high computational requirements. In contrast, tree-independent methods like MI and SCA have found wider applications. The phylogenetic dependency of sequences is acknowledged in the evolutionary biology literature, but is often not suitably accounted for. The tree-based methods such as LnLCorr and GCTMPCA are unquestionably the best options available for modeling and studying the process of coevolution in biological sequences, but suffer from computer-resource demands, which prevent their use on large and/or numerous data sets (Dutheil, 2011). Several tree-independent methods such as Normalized mutual information (NMI), Resample mutual information (RMI), Corrected mutual information (MIp) and SCA incorporate means to compare the coevolution statistics to a background distribution of scores with the same underlying phylogeny, which reduces false positives arising from phylogenetic effects. Many studies (Caporaso et al., 2008, Horner et al., 2008, Dutheil, 2011) have compared coevolution detection algorithms, but no method comes across as best on a consistent basis.

Table 3.2 Comparison of tree-based methods

Method	Short Description	Strong points	Weak points
1. Ancestral States	"Ancestral States" (Shindyalov et al., 1994) improved by (Tuff and Darlu, 2000) is one of the earliest methods developed to reveal coevolution in proteins. In this method, ancestral states are inferred for each internal node of the provided phylogenetic tree using maximum likelihood with a substitution model calculated from the alignment and tree. For each pair of positions in the alignment, all pairs of organisms were evaluated to score position pairs based on the number of times both underwent a substitution since their last common ancestor.	Coevolution statistic includes shared ancestry. Null hypothesis accounts for shared ancestry. Coevolution statistics uses explicit phylogenies.	High compute time
2. Comap	The CoMap algorithm (Dutheil et al., 2005) is similar to Ancestral States in that it relies on reconstruction of the ancestral states of all positions in the alignment. However, instead of simply counting the number of co-occurring substitutions, CoMap builds 'substitution vectors' for each position, where each element in the vector represents a change in a corresponding branch of the phylogenetic tree. Coevolving positions are identified as those with correlated substitution vectors.	Coevolution statistic includes shared ancestry. Null hypothesis accounts for shared ancestry. Coevolution statistic uses explicit phylogenies.	Moderate to high compute time
3. Generalized Continuous-Time Markov Process Coevolutionary Algorithm (GCTMPCA)	The generalized continuous-time Markov process coevolutionary algorithm (GCTMPCA) employs maximum likelihood to determine if pairs of substitution events are more likely under a dependent or independent model of evolution (Yeang and Haussler, 2007). GCTMPCA requires a single parameter, ϵ, the penalty incurred for a single residue change (as opposed to a correlated change between two residues), be provided by the user. Based on empirical evidence, the authors	Coevolution statistic includes shared ancestry. Null hypothesis accounts for shared ancestry. Coevolution statistic uses explicit phylogenies.	High compute time

Method	Short Description	Strong points	Weak points
	define 0.7 as the optimal value of the parameter.		
4. LnLCorr	LnLCorr uses a likelihood ratio (LR) to compare the probability of the data under independent and dependent models of evolution. In this implementation, a larger LR between a pair of positions suggests coevolution (Pollock and Taylor, 1997, Wang and Pollock, 2007). LnLCorr differs from the other methods in that it incorporates its own amino acid alphabet reduction step based on a residue metric provided by the user. The method uses a simple two-state evolutionary model allowing, for example, discrimination only between positively and negatively charged amino acids or large and small amino acids.	Coevolution statistic includes shared ancestry. Null hypothesis accounts for shared ancestry. Coevolution statistic uses explicit phylogenies.	High compute time

3.1.5 Rubisco large subunit

3.1.5.1 Name conventions used in this study

Rubisco- specific abbreviations used in this chapter are summarized in Table 3.3.

Table 3.3 Rubisco specific abbreviation used in this study

Abbreviation	Full name
rbcL	Rubisco large subunit gene
rbcS	Rubisco small subunit gene
R-LSU	Rubisco large subunit protein
R-SSU	Rubisco small subunit protein
RA	Rubisco activase protein
RbcX	RbcX protein (Rubisco's chaperone)
Superscript ^{L}	To indicate a R-LSU site
Superscript ^{S}	To indicate a R-SSU site
Superscript ^{X}	To indicate a RbcX site
Superscript ^{RA}	To indicate a RA site

3.1.5.2 Rubisco large subunit, an ideal system to study coevolution?

Rubisco large subunit is part of the Rubisco holoenzyme in higher plants. The holoenzyme consists of eight large subunits (LSUs), encoded by the chloroplast gene *rbcL,* assembled into four dimers, and eight small subunits (SSUs) encoded by the nuclear gene *rbcS*. Two active sites are formed at the intra-dimer interface from the C-terminal, α/β barrel domain of one LSU and the N-terminal domain of another, thus making the L_2 dimer the basic catalytic unit of the enzyme.

It has been noted by plant systematists (Albert et al., 1994) that *rbcL* evolution appears to be strongly constrained by its function. Estimates of synonymous nucleotide substitution rates for *rbcL* sequences are approximately 4-5 fold lower than estimates from plant nuclear protein-coding genes (Clegg, 1993).

Factors that might underlie the slow evolution of *rbcL* are the complex tertiary structure of Rubisco, the requirement to catalyze a complex multistep series of chemical reactions, and its interactions with other proteins during the course of its assembly, activation and re-activation. Within the Rubisco holoenzyme, R-LSU has to deal with selection forces acting against mutations that could destabilize intra-dimer (LSU-LSU), inter-dimer (L_2-L_2) and inter-subunit (LSU-SSU) interactions. Interactions with Rubisco activase (RA), Rubisco- LSU methyl-transferase and chaperonins, such as RbcX, further reduce the already constrained "residue space" that R-LSU can sample evolutionarily by mutation while maintaining sufficient activity in all catalytic steps to constitute a viable enzyme. Because of these inherent functional/ structural constraints, it can be expected that R-LSU has evolved only slowly.

Consequently, every evolutionary change optimizing Rubisco's function has likely been subjected to strong selection forces, due to the tight link between its function and the biological fitness of the plant (Sen et al., 2011). In accordance with the neutral theory of molecular evolution (Kimura, 1983), it can be assumed that advantageous mutations in Rubisco would be favored by adaptive evolution, while deleterious mutations would be removed by purifying selection.

Significant positive selection events have been identified in the *rbcL* genes of most land plant lineages (Kapralov and Filatov, 2006, Kapralov and Filatov, 2007, Christin et al., 2008, Kapralov et al., 2011). How are these developments manifested at the molecular level? To understand Rubisco's functional landscape, adaptive evolution analysis of *rbcL* alone will not suffice, in view of its complex structural and functional constraints and its reliance on interactions with other proteins to accomplish its function. The identification of complex coevolutionary processes both within R-LSU and between R-LSU and its interacting partners will provide a better understanding of R-LSU's fine-tuning at molecular level.

Coevolutionary studies have been applied to many protein families e.g. cytochrome c oxidase (Wang and Pollock, 2007), dihydrofolate reductase, cyclophilin and formyl-transferase (Saraf et al., 2003), and 91 protein families from HSSP (database of homology-derived protein structures) (Shindyalov et al., 1994); these provided new information about protein-protein interactions, ligand-receptor binding, and 3D protein structure. Two recent studies on *rbcL* (Sen et al., 2011, Wang et al., 2011), one on Gymnosperm *rbcL*, and other on Angiosperm *rbcL,* attempted to uncover correlated interactions within the R-LSU. In this study I have looked into coevolution both within R-LSU and with its interacting partners (RA, R-SSU and RbcX).

There is a wealth of sequences available for R-LSU (~80,000) in public databases from eukaryotes (plants and algae), archaea and prokaryotes (autotrophic bacteria). The number of sequences for R-LSU's interacting partners are comparatively small: R-SSU (~1000), RbcX(~600) and RA(~200) are an adequate starting point. This wealth of sequences for R-LSU and availability of its interacting partner's sequences in reasonable numbers makes it a good candidate for study of coevolutionary processes by coevolution-detection algorithms. Note that availability of a large number of *rbcL* sequences doesn't necessarily translate into a balanced dataset; due care had been taken to construct a sufficiently balanced and diverse dataset for the current study.

3.2 Methods

3.2.1 Data Preparation

Sequences were downloaded from public databases; species name and accession numbers are given in Appendix 3. Angiosperm R-LSU sequences were organized into 47 monophyletic groups, according to the taxonomic classification downloaded from NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/). The assembled sequences were edited using BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html). Sequences were aligned using ClustalW (Thompson et al., 1994); alignments of more than 200 sequences were performed using a parallel version of ClustalW (Li, 2003) on the NCI (National Computing Infrastructure) Oracle/Sun Constellation Cluster at located at ANU. All alignments were straightforward, consistent with the highly conserved nature of the R-LSU. I found only one insertion at position 469 in the alignment of Angiosperms, which differentiated a few C_4 plants (22) from the rest of the analyzed lineages. It should be noted that many sequences are incomplete and lack residues at the N-terminal and/or C-terminal end. Sequences < 450 residues in length were excluded from analysis. Also, sequences lacking residues at the C-terminus were excluded from analysis as the C-terminal tail is known to have a significant functional role both within the R-LSU (opening and closing of loop 6) and also with its interacting partners as a recognition motif (Knight et al., 1990, Bracher et al., 2011).

Additionally, sequences of R-LSU interacting partners, R-SSU, RbcX and RA were also downloaded from the NCBI Genbank and aligned. Both R-SSU and RA contain N-terminal signal-peptide sequence (required for transport into chloroplast), which was removed from their respective sequences before further analysis.

3.2.2 Coevolution Analysis

Although a single method did not emerge as the overall best choice from method comparison in the literature (Caporaso et al., 2008, Horner et al., 2008), I adopted the joint-entropy-normalized mutual information (NMI) as my method-of-choice for coevolutionary analysis of R-LSU for a variety of reasons (explained in more detail in the next section). I employed the Caporaso et al. (2008) implementation of NMI in PyCogent (http://pycogent.sourceforge.net/) developed by Knight et al. (2007) for the analysis. In a given MSA, only sites with entropy >0.3 were selected for further analysis. An entropy cutoff of 0.5 was used for inter-protein analyses as the datasets for RA, RbcX and R-SSU are small. NMI scores of sites were standardized (by calculating z-scores) and only sites which have z-score > 6 were identified as coevolving sites or as otherwise noted.

I begin by introducing in more detail the joint-entropy-normalized mutual information metric.

3.2.2.1 Joint-entropy-normalized mutual information

The Shannon entropy (*H*) of a position *a* in a multiple sequence alignment is a measure of its variability. For a set of discrete states $X= \{x_1, x_2 ....x_n\}$, Shannon entropy (Shenkin et al., 1991) is computed as:

$$H_a = -\sum_{i=1}^{n} p(x_i) \, . \, log_2 p(x_i) \tag{3.1}$$

In the case of protein sequence alignments, the states are the amino acid residues, and the probability for observing each state ($p(x_i)$) is computed as the frequency of that state at position *a* in the alignment. In practice, the base of the logarithm is not important as long as it is consistent; conventionally, base 2 is used making bits the units of *H*. If one of the states is not observed at position *a*, as is nearly always the case in protein sequence alignments, it is taken that $0 \log_2 0 = 0$. The entropy at a position decreases with conservation, so a perfectly conserved position has $H = 0$.

The Shannon entropy for a pair of positions *a* and *b*, or the joint entropy, is computed similarly except that the set of states is now all possible pairs of states: XY = $\{x_1y_1, x_2y_2..........x_my_n\}$. The joint entropy (Martin et al., 2005) calculation is:

$$H_{ab} = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_iy_j) \,.\, log_2 p(x_iy_j) \tag{3.2}$$

In the context of a multiple sequence alignment, the Mutual Information for a pair of positions *a* and *b* (MI_{ab}) is a measure of the degree to which knowing the identity of the residue at position *a* provides information of the residue at position *b* (or vice versa: $MI_{ab} = MI_{ba}$). More generally, *MI* is a measure of the degree to which knowing the value of one discrete random variable provides information about the value of another discrete random variable. MI_{ab} is calculated as the sum of the Shannon entropies (H_a *and* H_b) at each position minus the joint entropy (Martin et al., 2005) of the positions (H_{ab}).

$$MI_{ab} = H_a + H_b - H_{ab} \tag{3.3}$$

The joint-entropy normalized Mutation Information for positions *a* and *b* (NMI_{ab}) is simply (Martin et al., 2005):

$$NMI_{ab} = \frac{MI_{ab}}{H_{ab}} \tag{3.4}$$

3.2.2.2 Rationale for selecting NMI for this analysis

3.2.2.2.1 NMI removes the effect of evolutionary rate heterogeneity among sites

Because mutual information is normalized by the joint entropy of the pair of sites, rate heterogeneity among sites is controlled for, and therefore does not affect the covariation statistic. It has been noted that rate heterogeneity is an inherent problem with many coevolution detection algorithms, including MI and SCA (Fodor and Aldrich, 2004) .

3.2.2.2.2 Minimum and maximum NMI scores are clearly defined

As with standard mutual information (MI), the minimum NMI value is 0.0. As the maximum mutual information score for a pair of positions *a* and *b* is obtained when the residue patterns of the two positions are identical, and are therefore both identical to the pattern of the combined positions,

$$H_a = H_b = H_{ab} \tag{3.5}$$

If we call this quantity H_0, the NMI calculation follows as:

$$MI_{ab} = H_a + H_b - H_{ab} = H_0 + H_0 - H_0 = H_0 \tag{3.6}$$

$$NMI_{ab} = \frac{MI_{ab}}{H_{ab}} = \frac{H_0}{H_0} = 1.0 \tag{3.7}$$

NMI therefore has a maximum value of 1.0, and is interpreted as the proportion of the maximum possible MI at a pair of positions which is observed. The clear upper and lower bounds on NMI make it a convenient statistic to work with.

3.2.2.2.3 NMI does not require an evolutionary model

Because NMI does not require an evolutionary model (as the tree-based methods do) it is not possible to miss-specify the evolutionary model. Additionally, gap characters do not pose a problem for the analysis, as they can be treated simply as any other alignment character.

3.2.2.2.4 NMI is fast

NMI is relatively fast to compute for all pairs of positions in an alignment. When run on the R-LSU data sets, it was consistently among the fastest methods.

3.2.2.3 *z*-score calculation

z-score or standard score indicates how many standard deviations an observation is above or below the mean. It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This conversion process is called standardizing. The standard score is

$$Z = \frac{x-\mu}{\sigma} \tag{3.8}$$

where:

x is a raw score to be standardized;

μ is the mean of the population;

σ is the standard deviation of the population.

The quantity *z* represents the distance between the raw score and the population mean in units of the standard deviation. *z* is negative when the raw score is below the mean, positive when above.

3.2.2.4 Analyses performed in this study

I performed three different coevolution analyses in this study: i) coevolution analysis using all Angiosperm R-LSU sequences, ii) coevolution analysis of R-LSU at order level (based on NCBI taxonomy), and iii) Inter-protein analysis involving R-LSU-R-SSU, R-LSU-RbcX and R-LSU-RA.

For coevolution analysis of all Angiosperm R-LSU sequences, all available Angiosperm R-LSU sequences from NCBI were used, with sequences less than 450 residues in length or lacking residues at C-terminus excluded from the analysis (see section 3.2.1). The final alignment comprised 5052 sequences (see Appendix 3.4 for sequence ids) and 450 residue positions of the R-LSU, (residues 26 to 475) as many sequences were missing residues in the N-terminal region.

At order level, the coevolution analysis of four orders (Solanales, Gentianales, Poales and Caryophyllales, for sequence ids, see Appendix 3.4), chosen due to the presence of unique sequence signatures in the R-LSU protein as found in the literature or observed in the course of this study, was performed with the background dataset. Details are summarized in Table 3.3.

Table 3.3 Details of number of sequences included in Coevolution Analysis at order level

Order	Residue positions differing from highest-frequency residue of Angiosperms	Number of sequences from order	Number of sequences from background dataset
Solanales	89 and 94	91	50
Gentianales	95	330	110
Caryophyllales	32	155	52
Poales	91 and 464	154	59

To define the background dataset, I utilized position-wise residue frequency statistics of > 10,000 Angiosperm plant R-LSU protein sequences compiled by Dr. Babu Kannappan in our lab to identify the residue with the highest frequency for each position of R-LSU. The background dataset was created from sequences conforming to the highest frequency residue in each position. For instance, in order Solanales, R-LSU positions 89 and 94 are known to be Arg and Lys (Larson et al., 1997, Ott et al., 2000), whereas the highest frequency residues for these sites in Angiosperm R-LSU as a whole are Pro and Glu, respectively. So, the background dataset will include sequences with residue Pro in position 89, and Glu in position 94. The rationale behind this exercise is to have variation in positions 89 and 94 in the Solanales coevolution dataset. This helps in identifying the other positions in the alignment that are unique to Solanales (See Figure 3.1 for illustration). In general the number of sequences in the background dataset is one half to one third of the number of sequences utilized in analyses of individual orders.

	Residue positions in R-LSU alignment			
	86 (found)	89 (known)	94 (known)	95 (found)
Angiosperms	H	P	E	N
Solanales	R	R	K	D

Figure 3.1 Rationale behind using background dataset. In order Solanales, R-LSU positions 89 and 94 are known to be Arg and Lys (from literature), whereas highest-frequency residues for these sites in Angiosperm R-LSU are Pro and Glu, respectively. Using background dataset with variation in positions 89 and 94, allow identification of variation in positions 86 and 95.

In the inter-protein analysis, R-LSU-RA (23 sequences, see Appendix 3.1), R-LSU-RbcX (14 sequences, see Appendix 3.3) and R-LSU-R-SSU (44 sequences, see Appendix 3.2) are included in the analysis.

3.3 Results

3.3.1 Coevolution Analysis using all-Angiosperm R-LSU sequences

Altogether 15 sites were found to be coevolving in all-Angiosperm R-LSU sequences (Figures 2A, B). The identified coevolving residues are clustered in groups of 2 - 7 residues, and are mostly located in the C-terminal domain. In the N-terminal domain one single-site pair (95, 97) was found to be coevolving. Site pairs (247, 282), (439, 466) and (466, 468) are the only three single-site pairs coevolving in the C-terminal domain. The remainder of the sites formed a network of coevolving sites in the C-terminal domain, which also included site 91 from the N-terminal domain.

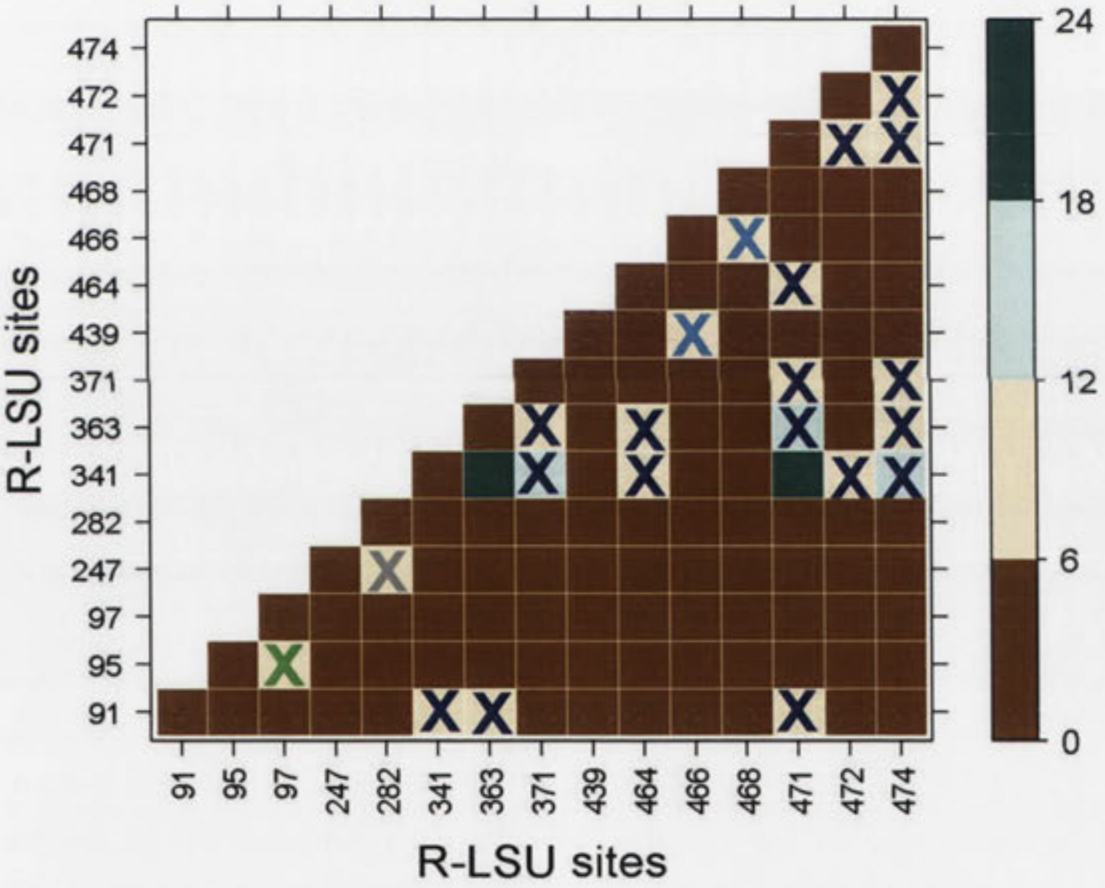


Figure 3.2A. Coevolution analysis of all-Angiosperm R-LSU sequences using NMI. The NMI *z*-scores matrix (*z*>6) of the all-Angiosperm R-LSU MSA is plotted. On the colour scale, the z-score ranges from 0 to 24. The minimum and maximum values in the matrix are 6.1 and 22.6, respectively. The x and y axes show R-LSU sites. Spinach R-LSU numbering is used for cross comparison convenience. In total, 15 coevolving sites were detected. All clustered and single pair residues are marked "X" in same color as shown in cluster diagram in Figure 3.2B.

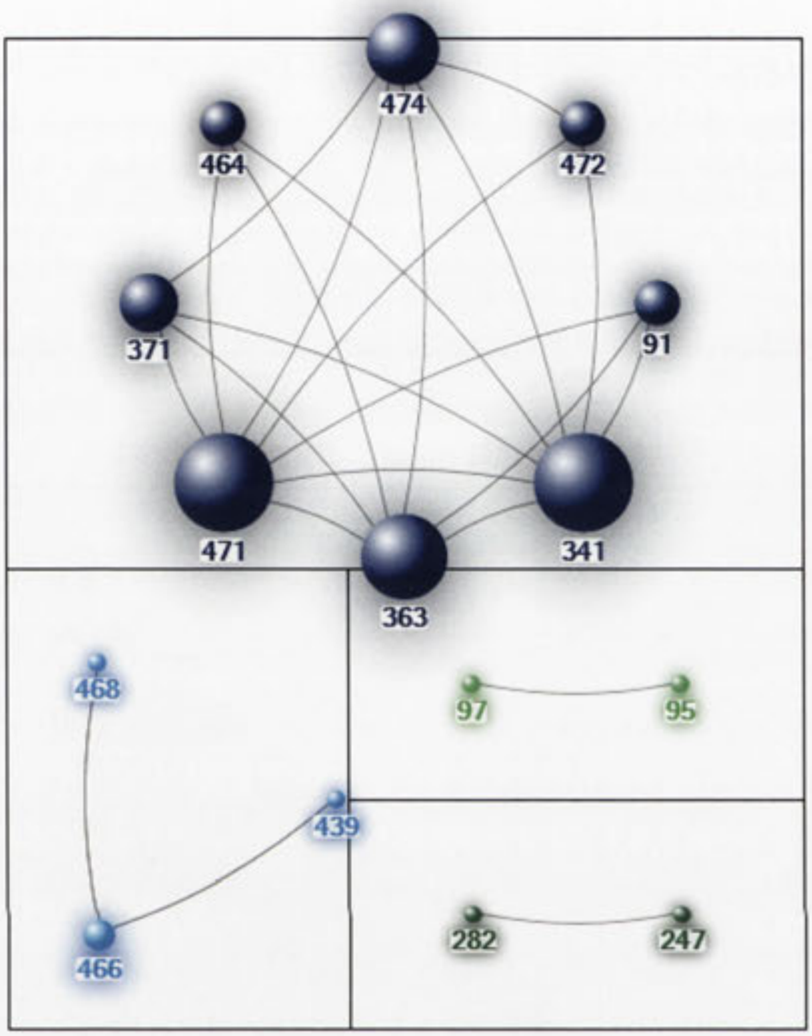


Figure 3.2B Network plot of coevolution Analysis of all-Angiosperm R-LSU sequences. Single-pair sites (95, 97), (247, 282) and (439, 466) and (466, 468) can be seen as isolated single pairs, whereas sites 91, 341, 363, 371, 464, 471, 472 and 474 can be seen as clustered together. The Girvan-Newman algorithm was used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

Sites 91, 341, 363, 371, 464, 471,472 and 474 are found to be strongly coupled in this analysis (Figure 3.3). Most of these sites are more hydrophobic {91(Ala/Pro/Val), 341(Ile/Met), 363(Tyr/Phe), 371(Leu/Met), 471(Ala/Pro), 472(Met/Val)}, except two of the sites {464(Glu/Ala) and 474 (Thr/Lys)}. Site 341 is part of Helix 6 in the C-terminal domain, very close to loop 6; a conformational change in this loop is known to be required to release tightly-bound inhibitor, thus making Rubisco ready for catalysis (Knight et al., 1990). Sites 464, 471,472 and 474 are part of the C-terminal tail, whereas site 91 is located in a region identified as involved in RA recognition in the N-terminal domain (Andersson and Backlund, 2008).

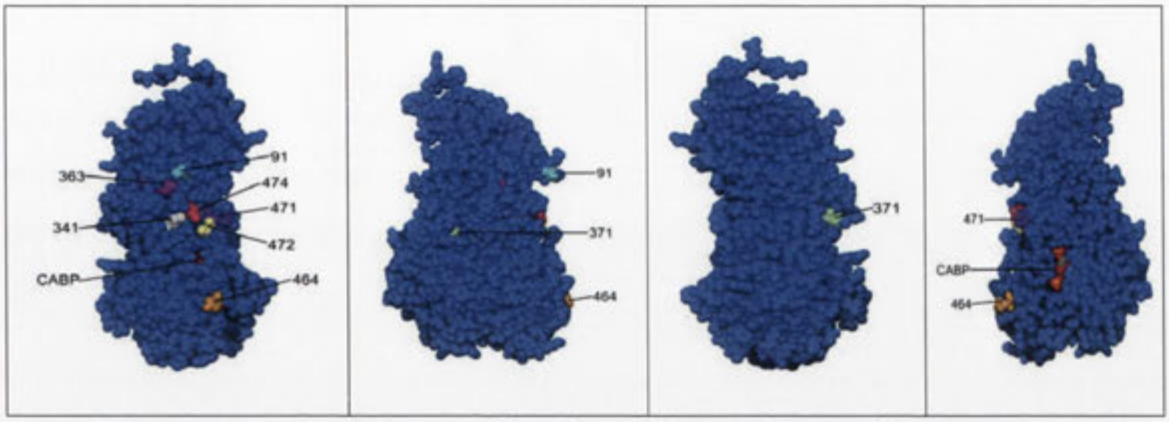


Figure 3.3 Cartoon representations of the clusters detected in the all-Angiosperm R-LSU coevolution analysis (shown in 4 different orientation obtained by 90 deg rotation along the vertical axis, pdb id 8RUC, CABP stands for 2-carboxyarabinitol-1,5-diphosphate, an inhibitor of Rubisco's catalytic reaction). R-LSU sites 91, 341, 363, 371, 464, 471, 472 and 474 are depicted on the monomer of R-LSU. Except for site 371, all the clustered sites are on one face of the R-LSU. Site 341 is part of helix 6 in the C-terminal domain, very close to loop 6. Sites 464, 471,472 and 474 are part of the C-terminal tail.

3.3.2 Coevolution analysis of R-LSU at order level

Coevolution analysis was also carried out at order level, based upon NCBI taxonomy. It is generally understood that coevolution analysis based on taxonomy will suffer from noise from shared ancestry, a noise factor NMI endeavors to eliminate. But in the case of Rubisco, the reaction mechanism, role of active-site residues and structure-function relationship have been studied in some detail, so noise arising from shared ancestry can be filtered out. Moreover, it is known that Rubiscos with unique sequence signatures in the R-LSU from a few plant groups show variation in inter-protein interactions (Portis, 2003). Coevolution analysis was conducted to see if it is possible to trace the basis of this variation at order level. As mentioned in Methods (section 3.2.2.4), this analysis was carried out in 4 plant orders with unique sequence signatures.

3.3.2.1 Solanales

Solanales is an order in flowering plant which includes tomato, potato, tobacco and capsicum as its members. In *Solanaceae* (a family in order Solanales), it has been deduced from mutagenesis studies that R-LSU sites 89 and 94 interact with RA (Larson et al., 1997, Ott et al., 2000). During the course of this analysis I found that not only in family

Solanaceae but more generally in the order Solanales, the highest-frequency residue at position 89 is Arg, instead of Pro (highest-frequency residue in the Angiosperm dataset), and at position 94 it is Lys, instead of Glu (again highest-frequency residue in Angiosperm dataset). This is intriguing because both are non-conservative substitutions and position 94 in particular is dominated (over 90% Asp or Glu) by a negatively charged residue in our all-Angiosperm sequence dataset. Interestingly, the highest-frequency residue at position 95 in Solanales is Asp, (highest-frequency residue at position 95 is Asn/Ser in the Angiosperm dataset) a negatively charged residue, probably to compensate for charge imbalance at position 94.

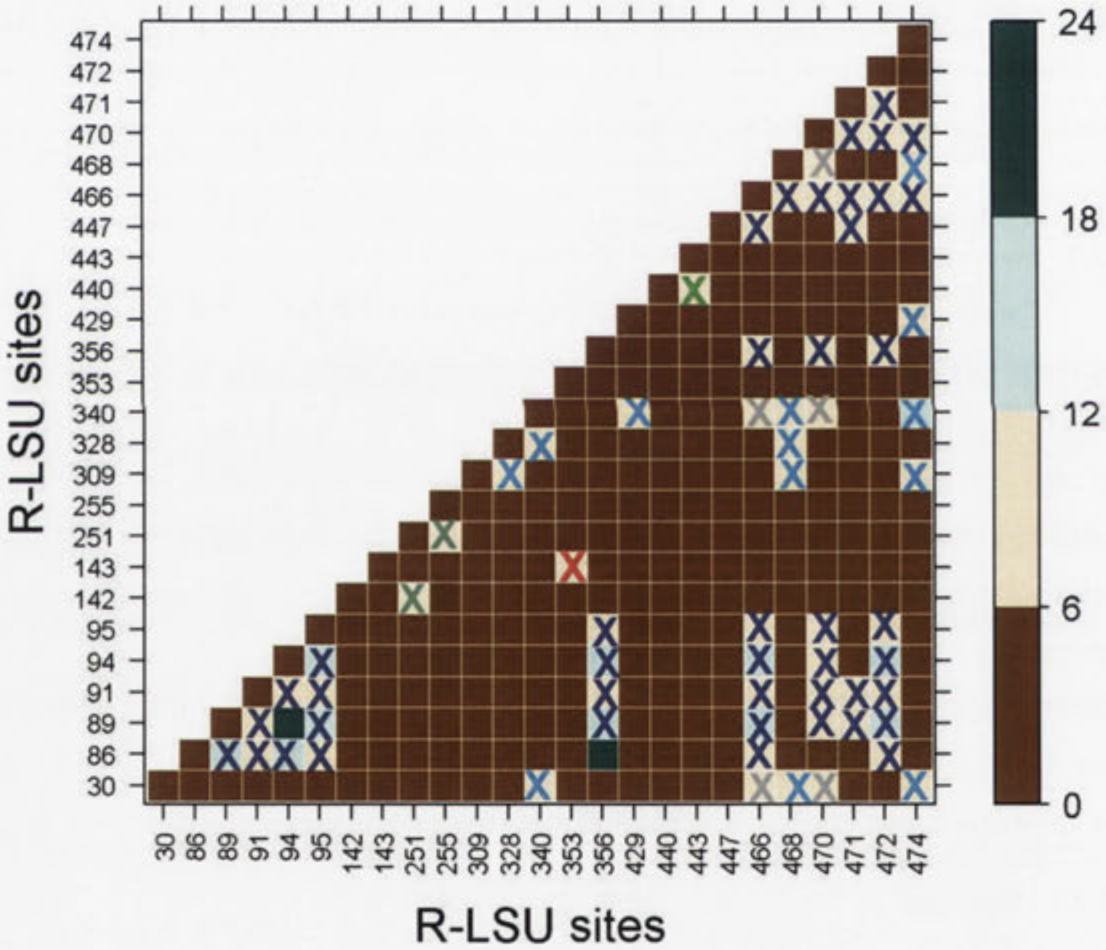


Figure 3.4A Coevolution analysis of Solanales R-LSU sequences against the background dataset using NMI. The *z*-scores of NMI matrix (*z*>6) of Solanales MSA is plotted. On the color scale, the z-score ranges from 0 to 24. The minimum and maximum values in the matrix are 6.1 and 21.6, respectively. The x and y axes show R-LSU sites. Spinach R-LSU numbering is used for cross comparison convenience. In total, 25 coevolving sites were detected. All clustered and single pair residues are marked "X" in same color as shown in cluster diagram in Figure 3.4B. Inter-cluster connections are shown in grey.

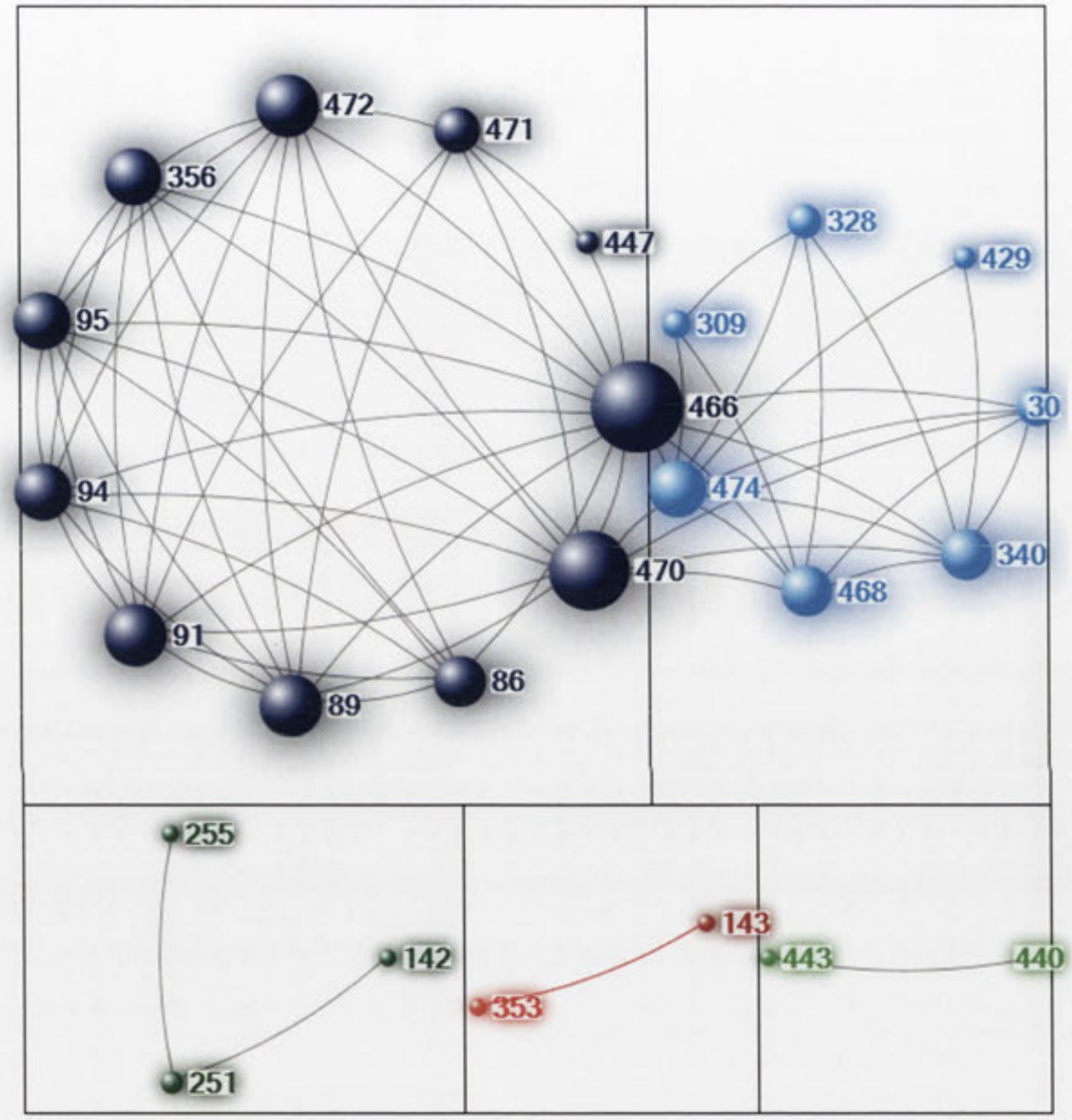


Figure 3.4B Network plot of coevolution analysis of Solanales R-LSU sequences against the background dataset. Single pair sites (142,251), (143, 353), (251, 255) and (440,443) can be seen as isolated single pairs, whereas two major clusters i.e. cluster1 {86, 89, 91, 94, 95, 356, 447, 466, 470, 471 and 472} and cluster2 {30, 309, 328, 340, 429, 468 and 474} can be seen as having co-cluster interactions. In total, 25 coevolving sites were detected in this analysis. The Girvan-Newman algorithm was used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

In total, 25 sites were found to be coevolving in Solanales (Figure 3.4A and B). The analysis revealed two major clusters, i.e. cluster1 {86, 89, 91, 94, 95, 356, 447, 466, 470, 471 and 472} and cluster2 {30, 309, 328, 340, 429, 468 and 474} of 11 and 7 sites, 6 from the N-terminal domain and 12 from the C-terminal domain. All 6 sites from the N-terminal domain, i.e. 30, 86, 89, 91, 94 and 95 are surface accessible. Moreover site 30 is part of a loop between β-strands A and B, whereas the other N-terminal domain sites 86, 89, 91, 94 and 95 are flanking the loop between β-strands C and D. Most of the clustered C-terminal sites 466, 468, 470, 471, 472 and 474 are from the C-terminal tail. Of the other C-terminal sites, 340 is spatially close to loop 6; 356 is part of β-strand G; 429 is part of helix 8; and 447 is part of helix G. The site pair (328, 340) is notable among cluster2 residues, as both these sites flank the start and end of loop 6. Site 309 is part of β-strand F and has recently been shown to act as a catalytic switch between C_3 and C_4 Rubiscos (Whitney et al., 2011b).There are also four single coevolving pairs (142, 251), (143, 353), (251, 255) and (440, 443). Additionally, sites 142 and 143 are located on the inter-dimer interface and sites 251 and 255 are at R-SSU interface of R-LSU suggesting these sites co-evolve as part of evolving inter-protein interactions.

3.3.2.2 Poales

The plant order Poales is the most economically significant order of monocots and possibly the most crucial order of plants in general as it include the major food cereals rice, wheat, barley, maize and millet.

In Poales, the highest-frequency residue is Lys for position 14 but significant numbers of sequences with residue Gln are also found. Interestingly, with only a few exceptions, Angiosperm sequences with Gln14 belong to Poales (from Angiosperm position-wise frequency statistics compiled by Dr. Babu Kannappan). Site 14 has been shown to be methylated in tobacco (Raunser et al., 2009), and could be functionally important. There are also reports linking positions 14, 95 and 477 to kinetic properties of Rubisco (Terachi et al., 1987).

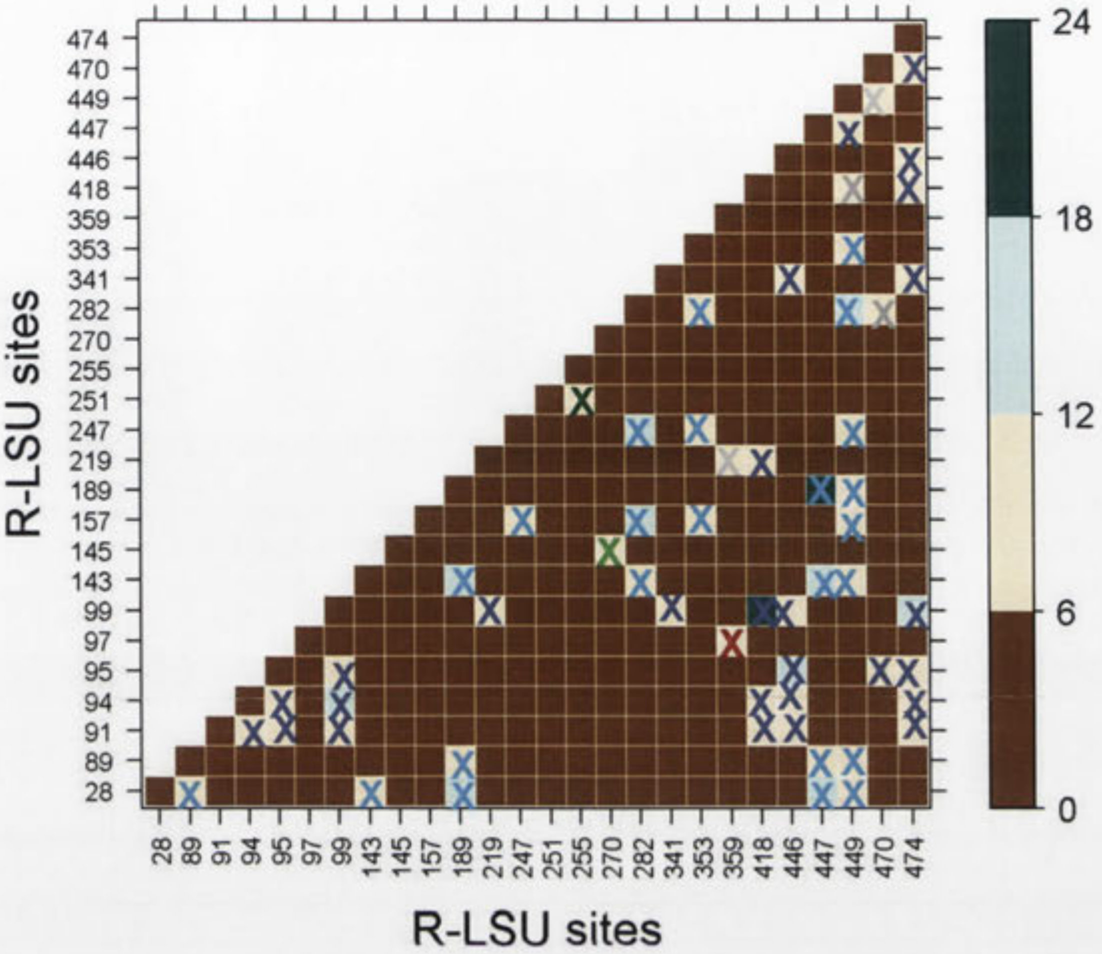


Figure 3.5A Coevolution analysis of Poales R-LSU sequences against the background dataset using NMI. The NMI *z*-scores matrix (*z*>6) of Poales MSA is plotted. In the colour scale, the z-score ranges from 0 to 24. The minimum and maximum values in the matrix are 6.1 and 21.3, respectively. The x and y axes show R-LSU sites. Spinach R-LSU numbering is used for cross comparison convenience. In total 26 coevolving sites were detected. All clustered and single pair residues are marked "X" in same color as shown in cluster diagram in Figure 3.5B. Inter-cluster connections are shown in grey.

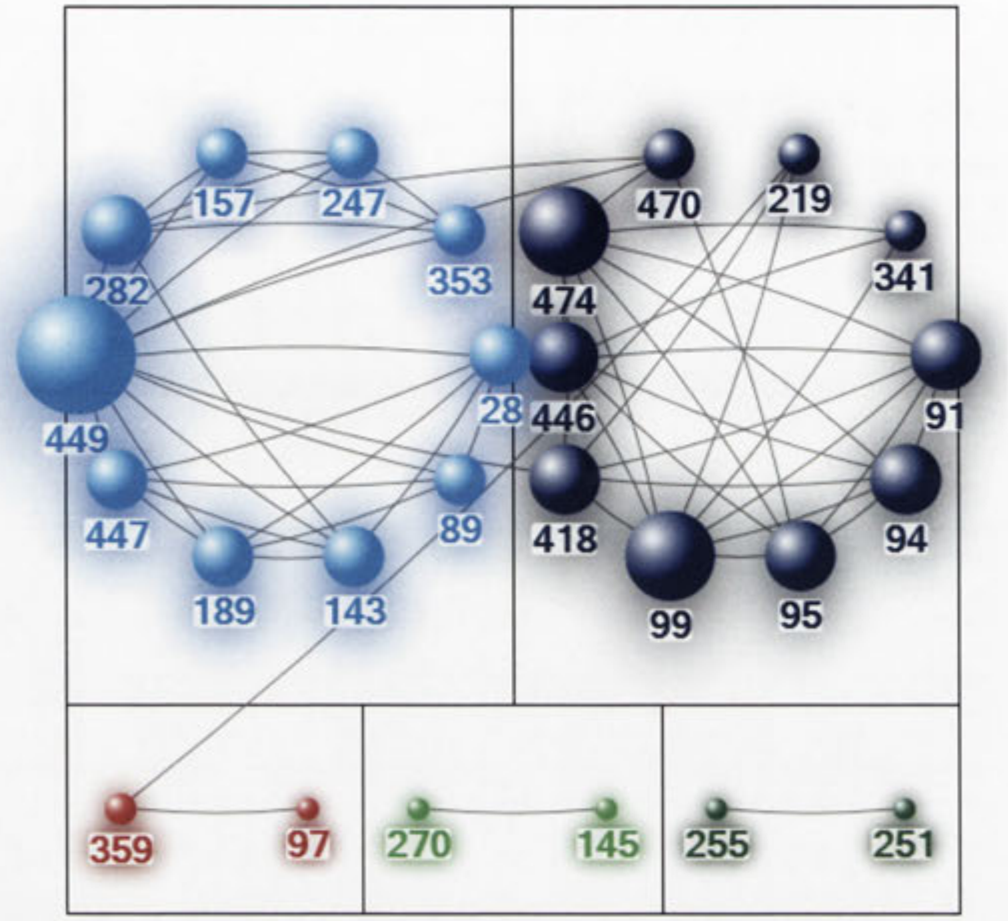


Figure 3.5B Network plot of Coevolution analysis of Poales R-LSU sequences against the background dataset. Single pair sites (97,359), (145,270), (219,359), (219,418) and (251,255) are seen as isolated single pairs, whereas two major clusters i.e. cluster1 {91, 94, 95, 99, 219, 341, 418, 446, 470, 474} and cluster2 {28, 89, 143, 157, 189, 247, 282, 353, 447 and 449} can be seen as having co-cluster interactions. The Girvan-Newman algorithm has been used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

Coevolution analysis of R-LSU in Poales identified 26 sites (Figure 3.5A and B). Two major clusters, cluster1 {91, 94, 95, 99, 219, 341, 418, 446, 470, 474} and cluster2 {28, 89, 143, 157, 189, 247, 282, 353, 447 and 449} were identified. Cluster1 sites 91, 94, 95 and 99 are located in the RA interaction region (Andersson and Backlund, 2008) in the N-terminal domain, whereas the C-terminal domain sites 446 and 474 are surface accessible. Sites 157, 189, 247, 282, 353, 418, and 449 in cluster2 flank the hydrophobic core in the C-terminal domain but site 447 is surface accessible. The N-terminal domain sites in cluster2, 28 and 89, are surface accessible, whereas 143 is at the intra-dimer interface.

Additionally four single-pair sites (97,359), (145,270), (219,359), and (251,255) are detected to be coevolving.

3.3.2.3 Gentianales

The most well known member of plant order Gentianales is coffee. The order Gentianales is noteworthy because it shows similar variations as Solanales at site 95. It is the only order other than Solanales which has site 95 dominated by Asp; sequences from all the other Angiosperm orders have Asn or Ser. The question is does Gentianales show similar coevolution patterns as Solanales?

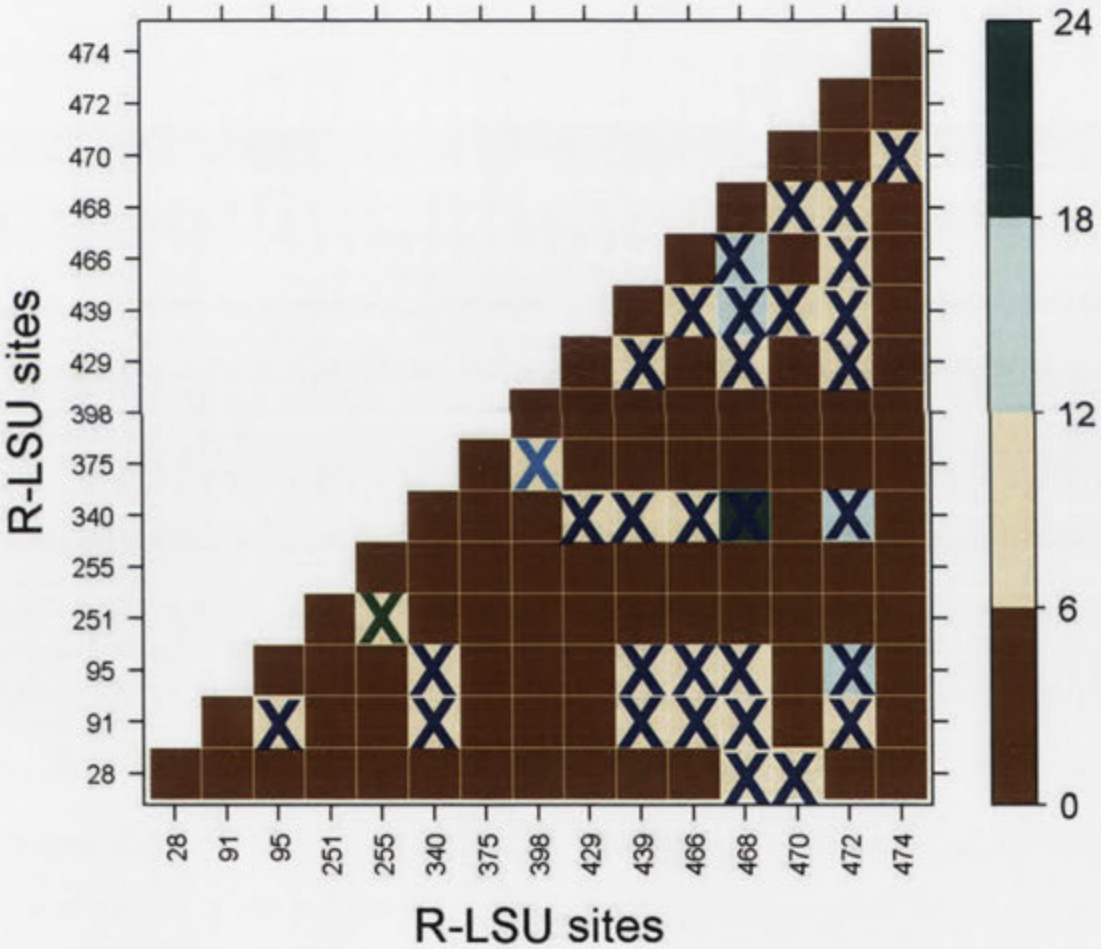


Figure 3.6A Coevolution analysis of Gentianales R-LSU sequences against the background dataset using NMI. The *z*-scores of NMI (*z*>6) of Gentianales MSA is plotted. In the colour scale, the z-score ranges from 0 to 24. The minimum and maximum values in the matrix are 6.0 and 19.7, respectively. The x and y axes show R-LSU sites. Spinach R-LSU numbering is used for cross comparison convenience. In total, 15 coevolving sites were detected. All clustered and single pair residues are marked "X" in same color as shown in cluster diagram in Figure 3.6B.

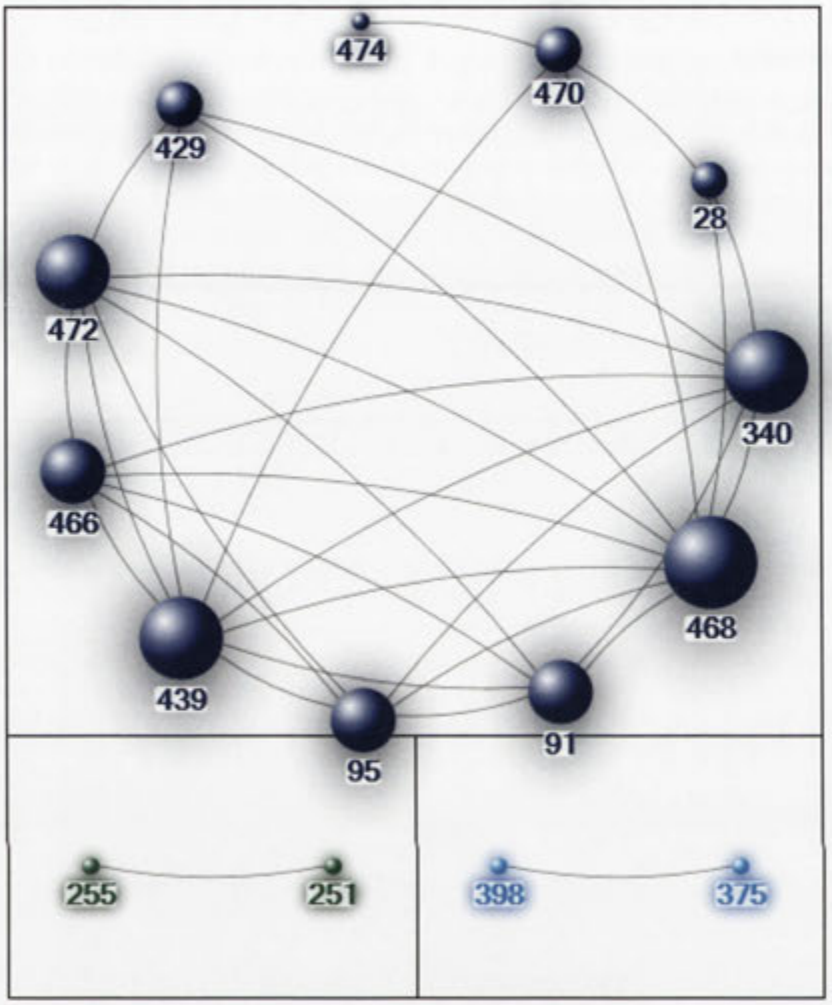


Figure 3.6B Network plot of coevolution analysis of Gentianales R-LSU sequences against the background dataset. Single-pair sites (251, 255), (375,398) and (470,474) can be seen as isolated single pairs, whereas sites 28, 91, 95, 340, 429, 439, 466,468, 470 and 472 can be seen as clustered together. In total, 15 coevolving sites were detected in this analysis. The Girvan-Newman algorithm was used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

A major cluster of 10 coevolving sites i.e. {28, 91, 95, 340, 429, 439, 466,468, 470 and 472} is identified in Gentianales (Figure s 6A and B), which is similar to the Solanales cluster in terms of location of sites with few exceptions (sites 28 and 439 being absent in Solanales). This cluster included 3 sites in the N-terminal domain, 28 (site 30 in same region) being solvent-surface accessible and 91 and 95 located in the RA interaction region (loop between βC and βD), whereas the other 7 sites are in the C-terminal domain. Among these 7 sites, site 340 flanks loop 6, 429 is close to the surface, 439 is surface accessible, while the other 4 sites {466, 468, 470 and 472} are in the C-terminal tail. Three single coevolving site pairs (251, 255), (375,398) and (470,474) were also detected in the C-terminal domain.

3.3.2.4 Caryophyllales

Caryophyllales is important as a source of food plants, including amaranth, rhubarb, quinoa, and spinach, and ornamentals such as cacti, carnations, four-o'clocks, ice plants, and globe amaranths. Coevolution analysis of the order produced some interesting results. At *z*>6, very few coevolving sites were detected. The *z* cut off had to be reduced to 3 instead of 6, to detect coevolving sites. This could be due to high sequence diversity within the order. In total, 31 sites were found to be coevolving in Caryophyllales (Figure 3.7A). These sites included three major clusters and few isolated pair of coevolving sites.

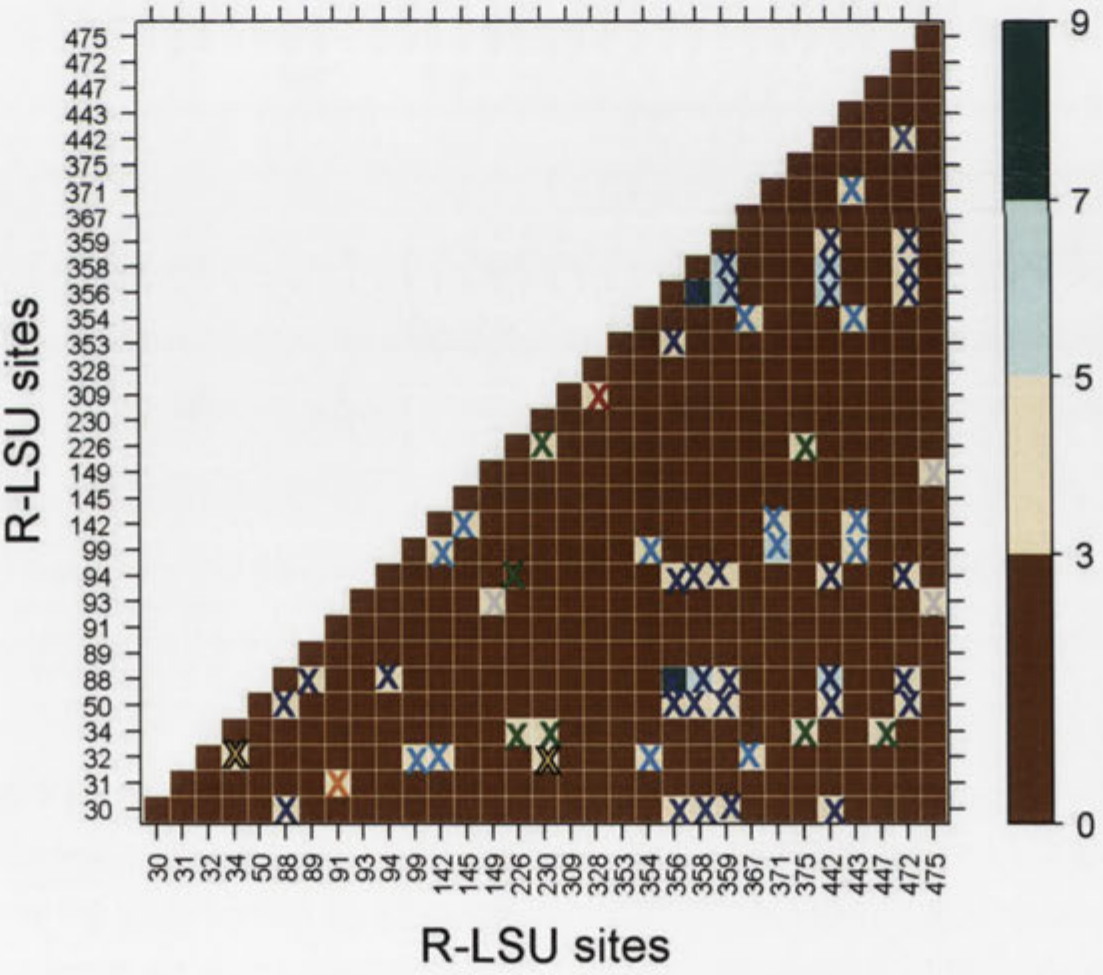


Figure 3.7A Coevolution analysis of Caryophyllales R-LSU sequences against the background dataset using NMI. The NMI *z*-scores matrix (*z*>3) of Caryophyllales MSA is plotted. In the colour scale, the z-score ranges from 0 to 8. The minimum and maximum values in the matrix are 3.0 and 7.8, respectively. The x and y axes show R-LSU sites. Spinach R-LSU numbering is used for cross comparison convenience. In total, 31 coevolving sites were detected in this analysis. All clustered and single pair residues are marked "X" in same color as shown in cluster diagram in Figure 3.7B. Inter-cluster connections are shown in black outlined yellow.

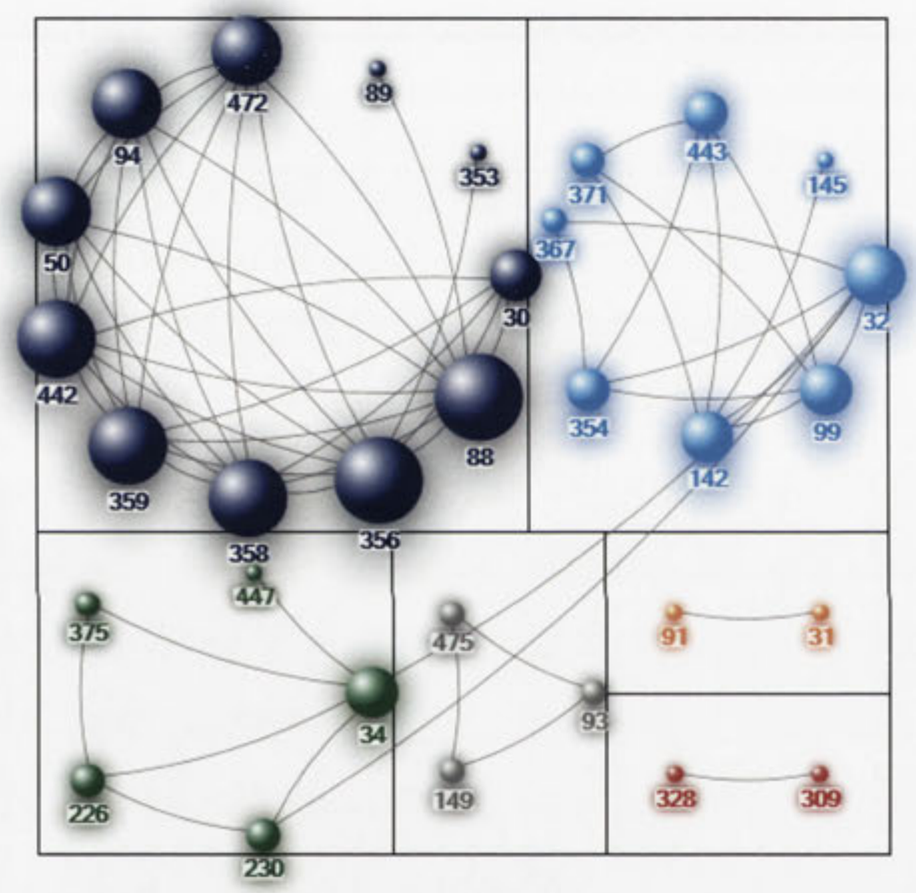


Figure 3.7B Network plot of coevolution analysis of Caryophyllales R-LSU sequences against the background. Single-pair sites (31, 91), (93,149), (93, 475), (149, 475) and (309,328) can be seen as isolated single pairs, whereas three major clusters, i.e. cluster1 {30, 50, 88, 89, 94, 353, 356, 358, 359, 442 and 472}, cluster2 {32, 99, 142, 145, 354, 367, 371 and 443} and cluster 3{34, 226, 230, 375 and 447} can be seen as clustered together. The Girvan-Newman algorithm was used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

Three clusters, cluster1 {30, 50, 88, 89, 94, 353, 356, 358, 359, 442 and 472}, cluster2 {32, 99, 142, 145, 354, 367, 371 and 443} and cluster {34, 226, 230, 375 and 447} were detected in Caryophyllales. Five sites from cluster1, i.e. 30, 89, 94, 356 and 472 were also identified in Solanales and could be part of evolving inter-protein interaction interface with RA. The single-pair site coevolving pairs are (31, 91), (93,149), (149,475) and (309,328). Interestingly, some of the coevolving sites (31^L, 32^L, 34^L, 88^L, 358^L, 359^L and 442^L) detected in this analysis are restricted to Caryophyllales, i.e. these variations are only present in Caryophyllales and may be attributed to phylogenetic correlation.

3.3.3 Inter-protein Coevolution Analysis

As discussed above, the Rubisco large subunit(R-LSU) interacts with many proteins during its life cycle; RbcX helps in assembly, R-SSU is part of the holoenzyme and RA assists in activation by releasing the inhibitors. It is likely that residues at binding interfaces in R-LSU may be coevolving with its interacting partners. To see if coevolution methods can detect this signal if applied in the inter-molecular context, I carried out seperate coevolution analyses of R-LSU with RA, RbcX and R-SSU.

3.3.3.1 Coevolution Analysis of Rubisco large subunit and Rubisco activase

Altogether, 21 sites from R-LSU were found to be coevolving with 21 sites from RA (Figure 3.8A). Sites 86^L, 89^L, 94^L, 356^L and 466^L are amongst the most prominent sites in R-LSU found to be coevolving with a number of sites from RA. In RA, in addition to sites 311^{RA} and 314^{RA}, 6 more sites 50^{RA}, 86^{RA}, 120^{RA}, 155^{RA}, 161^{RA} and 370^{RA} are notable among the sites coevolving with R-LSU.

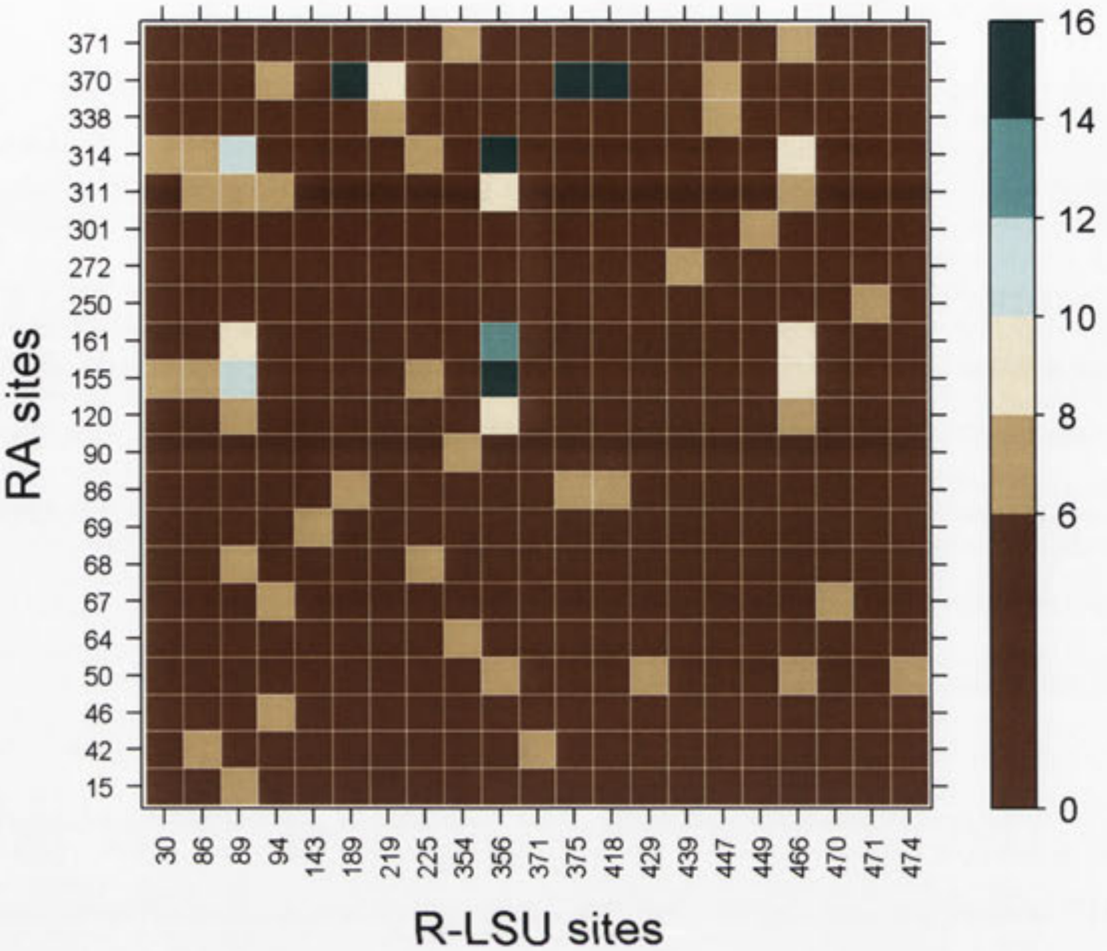


Figure 3.8A Inter-protein coevolutionary analysis of R-LSU-RA using NMI. The NMI *z*-scores matrix (*z*>6) of the analysis is plotted. On the colour scale, the z-score ranges from 0 to 16. The minimum and maximum values in the matrix are 6.0 and 15.4, respectively. The x axis shows R-LSU sites, while the y axis shows RA sites. Spinach sequence numbering is used for cross comparison convenience in both R-LSU and RA.

In R-LSU, 13 sites (30^L, 86^L, 89^L, 94^L, 356^L, 429^L, 439^L, 447^L, 449^L, 466^L, 470^L, 471^L, 474^L) out of the total of 21 coevolving sites are surface accessible and most of them are also charged/polar (Figure 3.8B).

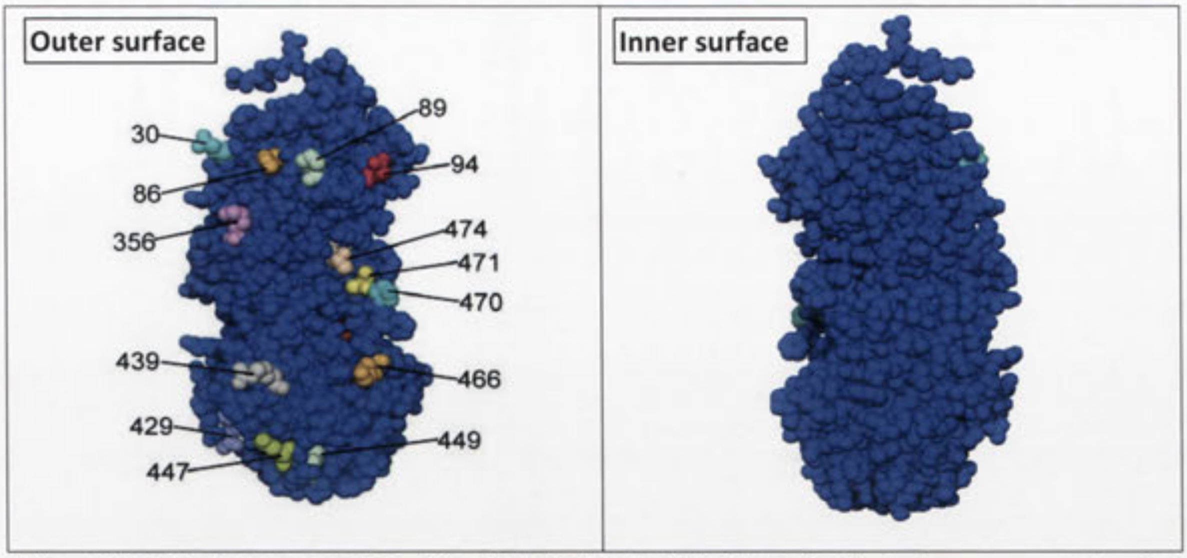


Figure 3.8B Surface accessible sites detected in R-LSU-RA analysis. Cartoon representation of sites 30, 86, 89, 94, 356, 429, 439, 447, 449, 466, 470, 471 and 474 depicted on a monomer of R-LSU. All of these coevolving sites in R-LSU are on one face (outer surface) of R-LSU. R-LSU is shown in 2 different orientation obtained by 180 degree rotation along the vertical axis, PDB id 8RUC. The sites 86, 89, 94 in the RA interaction region are in the N-terminal domain. The other N-terminal domain site 30 is also spatially proximal to this region. In C-terminal domain, some of the coevolving sites, i.e. sites 466, 470, 471 and 474, are part of the C-terminal tail.

Moreover 3 sites in R-LSU (86^{L}, 89^{L}, 94^{L}) from the N-terminal domain, are spatially proximal to the loop between β-strands C and D, shown to be part of the activase-recognition region (Andersson and Backlund, 2008). Sites 466^{L}, 470^{L}, 471^{L} and 474^{L} are in the C-terminal tail. The other 8 sites (143^{L}, $189^{L,}$ 219^{L}, 225^{L}, 354^{L}, 371^{L}, 375^{L}, and 418L) are part of the hydrophobic core, with the exception of 143^{L}, which is at the inter-dimer interface. In the case of RA, sites 50^{RA}, 86^{RA}, 155^{RA}, 161^{RA}, 311^{RA} are charged /polar, while 120^{RA}, 314^{RA} and 370^{RA} are hydrophobic.

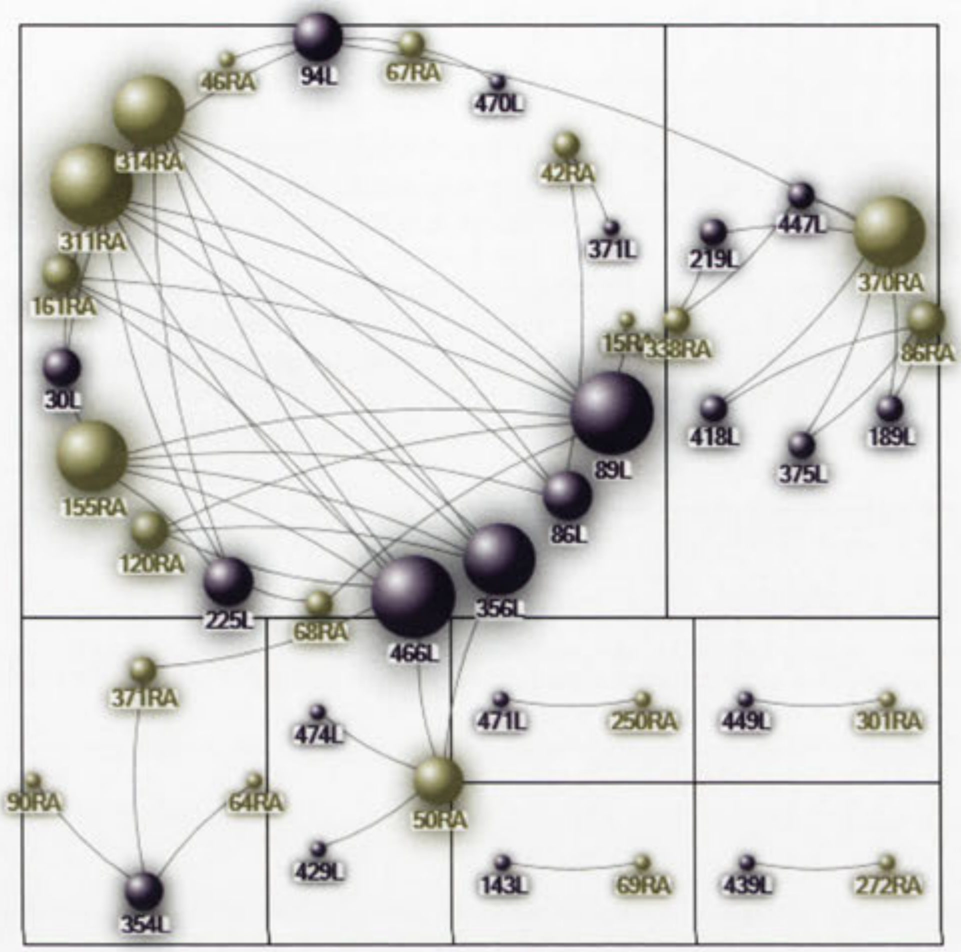


Figure 3.8C Network plot of inter-protein coevolutionary analysis of R-LSU-RA. Single-pair sites (69^{RA}, 143^{L}), (250^{RA}, 471^{L}), (272^{RA}, 439^{L}) (301^{RA}, 449^{L}), (370^{RA}, 94^{L}) and (371^{RA}, 466^{L}) can be seen as isolated single pairs. Sites 15^{RA} , 42^{RA} , 46^{RA} , 67^{RA} , 68^{RA} , 120^{RA}, 155^{RA}, 161^{RA} 311^{RA} and 314^{RA} from RA were coupled with R-LSU sites 30^{L},86^{L}, 89^{L}, 94^{L}, 225^{L}, 356^{L}, 466^{L} and 470^{L}, whereas sites 86^{RA}, 338^{RA} and 370^{RA} were found to be coupled with 189^{L}, 219^{L}, 375^{L}, 418^{L} and 447^{L}. Site 354^{L} was found to be coupled with 64^{RA}, 90^{RA} and 371^{RA}. Similarly site 50^{RA} was found to be coupled with 356^{L}, 429^{L}, 466^{L} and 474^{L}. In total, 21 sites from R-LSU are found to be coevolving with 21 sites from RA. The Girvan-Newman algorithm was used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

Sites 15^{RA}, 42^{RA}, 46^{RA}, 67^{RA}, 68^{RA}, 120^{RA}, 155^{RA}, 161^{RA} 311^{RA} and 314^{RA} from RA form a cluster with R-LSU sites 30^{L}, 86^{L}, 89^{L}, 94^{L}, 225^{L}, 356^{L}, 466^{L} and 470^{L}, whereas sites 86^{RA}, 338^{RA} and 370^{RA} were found to be coupled with 94^{L},189^{L}, 219^{L}, 375^{L}, 418^{L} and 447^{L}, mostly hydrophobic sites from R-LSU except for 94^{L}, which is charged. Additionally sites 50^{RA} with (356^{L}, 429^{L}, 466^{L}, 474^{L}) and 354^{L} with (64^{RA}, 90^{RA} and 371^{RA}) were also found to be coupled. A small number of single-pair coevolving sites were also detected (69^{RA}, 143^{L}), (250^{RA}, 471^{L}), (272^{RA}, 439^{L}) (301^{RA}, 449^{L}), (370^{RA}, 94^{L}) and (371^{RA}, 466^{L}).

3.3.3.2 Coevolution analysis of Rubisco large subunit and RbcX

The coevolution analysis of R-LSU-RbcX identified 25 sites in R-LSU to be coevolving with 26 sites in RbcX (Figures 3.9A and B). The most frequently occurring coevolving sites in R-LSU are 189^{L}, 341^{L}, 363^{L}, 375^{L}, 418^{L}, 449^{L}, 470^{L} and 471^{L}. Among these sites, except for 449^{L}, 470^{L} and 471^{L}, all sites form the hydrophobic core of R-LSU. Site 449^{L} is surface accessible while 470^{L} and 471^{L} are located in the C-terminal tail. Other coevolving sites in R-LSU included surface accessible sites (86^{L}, 91^{L}, 94^{L}, 95^{L} from the N-terminal domain, and 447^{L}, 461^{L} and 464^{L} from the C-terminal domain), sites at the inter-dimer interface (143^{L},145^{L}) and hydrophobic core sites 219^{L},340^{L},353^{L}, 359^{L} from the C-terminal domain.

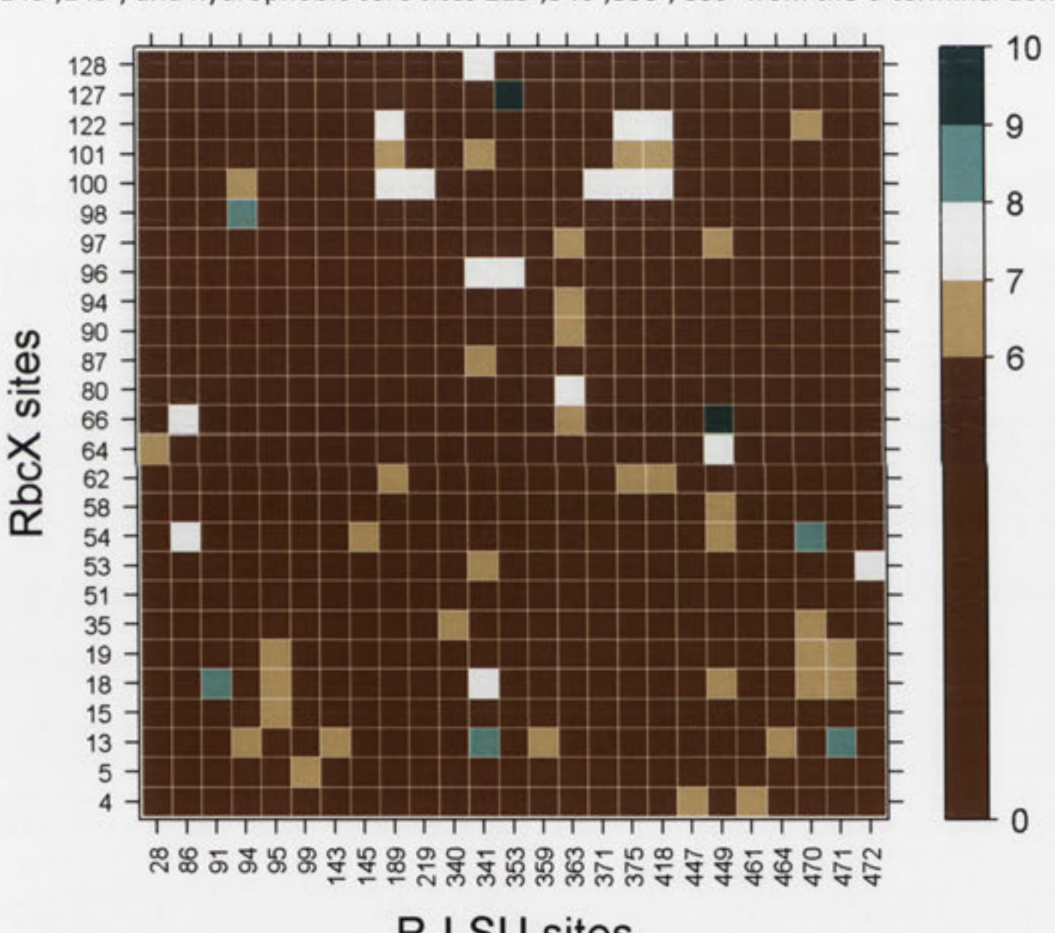


Figure 3.9A Inter-protein coevolutionary analysis of R-LSU-RbcX using NMI. The NMI *z*-scores matrix (*z*>6) of the analysis is plotted. On the colour scale, the *z*-score ranges from 0 to 10. The minimum and maximum values in the matrix are 6.0 and 9.8, respectively. The x axis shows R-LSU sites, while the y axis shows RbcX sites. Spinach sequence numbering is used for cross comparison convenience in R-LSU, but for RbcX, *Arabidopsis* sequence numbering is used, as spinach RbcX sequence is not available. In total, 25 sites from R-LSU were found to be coevolving with 26 sites from RbcX.

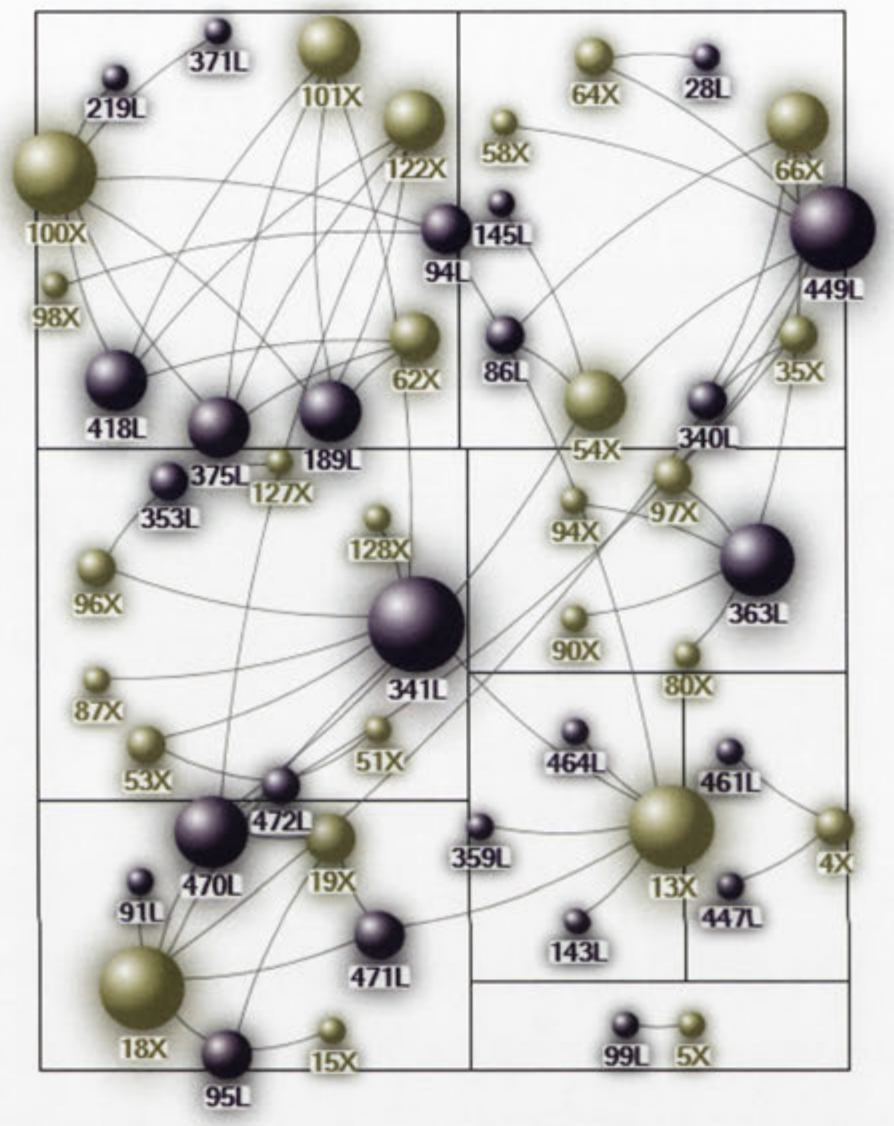


Figure 3.9B Network plot of inter-protein coevolutionary analysis of R-LSU-RbcX. Single-pair sites (28^{L}, 64^{X}), (91^{L}, 18^{X}), (99^{L}, 5^{X}), (145^{L}, 54^{X}), (219$^{L\cdot}$ 100^{X}), (359^{L}, 13^{X}), (341^{L}, 18^{X}), (341^{L}, 62^{X}), (371^{L}, 100^{X}), (447^{L}, 4^{X}), (461^{L}, 4^{X}), (464^{L}, 13^{X}), (470^{L}, 35^{X}) and (470^{L}, 54^{X}) can be seen as isolated single pairs. Many clusters were identified: cluster 1 {94^{L}, 189^{L}, 375^{L}, 418^{L}}with {62^{X}, 98^{X}, 100^{X}, 101^{X}, 122^{X}}, cluster2 {341^{L}, 472^{L}} with {51^{X}, 53^{X}, 87^{X}, 96$^{X\cdot}$ 127^{X}, 128^{X}}, cluster3 {95^{L}, 470^{L}, 471^{L}} with {15^{X}, 18^{X}, 19^{X}), cluster4 {363^{L}} with {66^{X}, 90^{X}, 94^{X}, 97^{X}}, cluster5 {449^{L}} with {18^{X}, 54^{X}, 58^{X}, 64^{X}, 66^{X}, 97^{X}}, cluster6 {13^{X}} with {94^{L}, 143^{L}, 341^{L} , 359^{L}, 464^{L}, 471^{L}} and cluster7 {4^{X}} with {447^{L}, 461^{L}}. The Girvan-Newman algorithm was used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

In total, inter-protein coevolutionary analysis of R-LSU-RbcX identified 7 clusters of coevolving sites. Amongst the 7 clusters, cluster 3 i.e. {95^{L}, 470^{L}, 471^{L}} with {15^{X}, 18^{X}, 19^{X}} is noteworthy, as it included sites 470^{L} and 471^{L} from the C-terminal tail of R-LSU, which has been reported to be the major interface between RbcX$_2$ and the R-LSU subunits

(Bracher et al., 2011). Other major clusters include numerous R-LSU sites located in regions of structural and functional importance coevolving with several RbcX sites: cluster1 {94^{L}, 189^{L}, 375^{L}, 418^{L}} with {62^{X}, 98^{X}, 100^{X}, 101^{X}, 122^{X}}; cluster2 {341^{L}, 472^{L}} with {51^{X}, 53^{X}, 87^{X}, $96^{X,}$ 127^{X}, 128^{X}}, which includes the loop 6 flanking site 341^{L} ; and, clustre5, the surface accessible R-LSU site 449^{L} with {18^{X}, 54^{X}, 58^{X}, 64^{X}, 66^{X}, 97^{X}}. Cluster4 that included RbcX sites, 66^{X}, 90^{X}, 94^{X} and 97^{X}, was found to be coevolving with site 363^{L} in R-LSU. Fourteen single-pair sites (28^{L}, 64^{X}), (91^{L}, 18^{X}), (99^{L}, 5^{X}), (145^{L}, 54^{X}), ($219^{L,}$ 100^{X}), (359^{L}, 13^{X}), (341^{L}, 18^{X}), (341^{L}, 62^{X}), (371^{L}, 100^{X}), (447^{L}, 4^{X}), (461^{L}, 4^{X}), (464^{L}, 13^{X}), (470^{L}, 35^{X}) and (470^{L}, 54^{X}) were also found to be coevolving between R-LSU and RbcX.

3.3.3.3 Coevolution analysis of R-LSU and R-SSU

The interface between R-LSU and R-SSU covers a large buried area; each small subunit is in contact with three different large subunits from two different L_2 dimers as well as with two neighbouring small subunits. The interface shows some interesting general features; although the contact area of the small subunit shows the normal distribution between non-polar, polar and charged atoms (Janin et al., 1988), the corresponding areas from the large subunits are enriched in charged and polar atoms (Knight et al., 1990). In my analysis, the main aim was to understand the rules of coevolutionary dynamics between the two types of subunit.

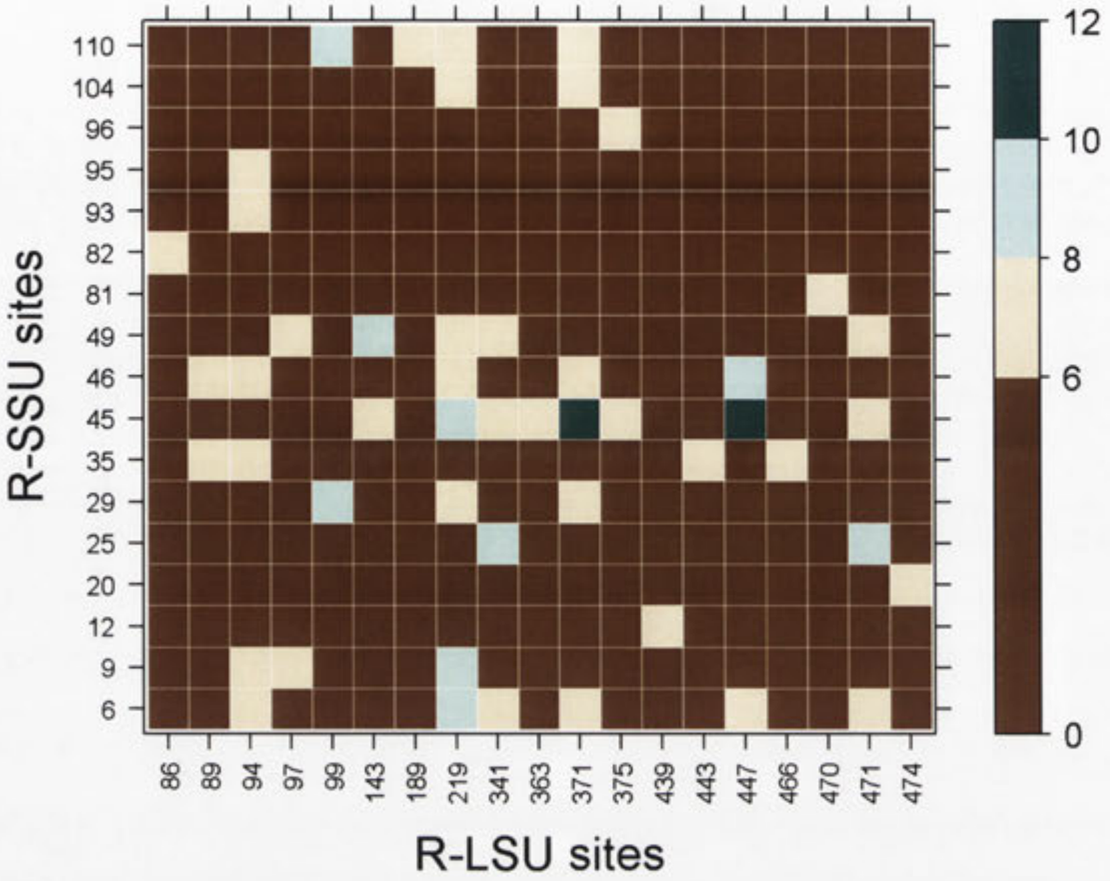


Figure 3.10A Inter-protein coevolutionary analysis of R-LSU-R-SSU using NMI. The NMI *z*-scores matrix (*z*>6) of analysis is plotted. On the colour scale, the z-score ranges from 0 to 12. The minimum and maximum values in the matrix are 6.0 and 10.7, respectively. The x axis shows R-LSU sites, while the y axis shows R-SSU sites. Spinach sequence numbering is used for cross comparison convenience in both R-LSU and R-SSU. In total, 19 sites from R-LSU were found to be coevolving with 17 sites from R-SSU.

The analysis identified 19 sites in the R-LSU to be coevolving with 17 sites in the R-SSU (Figure 3.10A). Most of the coevolving sites in the R-LSU are located in the C-terminal domain, the most notable being 219^L, 341^L, 371^L and 471^L, as site 219^L is known be on the R-LSU-R-SSU interface (Knight et al., 1990), site 471^L is part of the C-terminal tail, whereas sites 341^L and 371^L are part of the hydrophobic core of the C-terminal domain.

In total, inter-protein coevolutionary analysis of R-LSU-R-SSU identified 8 clusters of coevolving sites (Figure 3.10B). R-LSU sites 219^L, 371^L and 447^L formed a major cluster (cluster1) with R-SSU sites 6^S, 29^S, 45^S, 46^S and 104^S. Also R-LSU sites 341^L and 471^L were found to be coevolving with 6^S, 25^S, 45^S and 49^S from R-SSU.

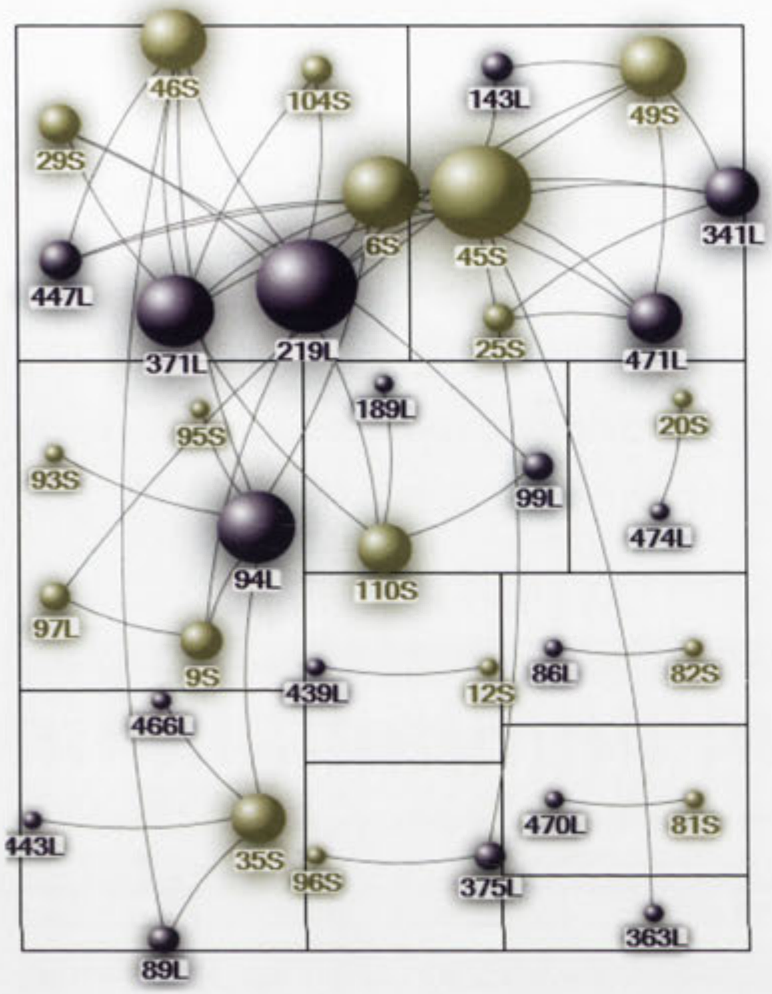


Figure 3.10B Network plot of inter-protein coevolutionary analysis of R-LSU-R-SSU. Single-pair sites (9^S, 219^L), (12^S, 439^L), (20^S, 474^L), (29^S, 99^L), (49^S, 97^L), (45^S, 363^L), (81^S, 470^L), (82^S, 86^L) and (96^S, 375^L) can be seen as isolated single pairs. Many clusters were identified: cluster1 {219^L, 371^L, 447^L} with {6^S, 29^S, 45^S, 46^S, 104^S}; cluster2 {94^L, 97^L} with {9^S, 93^S, 95^S}; cluster3 {35^S} with {89^L, 443^L, 466^L}; cluster4 {143^L, 341^L, 471^L} with {25^S, 45^S and 49^S}; cluster5 {110^S} with {99^L, 189^L, 219^L, 371^L}; cluster6 {375^L} with {45^S, 96^S}; cluster7 {94^L} with {6^S, 35^S, 46^S} and cluster8 {6^S} with {341^L, 471^L}. The Girvan-Newman algorithm was used for cluster detection. As evident in figure, the size of circle scales with number of interactions.

Several N-terminal domain coevolving sites (86^L, 89^L, 94^L, 97^L, 99^L and 143^L) were detected in the R-LSU. Site 94^L was found to be most prominent, coevolving with R-SSU sites 6^S, 9^S, 35^S, 46^S, 93^S and 95^S. In the R-SSU, sites 6^S, 9^S, 35^S, 45^S, 46^S, 49^S, 104^S and 110^S were identified as the most frequent coevolving sites. R-SSU sites 45^S, 46^S and 49^S are close to the hairpin loop which shapes the surface of the central solvent channel. Nine single-site pairs (9^S, 219^L), (12^S, 439^L), (20^S, 474^L), (29^S, 99^L), (45^S, 363^L), (81^S, 470^L), (82^S, 86^L) and (96^S, 375^L) were found to be coevolving between R-LSU and R-SSU.

3.3.4 Summary of Intra/Inter-protein coevolutionary analysis of R-LSU

I have summarized the result of both intra and inter-protein coevolutionary analysis of R-LSU in Table 3.4 for Discussion.

Table 3.4 Summary of single pair sites and clusters found in Coevolution Analyses

Source	Analysis	Single pair sites	Clusters
All-Angiosperm R-LSU sequences	Intra	(95, 97), (247, 282) and (439, 466) and (466, 468)	All-Angiosperm _cluster {91, 341, 363, 371, 464, 471, 472, 474}
Solanales	Intra	(142,251), (143, 353), (251, 255) and (440,443)	Solanales_cluster1 {86, 89, 91, 94, 95, 356, 447, 466, 470, 471 and 472} and Solanales_cluster2 {30, 309, 328, 340, 429, 468 and 474}
Poales	Intra	(97,359), (145,270), (219,359), (219,418) and (251,255)	Poales_cluster1 {91, 94, 95, 99, 219, 341, 418, 446, 470, 474} and Poales_cluster2 {28, 89, 143, 157, 189, 247, 282, 353, 447 and 449}
Gentianales	Intra	(251, 255), (375,398) and (470,474)	Gentianales_cluster {28, 91, 95, 340, 429, 439, 466,468, 470 and 472}

Caryophyllales	Intra	(31, 91), (93,149), (149, 475) and (309,328)	Caryophyllales_cluster1{30, 50, 88, 89, 94, 353, 356, 358, 359, 442 and 472},
			Caryophyllales_cluster2{32, 99, 142, 145, 354, 367, 371 and 443}
			Caryophyllales_cluster3{34, 226, 230, 375 and 447}

R-LSU-RA	Inter	(69^{RA}, 143^{L}), (250^{RA}, 471^{L}), (272^{RA}, 439^{L}) (301^{RA}, 449^{L}), (370^{RA}, 94^{L}) and (371^{RA}, 466^{L})	R-LSU-RA_cluster1 {15^{RA}, 42^{RA}, 46^{RA}, 67^{RA}, 68^{RA}, 120^{RA}, 155^{RA}, 161^{RA} 311^{RA}, 314^{RA}} with {30^{L},86^{L}, 89^{L}, 94^{L}, 225^{L}, 356^{L},466^{L}}
			R-LSU-RA_cluster2 {86^{RA}, 370^{RA} } with {189^{L}, 219^{L}, 375^{L}, 418^{L}, 447^{L}}
			R-LSU-RA_cluster3 {354^{L}} with {64^{RA}, 90^{RA}, 371^{RA}}
			R-LSU-RA_cluster4 {50^{RA}} with {356^{L}, 429^{L}, 466^{L} and 474^{L}}

R-LSU-RbcX	Inter	(28^{L}, 64^{X}), (91^{L}, 18^{X}), (99^{L}, 5^{X}), (145^{L}, 54^{X}), (219^{L} 100^{X}), (359^{L}, 13^{X}), (341^{L}, 18^{X}), (341^{L}, 62^{X}), (371^{L}, 100^{X}), (447^{L}, 4^{X}), (461^{L}, 4^{X}), (464^{L}, 13^{X}), (470^{L}, 35^{X}) and (470^{L}, 54^{X})	R-LSU-RbcX _cluster 1 {94^{L}, 189^{L}, 375^{L}, 418^{L}} with {62^{X}, 98^{X}, 100^{X}, 101^{X}, 122^{X}}
			R-LSU-RbcX _cluster2 {341^{L}, 472^{L}} with {51^{X}, 53^{X}, 87^{X}, 96^{X} 127^{X}, 128^{X}}
			R-LSU-RbcX _cluster3 {95^{L}, 470^{L}, 471^{L}} with {15^{X}, 18^{X}, 19^{X})
			R-LSU-RbcX _cluster4 {363^{L}} with {66^{X}, 90^{X}, 94^{X}, 97^{X}}
			R-LSU-RbcX _cluster5 {449^{L}} with {18^{X}, 54^{X}, 58^{X}, 64^{X}, 66^{X}, 97^{X}}
			R-LSU-RbcX _cluster6 {13^{X}} with {94^{L}, 143^{L}, 341^{L} , 359^{L}, 464^{L}, 471^{L}}
			R-LSU-RbcX _cluster7 {4^{X}} with {447^{L}, 461^{L}}

R-LSU-R-SSU	Inter	(9^{S}, 219^{L}), (12^{S}, 439^{L}), (20^{S}, 474^{L}), (29^{S}, 99^{L}), (49^{S}, 97^{L}), (45^{S}, 363^{L}), (81^{S}, 470^{L}), (82^{S}, 86^{L}) and (96^{S}, 375^{L})	R-LSU-R-SSU_cluster1 {219^{L}, 371^{L}, 447^{L}} with {6^{S}, 29^{S}, 45^{S}, 46^{S}, 104^{S}}
			R-LSU-R-SSU_cluster2 {94^{L}, 97^{L}} with {9^{S}, 93^{S}, 95^{S}}
			R-LSU-R-SSU_cluster3 {35^{S}} with {89^{L}, 443^{L}, 466^{L}}
			R-LSU-R-SSU_cluster4 {143^{L}, 341^{L}, 471^{L}} with {25^{S}, 45^{S} and 49^{S}}
			R-LSU-R-SSU_cluster5 {110^{S}} with {99^{L}, 189^{L}, 219^{L}, 371^{L}}; cluster6 {375^{L}} with {45^{S}, 96^{S}}
			R-LSU-R-SSU_cluster7 {94^{L}} with {6^{S}, 35^{S}, 46^{S}}
			R-LSU-R-SSU_cluster8 {6^{S}} with {341^{L}, 471^{L}}

3.4 Discussion

3.4.1 A large number of coevolving sites in Rubisco are found in clusters

The coevolution analysis of R-LSU sequences revealed two broad groups of coevolving sites as summarized in Table 3.4. One group contains sites that coevolve with only one or two other sites. These positions tend to be spatially close and display high probability of direct amino acid side-chain interactions with their coevolving partner. The other group included positions that coevolve with many others. They are often found in regions crucial for Rubisco function, such as areas surrounding the active-site and surfaces involved in intermolecular interactions and recognition. Most coevolving sites found in the analysis showed a tendency to participate in a cluster/network of coevolving residues. Many such clusters were identified. The cluster of residues identified in All-Angiosperm _cluster in Table 3.4 {91^L, 341^L, 363^L, 371^L, 464^L, 471^L, 472^L and 474^L} are spatially proximal to loop 6 and the C-terminal tail, and may have an indirect role in conformational changes required to release inhibitors from the Rubisco active site. Clusters (Table 3.4) detected in Solanales_cluster1 {86, 89, 91, 94, 95, 356, 447, 466, 470, 471 and 472}, Gentianales_cluster {28, 91, 95, 340, 429, 439, 466,468, 470 and 472} and Poales_cluster1 {91, 94, 95, 99, 219, 341, 418, 446, 470, 474} could be part of the activase recognition region in Rubisco. These observations are consistent with findings of Gloor et al. (2005), who documented similar patterns of coevolving sites in many families of proteins.

3.4.2 Network of coevolving sites flanking loop 6 of R-LSU

As noted above, the coevolution analysis of the all-Angiosperm plant R-LSU sequences uncovered one major cluster of coevolving sites, i.e. All-Angiosperm_cluster as shown in Table 3.4. This cluster includes 7 sites, 341^L, 363^L, 371^L, 464^L, 471^L, 472^L and 474^L from the C-terminal domain. Site 341^L is spatially proximal to both loop 6 and the C-terminal tail (within 4Å). Mutations in and around loop 6 have been studied extensively. Several investigators have changed the *Synechococcus* α-helix 6 sequence $D^LK^LA^LS^L$ (residues 338^L–341^L) to the $E^LR^LE^LI^L$ or $E^LR^LD^LI^L$ sequence characteristic of land plants (Gutteridge et al., 1993, Kane et al., 1994, Parry et al., 1992). Mutations were also made in

Chlamydomonas (Leu-326^L, Val-341^L, Met-349^L) to imitate the land-plant loop 6 (Ile-326^L,Ile-341^L, Leu-349^L) (Zhu and Spreitzer, 1996). All of these studies reported that these mutations impaired the holoenzyme stability and/or catalytic properties of Rubisco. One inference of these outcomes could be that sites in the loop-6 region are coevolving with additional sites outside of this loop and any mutation in this region needs to be complemented, appropriately.

The cluster also includes 4 sites (464^L, 471^L, 472^L and 474^L) from the C-terminal tail of Rubisco. In the crystal structure the C-terminal tail packs on top of loop 6 with numerous hydrogen bonds and ion-pair interactions keeping it fixed. This has been interpreted as acting as a "bolt" that locks the initially flexible loop 6 in position (Andersson et al., 1989, Knight et al., 1990, Curmi et al., 1992) . Any mutation in the C-terminal tail could disturb this network of hydrogen bonds and ion-pair interactions.

Thus, overall the results of this analysis suggest that mutations in loop 6 or the C-terminal tail need to be complemented by other mutations from this cluster of residues (341^L, 363^L, 371^L, 464^L, 471^L, 472^L and 474^L) of R-LSU to maintain the coevolutionary dynamics among these sites. To summarize, the network of coevolving sites discovered in this study points to complementary changes required to maintain the catalytic efficiency and specificity of Rubisco, in case of mutations in and around loop 6 and the C-terminal tail.

3.4.3 Coevolving sites as potential targets of RA

3.4.3.1 R-LSU-RA inter-protein analysis identified coevolving sites in activase recognition region of R-LSU

The highlight of the inter-protein coevolution analysis between R-LSU and RA is the detection of a number of coevolving charged sites on the outer surface of the R-LSU. These charged residues are on one face of the solvent accessible surface (Figure 3.8B) of the R-LSU, with side chains protruding outwards, making them a potential target for interaction with RA.

As a proof of concept, the analysis detected a strong coevolutionary signal between positions 89^{L} and 94^{L} from R-LSU and 311^{RA} and 314^{RA} from RA as experimentally shown by Li et al. (2005) as evident in R-LSU-RA_cluster1 in Table 3.4. In the N-terminal domain of R-LSU, side chains of coevolving sites 30^{L}, 89^{L} and 94^{L} are highly surface accessible and sites 30^{L} and 94^{L} are also charged. These sites are thus likely targets for protein interactions with the N-terminal domain of R-LSU. The side chain of site 30^{L} is fully exposed and its neighboring region is packed with totally conserved negatively charged residues (Asp/Glu 28^{L}, Asp 33^{L} and Asp 35^{L}). This region of the N-terminal domain could act as a "sticky recognition spot" for protein-protein interactions. The significance of electrostatic contributions of charge-charge interactions in protein-protein interactions are well documented in the literature (Sheinerman et al., 2000, Sinha and Smith-Gill, 2002, Keskin et al., 2005). These finding suggest that site 30^{L} of R-LSU could well be one of the anchor residues for RA interaction in N-terminal domain in addition to sites 89^{L} and 94^{L}.

Also C-terminal sites 356^{L}, 429^{L}, 439^{L}, 447^{L} and 449^{L} form a network of charged/polar sites on the solvent accessible surface of the R-LSU. Specifically, site Arg 439^{L} is fully exposed with a protruding side chain. It is also surrounded by conserved positively charged sites Arg 431^{L}, Arg 435^{L} and Arg 446^{L}. Arg 439^{L} could be the anchor in the C-terminal domain for RA interaction, supported by other coevolving sites found in this analysis. Site Lys 356^{L} also has high solvent accessibility and is located among a series of charged residues (Asp 351^{L}, Asp 352^{L}, Glu 355^{L}, Asp 357^{L}, Arg 359^{L} and Arg 360^{L}). Thus in summary, the analysis of coevolving sites revealed a highly charged region with high solvent accessibility in the R-LSU C-terminal domain that might act as a potential interface for RA interaction. Interestingly, mutagenesis studies in *Chlamydomonas* found evidence only for sites 89^{L} and 94^{L} to be important for activase interaction, whereas mutation in site 356^{L} (and 86^{L}) had little effect on the relative abilities of spinach and tobacco activase to activate the mutant Rubiscos (Larson et al., 1997, Ott et al., 2000).

The surface-accessible polar/charged sites (466^{L}, 470^{L}, 471^{L} and 474^{L}) form part of the C-terminal peptide tail of the R-LSU. This region has recently been shown to be the primary mode of engagement (Mueller-Cajar et al., 2011) between CbbX (red-type

Rubisco activase) and Rubisco holoenzyme, in red- type Rubisco. In RA, clusters of charged/polar (50^{RA}, 86^{RA}, 155^{RA}, 161^{RA}, 311^{RA}) and hydrophobic (120^{RA}, 314^{RA}, 370^{RA}) sites have been identified in the analysis. Of these sites, only 311^{RA} and 314^{RA} have been implicated so far in direct physical contact with the R-LSU (Li et al., 2005). Also, removal of the C-terminal extension in RA has been reported to cause the loss of the ATPase and activase functions (Stotz et al., 2011), indicating that site 370^{RA} may have a role in R-LSU-RA interaction.

3.4.3.2 Cluster of coevolving sites in Solanales

The coevolution analysis of Solanales R-LSU revealed two clusters of 11 and 7 sites as shown in Table 3.4. Most of the sites identified in the R-LSU-RA inter-protein analysis also showed up as coevolving within the R-LSU in Solanales, as compared with the all-Angiosperm background dataset. Compared with the R-LSU-RA analysis, sites 429^L, 439^L, 447^L and 449^L were not identified in Solanales R-LSU coevolution analysis, but it identified one additional site 468^L in the C-terminal domain. The mostly similar results in R-LSU-RA inter-protein analysis and Solanales intra-protein analysis suggest that these sites in R-LSU coevolve as part of evolving inter-protein interactions with RA.

3.4.3.3 Cluster of coevolving sites in Gentianales

The coevolution analysis of Gentianales also generated some interesting patterns. Ten coevolving sites 28^L, 91^L (Pro), 95^L (Asp), 340^L, 429^L, 439^L (Val/Ala), 466^L (Arg), 468^L (Asn), 470^L (Lys) and 472^L were found as shown in Gentianales_cluster in Table 3.4. The cluster of identified coevolving sites is similar to that for Solanales, except that sites 89^L and 94^L were not detected; also two additional sites 28^L and 439^L were identified. Coevolving sites in the C-terminal tail, 466^L, 468^L and 470^L, are all even-numbered, solvent exposed residues, whereas the odd-numbered, buried residues are totally conserved. As compared with the all-Angiosperm background dataset, Gentianales site 95^L acquires a negatively charged residue Asp, whereas site 439^L loses a positive charged residue, Arg, as for Solanales. Site 91^L differs from Solanales by recruiting Ala instead of Pro. Most of the coevolving sites identified in Gentinales_cluster are located in C-terminal tail, i.e. 466^L, 468^L, 470^L and 472^L along with two sites (91 and 95) in βC-βD loop in N-terminal domain.

Spatial location of these sites in R-LSU suggest that these R-LSU sites may be coevolving due to evolving inter-protein interactions with RA, as both of these structural regions are implicated in activase recognition in R-LSU (Andersson and Backlund, 2008, Mueller-Cajar et al., 2011). Unfortunately no RA sequence from Gentinales is available to examine the validity of these findings.

3.4.3.4 Cluster of coevolving sites in Poales

The first cluster of coevolving sites in Poales, 91^L, 94^L, 95^L and 99^L also flanks the activase recognition region in the N-terminal domain. The coevolving-residue set in Poales differs from that in Solanales (Lys 94^L and Asp 95^L) by exhibiting significant variability in sites 94^L (Asp/Glu/Pro/Ala) and 95^L (Asn/Ser/Asp). It appears that the R-LSU requires at least one negative charge at either site 94^L or site 95^L; 95^L is Asp whenever site 94^L is Pro/Ala, otherwise it is Asn/Ser. Schreuder et al. (1993) noted that Lys 94^L interacts with the side chain of Glu 93^L in the tobacco Rubisco x-ray structure, whereas the Glu 94^L side chain points in the opposite direction in the spinach Rubsisco x-ray structure. This is also the case in the rice Rubisco x-ray structure (PDB id 1WDD). This difference in side-chain direction (note the difference might be an artifact of crystal structure, i.e. may or may not exist in solution) at the site 94^L of Solanales, may be one of the reasons for differential structure specificities of Solanales and non-Solanales Rubisco activases.

3.4.3.5 Cluster of coevolving sites in Caryophyllales

Coevolution analysis of Caryophyllalels produced striking results with a high number of coevolving sites (31) as well as 3 clusters of coevolving sites (Table 3.4). As noted earlier, in terms of R-LSU sequence conservation, Caryophyllales is highly diverse; this is reflected in the outcomes of the coevolution analysis. Five sites from Caryophyllales_cluster1, i.e. 30, 89, 94, 356 and 472 could be part of activase recognition region, as noted in previous sections. Interestingly as noted in Results (section 3.3.2.4), several coevolving sites (31^L, 32^L, 34^L, 88^L, 358^L, 359^L and 442^L) identified in this analysis are specific to Caryophyllales; they were not found in the analysis for the other orders or all-Angiosperm analysis. As these sites are highly conserved in all the other plant orders, it appears that these variations are clade specific. Thus, some part of the coevolution signal observed in Caryophyllales may reflect shared ancestry. As a result, some of the coevolution signal observed in this study could be attributed to phylogenetic relationships within the order and may not reflect true coevolutionary processes.

Overall, coevolution analysis of R-LSU both within R-LSU in 4 plant orders and with RA revealed several novel coevolving sites. As discussed in section 3.4.3.1, coevolution analysis of R-LSU and RA identified many coevolving sites in both the N-terminal and C-terminal domain of R-LSU, most of which are located in solvent accessible charged surfaces of the R-LSU, hence making a strong case for these sites being the mediator of interaction between R-LSU and RA. Intra-protein analyses of Solanales (section 3.4.3.2), Gentianales (section 3.4.3.3), Poales (section 3.4.3.4) and Caryophyllales (section 3.4.3.5) identified coevolving sites are also located in the same activase-recognition regions. These findings are consistent with work of Pazos et al. (1997) who observed that analysis of coevolution within a protein can detect coevolution traces of protein-protein interactions. Moreover, some of the coevolution signal observed in Caryophyllales appears to be clade specific and could be attributed to phylogenetic noise.

3.4.4 R-LSU has highly conserved interaction interfaces with R-SSU and RbcX

The interaction regions of R-LSU with RbcX and R-SSU are very well defined. A recent study by (Bracher et al., 2011) reported the crystal structure of the RbcX-bound assembly intermediate of form I Rubisco, whereas the crystal structure of Rubisco holoenzyme (L_8S_8) was solved a long time ago (Andersson et al., 1989). Therefore, the interaction interfaces of R-LSU with RbcX and R-SSU have been studied in some detail and they seem to be fairly well conserved.

Bracher et al.(2011) solved the x-ray structure of *Synechococcus*6301 (R-LSU)$_8$-*Anabaena* sp. (RbcX$_2$)$_8$ complex (PDB id 3RG6) and identified three contact areas in the R-LSU for interaction with RbcX$_2$. Area I comprises the C-terminal peptide of the R-LSU (458^L–468^L), area II includes residues Leu332^L and Glu333^L and area III is the RbcX$_2$ interface with the adjacent R-LSU subunit of the R-LSU dimer (42^L–46^L, 49^L and 53^L) and residues 123^L–126^L. All these residues are highly conserved in sequences of form I R-LSU subunits, with the few exceptions being at the carboxy-terminus of the R-LSU. Altogether, only four coevolving sites 461^L, 464^L, 470^L and 471^L out of 23 sites identified in the coevolution analysis flank the contact area I of the R-LSU-$RbcX_2$ interface. In RbcX also most of the residues at the R-LSU-RbcX interface are highly conserved in the RbcX dataset and only one site 58^X was found to be coupled with surface-accessible residue 449^L in R-LSU-RbcX _cluster5 (Table 3.4). It should be noted that only 14 RbcX sequences were used in this analysis, due to limited availability of RbcX sequences in the public databases. A larger sequence set is necessary to increase the level of confidence for the prediction of coevolution sites between R-LSU and RbcX.

As noted previously, in the Rubisco holoenzyme, each small subunit is in contact with three different large subunits from two different L_2 dimers as well as with two neighboring small subunits. The R-LSU-R-SSU interface involves 49 residues from the R-LSU; virtually all of them are totally conserved (over 99% conservation) in Angiosperms. Of the few exceptions (76^L, 219^L, 226^L, 230^L, 429^L), the R-LSU-R-SSU coevolution analysis identified one of these sites, 219^L as coevolving with several sites from the R-SSU. Site 219^L along with sites 371^L and 447^L in the R-LSU were found to form a coevolving cluster with 5 SSU sites {6^S, 29^S, 45^S, 46^S, 104^S} in R-LSU-R-SSU_cluster1 as shown in Table 3.4. Several coevolving R-SSU sites, i.e. 45^S, 46^S and 49^S, are spatially close and part of a long hairpin loop (46^S to 67^S) which join strands β-A and β-B of the SSU and protrudes into the central solvent channel of the LSU. These coevolving sites can contribute to the hydrogen-bond network within the loop. The absence of this loop in the small subunit of cyanobacterial Rubisco (Knight et al., 1990), has generated a lot of interest in examining the contribution of these residues to R-LSU-R-SSU interactions in higher plant L_8S_8 molecules.

Interestingly, a large number of R-LSU coevolving sites identified in this analysis are located in the N-terminal domain (14^L, 86^L, 89^L, 94^L, 97^L, 99^L, 143^L), hydrophobic core (341^L, 359^L and 371^L) and C-terminal tail (466^L, 468^L, 470^L and 474^L). These regions of the large subunit are not spatially proximal to R-SSU interface regions of R-LSU; the origin/significance/reliability of this finding is unclear.

In summary, the results of inter-protein coevolution analysis of RbcX and R-SSU with R-LSU are consistent with experimental observations and also uncovered a few true positive coevolving sites (sites 461^L, 464^L, 470^L & 471^L with RbcX and site 219^L with R-SSU) in the R-LSU. As discussed above, this is not unexpected due to the presence of highly conserved residues at the known interaction interfaces.

3.5 Conclusion

In summary, coevolution analysis of the R-LSU and with its interacting partners has produced some interesting results. The All-Angiosperm _cluster {91^L, 341^L, 363^L, 371^L, 464^L, 471^L, 472^L and 474^L} is one of the most significant findings of the intra-protein analyses as coevolving sites identified in this cluster are located in known regions of functional and structural importance of R-LSU. Furthermore, many novel coevolving sites in the RA-interaction region of R-LSU were identified in Solanales_cluster1 {86, 89, 91, 94, 95, 356, 447, 466, 470, 471 and 472}, Gentianales_cluster {28, 91, 95, 340, 429, 439, 466,468, 470 and 472} and Poales_cluster1 {91, 94, 95, 99, 219, 341, 418, 446, 470, 474}. The identification of many novel coevolving sites (30^L, 429^L, 439^L, 447^L, 449^L, 466^L, 470^L, 471^L and 474^L) on the outer surface of R-LSU in the R-LSU-RA clusters is the highlight of the inter-protein coevolution analyses. The R-LSU-RbcX and R-LSU-R-SSU inter-protein analyses have resulted in a few true positive identifications because of the highly conserved binding interfaces.

4 Codon-usage analysis of *rbcL*

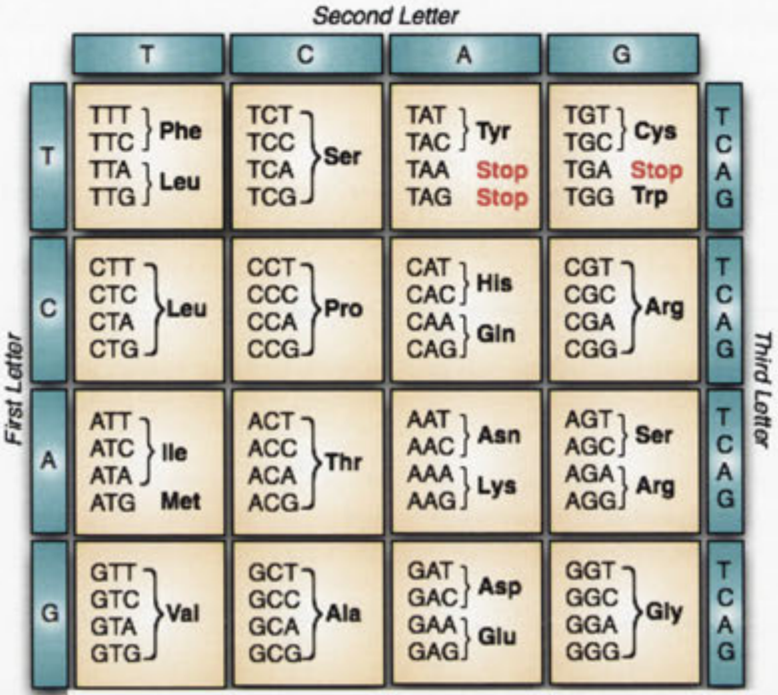
4.1 Background

In my thesis, I have performed wide-ranging computational studies to understand the functional significance of sequence variations in Angiosperm Rubisco-LSU sequences both at the protein and nucleotide levels. In the current chapter, I investigated codon-usage bias of the *rbcL* gene that encodes Rubisco-LSU to analyze the relationship between synonymous variations in *rbcL* with Rubisco's 3D structure. Furthermore, I consolidated tRNA and codon-usage data for all available Angiosperm chloroplast genomes in the public domain to examine the role of selection in shaping the codon-usage bias of *rbcL* and differences in codon-usage pattern of *rbcL* with other protein-coding genes in chloroplast genomes.

4.1.1 What is codon bias?

Codon-usage bias is a pattern of differential usage of codons for a particular amino acid, relative to codon frequencies expected by the degeneracy of the genetic code. Non-uniform use of synonymous codons is a general characteristic of coding sequences (Sharp and Li, 1986). It has been observed in almost every organism studied, both unicellular and multicellular (Grantham et al., 1986). Apart from Methionine and Tryptophan, all amino acids have codon redundancy that leads to the same amino acid when translated into the protein.

Amino acids can be categorized by their codon degeneracy. Because of the design of the genetic code, each amino acid (other than Met/Trp) has n synonymous codons that code for the same amino acid; there are 2-fold, 3-fold, 4-fold, and 6-fold classes (Figure 4.1). For 2-fold, 3-fold and 4-fold degenerate codons, the codons differ only at the third nucleotide position. All 6-fold degenerate amino acids (Leu, Ser, and Arg) can be further classified into a 4-fold degenerate group and a 2-fold degenerate group. Within each group codons vary at the third position nucleotide but between the 4-fold group and the synonymous 2-fold group they differ from each other at the first and/or second nucleotide. A consequence of this code structure is that most degeneracy occurs at the third nucleotide position.



	Second Letter				
First Letter	T	C	A	G	Third Letter
T	TTT TTC Phe TTA TTG Leu	TCT TCC TCA TCG Ser	TAT TAC Tyr TAA Stop TAG Stop	TGT TGC Cys TGA Stop TGG Trp	T C A G
C	CTT CTC CTA CTG Leu	CCT CCC CCA CCG Pro	CAT CAC His CAA CAG Gln	CGT CGC CGA CGG Arg	T C A G
A	ATT ATC ATA Ile ATG Met	ACT ACC ACA ACG Thr	AAT AAC Asn AAA AAG Lys	AGT AGC Ser AGA AGG Arg	T C A G
G	GTT GTC GTA GTG Val	GCT GCC GCA GCG Ala	GAT GAC Asp GAA GAG Glu	GGT GGC GGA GGG Gly	T C A G

Source: http://freethoughtlebanon.net/2011/12/mutaaion/

Figure 4.1 Codon table showing the different codons and their corresponding amino acid.

4.1.2 Instances of codon bias

Codon bias has been observed in bacteria, plants, yeast, fly, worm, and even mammals (Ikemura, 1981, Sharp et al., 1986, Akashi and Eyre-Walker, 1998a, Duret, 2002, Urrutia and Hurst, 2003, Comeron, 2004, Wright et al., 2004, Lavner and Kotlar, 2005). Evidence supporting codon adaptation in highly expressed genes has been found in several unicellular organisms (Ikemura, 1985, Sharp, 1991), *Drosophila* (Akashi, 1994, 1995), and plastid genomes (Morton, 1993, 1998, 2000).

4.1.3 Reasons for codon bias

Codon bias has been investigated extensively, because of its presumed connection between patterns of genome organization and gene and protein evolution. Generally, codon bias is believed to be the result of interplay of two forces, genome compositional bias (Grantham et al., 1980, 1981, 1986) and selection between synonymous codons for translational efficiency (Li, 1987, Akashi and Eyre-Walker, 1998b, Duret and Mouchiroud, 1999). There has been much interest in determining the relative contribution of these two forces in influencing codon bias.

4.1.3.1 Genome compositional bias

The genome hypothesis (Grantham et al., 1980, 1981, 1986) proposes that each genome has a strategy of codon use that is followed by all of its genes. This similarity in codon use within a genome has been shown for many species (Wada et al., 1990, Sharp and Li, 1986, Grantham et al., 1986). Bernardi and Bernardi (1986) have extended the hypothesis by suggesting that each genome (or compartment) has a "genome phenotype" resulting from compositional constraints acting on both coding and non-coding sequences. Constraints such as chromosome structure and CpG levels act on the genome as a unit to affect G + C composition mainly through selective fixation as opposed to random drift (Bernardi, 1986). Codon use by the genome, or compartment, is a result of these compositional constraints acting at the genome level (Bernardi, 1986).

4.1.3.2 Selection between synonymous codons for translational efficiency

There is now strong evidence in certain species that codon bias is a result of selection between synonymous codons due to differences in translation efficiency (Ikemura, 1985, Sharp, 1991, Akashi, 1995, Morton, 1998, 2000). Selection for translational efficiency may reflect selection for rapid translation (speed selection), selection for translation with high fidelity (accuracy selection), or both (Zhou et al., 2009). It has been shown that highly expressed genes of many organisms have a bias toward "major" codons (selection for rapid translation) that are complementary to abundant tRNAs (Ikemura, 1985, Andersson and Kurland, 1990, Bulmer, 1991). Akashi (1994) argued that selection for translational accuracy should lead to inhomogeneous codon-usage within genes. More important sites i.e., sites that are less robust to translation errors, should be encoded more frequently by codons with high fidelity than other sites; he found such a signal in *Drosophila*. Subsequently, similar signals were discovered in *Escherichia coli*, yeast, worm, and mammals (Stoletzki, 2008, Drummond and Wilke, 2008).

4.1.3.3 Codon bias model

The basic model for the way genome compositional bias (G_{CB}) and selection between synonymous codons for translational efficiency (S_T) are commonly thought to generate codon bias is simply (Morton, 2001)

$$G_{CB} + S_T \text{ -> codon bias}$$

However, there is a large and growing body of experimental evidence that suggests the possibility of a third force, the role of synonymous codons within the context of protein folding and function. A silent nucleotide polymorphism in the MDR1 gene leads to the synthesis of protein product with the same amino acid sequence but different structural and functional properties (Kimchi-Sarfaty et al., 2007). A link between synonymous codon-usage, protein production and protein structure has also been proposed (Thanaraj and Argos, 1996, Biro, 2006, Zhou et al., 2009). Numerous experiments have indicated that the speed and timing of translation may be critical to the formation of a protein's native structure (Komar et al., 1999, Kepes, 1996, Kim et al., 1991, Zama, 1995). *In vitro* experiments have shown that synonymous codon mutations can have a subtle but crucial effect on protein structure and/or function (Zhang et al., 2009, Hamano et al., 2007, Kimchi-Sarfaty et al., 2007, Komar, 2007, Cortazzo et al., 2002). Computational studies have found that synonymous codons have different secondary structure propensities in many species and this structural information seems to be species specific (Adzhubei et al., 1996, Murzin et al., 1995, Gu et al., 2003, Xie and Ding, 1998, Gupta et al., 2000). Furthermore, Zhou et al. (2009) linked optimal codons, those with near maximal translation speed, to buried residues.

Thus, there is a wealth of structural, biochemical, biophysical, and computational evidence that supports the critical role of synonymous codons within the context of protein structure/function. Therefore, the actual forces that interact to generate codon bias should be represented as

$$G_{CB} + S_T + S_P \rightarrow \text{codon bias}$$

where, S_P is defined as the role of synonymous codons within the context of protein structure.

4.1.4 Codon bias in *rbcL*

The chloroplast gene *rbcL* encodes the large subunit of Rubisco (Ribulose 1, 5-bisphosphate carboxlyase), an enzyme central to photosynthesis. The chloroplasts of plants and unicellular photosynthetic organisms contain a genome that codes for a fairly conserved set of fewer than 100 genes, most of which are involved in protein synthesis and photosynthesis. Genes of the plant chloroplast genome have a codon bias that appears to be the result of a strong compositional bias toward a high genomic A+T content, as synonymous codons with A or T at the third position are highly represented (Wolfe and Sharp, 1988).

The high functional significance and low rate of sequence divergence in *rbcL* have led authors to argue that *rbcL* does not show a codon-bias pattern reflective of mutational selection but, rather, one that reflects the low G+C content characteristic of the chloroplast genome (Albert et al., 1994, Morton, 1994, Morton and Levin, 1997). Wall and Herbeck (2003) concluded that codon bias in *rbcL* is heavily affected by background mutational biases and genetic drift. They also found evidence of weak selection in codon bias of *rbcL*. These studies addressed the questions of mutational dynamics, drift, and selection on the evolution of codon choice in *rbcL* but further work is required to define the contribution of synonymous codons within the context of its protein structure/function i.e. the third force.

Further research on codon preferences of residues of the Rubisco-LSU in relationship to secondary structure, solvent accessibility, and evolutionary conservation in a large family of orthologous sequences may help clarify the correlation between codon-usage bias and structural and/or functional importance of residues. This perhaps will provide a detailed framework from which to build more robust models to improve our understanding of molecular evolution of *rbcL*.

4.2 Methods

4.2.1 Strategy

I have compared the codon-usage of the *rbcL* gene with codon-usage of the total codon pool of all protein-coding genes in all available Angiosperm chloroplast genomes in public databases. My objective was to determine if there are significant differences between codon-usage patterns of *rbcL* and that of the whole chloroplast genome, and whether selection plays a role in shaping the codon choices of *rbcL*. The primary focus of the study was to investigate codon bias in *rbcL* within the context of structure and function of the Rubisco large subunit, i.e. at protein level.

Firstly, I define preferred codons in *rbcL* for each amino acid as those used more frequently than other synonymous codons. Against this background I address the following questions:

1) Are preferred codons more likely to be localized to code a particular secondary structure?

2) Are preferred codons more likely to be associated with conserved sites in orthologous sequences?

3) Are preferred codons more likely to encode residues in the core of proteins or on the surface?

4) Are preferred codons more likely to occur at sites for which computational modeling predicts that amino acid substitutions are particularly disruptive?

5) Are these associations, if any, a general characteristic of amino acids in Rubisco-LSU or do they depend on the type of amino acid encoded?

4.2.2 Data preparation

The sequences were downloaded from NCBI; the species name and accession number are given in Appendix 4. The sequences were then edited using BioEdit and aligned using ClustalW. Alignment of more than 200 sequences was done using a parallel version of ClustalW (Li, 2003) on the Sun supercomputing cluster at the National Computing Infrastructure located at the ANU supercomputing facility. Incomplete sequences of fewer than 450 codons in length were excluded from analysis; many *rbcL* sequences in the public databases are incomplete at the 5' and/or 3' ends. I used the Emboss CUSP package to calculate codon-usage. All statistical analyses were performed using the software R (R Development Core Team, 2008).

4.2.2.1 Dataset for comparison of *rbcL* and whole-chloroplast genome codon-usage

For comparative analysis of codon-usage of *rbcL* and all chloroplast genes, a set of chloroplast genes and their respective *rbcL* sequences for 132 Angiosperm species was downloaded from NCBI (Appendix 4.1). In this analysis, only protein-coding genes of the chloroplast genome were considered. I excluded the *rbcL* and *psbA* genes from the cumulative codon pool of all protein-coding genes of the chloroplast, as Morton (2001) has shown these genes to have significantly large CAI (Codon Adaptation Index).

4.2.2.2 Dataset for comparison of *rbcL* codon-usage

The downloaded *rbcL* sequences with more than 450 codons were further pruned at the 5' and 3' ends to produce a dataset of sequences with 453 codons (*rbcL* codon 21 to codon 473). This length was chosen as a compromise to create a dataset with a reasonable number of sequences for analysis. In total, the final dataset comprised 4944 Angiosperm *rbcL* sequences (Appendix 4.3).

4.2.2.3 Localizing codons in secondary structure

I used the spinach Rubisco-LSU x-ray structure (PDB id 8RUC) as a reference to map codons to secondary structures. Codons of the *rbcL* gene are categorized to be in helix, beta sheet or no secondary structure, based on the location of the corresponding amino acid in the 3D structure (Figure 4.2) of the Rubisco-LSU. For instance, codons 50-60 of the *rbcL* gene are categorized to be in helix. Likewise codons 24-26 are categorized to be in beta sheet.

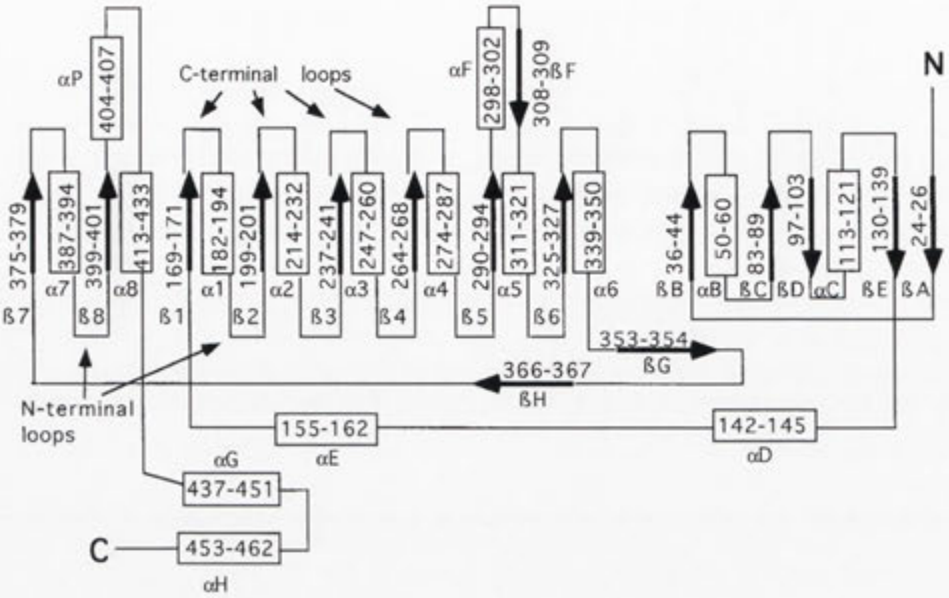


Figure 4.2 Connectivity diagram showing the secondary structure of the large subunit of Rubisco. Rectangles indicate helices, arrows indicate beta strands; numbering of helices and strands follows Knight et al. (1990). Numbers indicate amino acids included in helices or strands. Only some of the C-terminal and N-terminal loops are labeled. (Adapted from Kellogg and Juliano (1997))

4.2.2.4 Sequence conservation in *rbcL* sequences

I used the results of an analysis of conserved and variable residues in Angiosperm Rubisco-LSUs compiled by Dr. Babu Kanappan in our lab. This analysis used a previously compiled Multiple Sequence Alignment (MSA) of Angiosperm Rubisco-LSU sequences (~11,400 species). Based upon sequence conservation at a given residue position, I divided residue positions into two groups:

1. Conserved positions (conserved in > 99.5% Rubisco-LSU sequences)
2. Variable positions (conserved in < 99.5% Rubisco-LSU sequences)

4.2.2.5 Measure of structural sensitivity

I used the structural sensitivity measure developed by Zhou et al. (2009) in this study. They used the Rosetta ΔΔG module (Kortemme and Baker, 2002, Kortemme et al., 2004) to estimate the change in the free energy gap, ΔΔG, for all 19 possible single point amino acid substitutions at each site. They classified sites at which at least two mutations had ΔΔG >3.0 kcal/mol as important sites and all other sites as unimportant sites. The hypothesis is that if selection for translational accuracy acts to minimize mistranslation-induced protein misfolding then sites with higher structural importance should associate with preferred codons and *vice versa*.

4.2.2.6 Solvent accessibility

A web based tool "Get Area" located at the portal http://curie.utmb.edu/area.html was used for calculation of solvent accessible surface areas (SASA). This tool uses the method of Fraczkiewicz and Braun (1998). It takes the PDB file as input and calculates solvent accessible surface area of each residue in the protein. By default, residues at subunit-subunit interfaces are also considered as having large solvent accessible surface area. To correct this error, the input PDB file was modified to combine all the atoms in the hexadecamer into a single molecule by deleting the lines containing the "TER" keyword which indicates the end of records for a chain. SASA computed from this modified input was used to identify surface residues. Residues are considered to be solvent exposed if the ratio value exceeds 40% and to be buried if the ratio value is less than 40%.

4.2.3 Statistical tests of Association

4.2.3.1 Odds ratio

The odds ratio is a way of comparing whether the probability of a certain event is the same for two groups. Shown below is the typical 2×2 contingency table, Table 4.1.

Table 4.1 Example of a 2×2 contingency table

	X+	X-
Y+	a	b
Y-	c	d

The odds ratio can be understood by first noticing what the odds are in each row of the table. The odds for row Y+ are a/b. The odds for row Y- are c/d. The odds ratio (OR) is simply the ratio of the two odds

$$\mathrm{OR}=\frac{a/b}{c/d}$$

which can be simplified to

$$\mathrm{OR}=\frac{\mathrm{ad}}{\mathrm{bc}}.$$

An odds ratio of 1 implies that the event is equally likely in both groups; an odds ratio > 1 implies that the event is more likely in the first group; an odds ratio < 1 implies that the event is less likely in the first group. Notice that if the odds are the same in each row, then the odds ratio is 1. The odds ratio yields zero/undefined results in contingency tables where any of the values (a/b/c/d) in the contingency table is "0", so all such tables are excluded from the analyses.

4.2.3.2 Mantel–Haenszel procedure

To combine 2×2 contingency tables, the Mantel-Haenszel procedure (Mantel and Haenszel, 1959, Mantel, 1963) has been used. The basic principle is that all 2×2 contingency tables are independent. That is, indexing tables by i, with i^{th} table given by Table 4.2.

Table 4.2 Example of i^{th} 2×2 contingency table with index i

	X+	X-
Y+	a_i	b_i
Y-	c_i	d_i

The Mantel-Haenszel estimator for the common odds ratio (i.e., the single odds ratio ψ assumed to underlie all tables being analyzed) is

$$\text{MH Estimator } \psi = \left.\sum_i \frac{a_i d_i}{n_i}\middle/\sum_i \frac{b_i c_i}{n_i}\right.$$

where n_i is the total number of observations for the i^{th} 2×2 contingency table i.e.,

$$n_i = a_i+b_i+c_i+d_i$$

4.2.4 Analysis performed in this study

4.2.4.1 Codon propensities

Codon, Cdn, has propensity, P^{SS}_{Cdn}, for a secondary structure, SS, as calculated by Saunders and Deane (2010).

$$P^{SS}_{Cdn} = {N^{SS}_{Cdn}/N_{Cdn}} \Big/ {N_{SS}/N}$$

where

N^{SS}_{Cdn} = number of times Cdn is observed in secondary structure SS

N_{Cdn} = total occurrences of Cdn

N_{SS} = all observations of the secondary structure SS

N= total number of observations

A propensity >1 means that the codon is over-represented in the secondary structure and a propensity <1 indicates that the codon is under-represented.

4.2.4.2 Defining preferred codons

As an example, amino acid Ala has four synonymous codons. If there is no preference, all codons should be used as expected by the degeneracy of the genetic code (0.25 in case of Ala as 4-fold degenerate amino acid). So, for example as shown in Table 4.3, as the total number of Ala residues in spinach Rubisco-LSU is 43, all synonymous codons should be used equally i.e. 10.75 times (43 × 0.25). This is defined as the Expected codon count.

Table 4.3 Calculation of expected codon count for amino acid Alanine in the *rbcL* gene of spinach compared with observed count

	Observed codon count	Expected codon count
GCA	14	10.75
GCC	4	10.75
GCG	4	10.75
GCT	21	10.75
Total	43	43

I stratified the codon count data by synonymous codon family and constructed a separate 2×2 contingency table (see Table 4.4) for each synonymous codon family for each of 4944 species in my *rbcL* dataset, i.e. 4944 *rbcL* sequences. The codon-usage odds ratio for each codon for all the species in the *rbcL* dataset was combined into the common odds ratio using the Mantel-Haenszel procedure. The null hypothesis in this analysis assumes that synonymous codons should be used with equal frequency in any given species.

Table 4.4 Examples of 2×2 Contingency tables for all codons of Alanine in the spinach *rbcL* gene

2×2 contingency table for codon GCT				**2×2 contingency table for codon GCA**			
	Observed	Expected	Odds ratio		Observed	Expected	Odds ratio
GCT	21	10.75	**2.86**	**GCA**	14	10.75	**1.44**
GCC, GCA, GCG	22	32.25		GCC, GCG, GCT	29	32.25	
2×2 contingency table for codon GCC				**2×2 contingency table for codon GCG**			
	Observed	Expected	Odds ratio		Observed	Expected	Odds ratio
GCC	4	10.75	**0.33**	**GCG**	4	10.75	**0.33**
GCA, GCG, GCT	39	32.25		GCA, GCC, GCT	39	32.25	

Preferred Codons: odds ratio > 1, non-preferred Codons: odds ratio < 1

The odds ratio of codon-usage between observed and expected codon count for spinach is then, (21/10.75) /(22/32.25) = 2.86 for GCT, (14/10.75)/(29/32.25)= 1.44 for GCA, (4/10.75)/(39/32.25) =0.33 for GCC, and (4/10.75)/(39/32.25)=0.33 for GCG, respectively. This shows that the probability of GCT/GCA being used over the other 3 synonymous codons in spinach *rbcL* is 2.86/1.44 times more than the Expected codon count, making them the preferred codons for Ala in spinach; conversely, the probability of GCC/GCG being used over the other 3 synonymous codons in spinach *rbcL* is 0.33 (both GCC and GCG has same odds ratio) times less than the Expected codon count, making them the Non-preferred codons for Ala in spinach.

4.2.4.3 Comparison of *rbcL* and chloroplast codon-usage

I again utilized the 2×2 contingency table (Table 4.5) to compare the codon-usage of *rbcL* and the complete set of genes of the chloroplast genome (omitting genes *rbcL* and *psbA*), as noted in 4.2.2.1.

Table 4.5 Example of 2×2 contingency tables for amino acid Alanine to compare codon-usage in the *rbcL* gene and the genes of its respective chloroplast genome for spinach

2×2 contingency table for codon GCT				**2×2 contingency table for codon GCA**			
	GCT	GCC+ GCA+ GCG	Odds ratio		GCA	GCC+ GCT+ GCG	Odds ratio
rbcL	21	22	**1.27**	*rbcL*	14	29	**1.30**
Chloroplast	545	722		Chloroplast	344	923	
2×2 contingency table for codon GCC				**2×2 contingency table for codon GCG**			
	GCC	GCT+ GCA+ GCG	Odds ratio		GCG	GCC+ GCA+ GCT	Odds ratio
rbcL	4	39	**0.5**	*rbcL*	4	39	**0.67**
Chloroplast	215	1052		Chloroplast	163	1104	

The odds ratio of codon-usage between *rbcL* and the chloroplast-genome codon count is (21/22)/(545/722) = 1.27 for GCT, (14/29)/(344/923)= 1.3 for GCA, (4/39)/(215/1052) =0.5 for GCC, and (4/39)/(163/1104)=0.67 for GCG, respectively for spinach. This shows that the probability of GCT and GCA being used over the other 3 synonymous codons in spinach *rbcL* is 1.27 times and 1.30 times more in the gene *rbcL* as compared with genes of the chloroplast genome, whereas the probability of GCC and GCG being used over the other 3 synonymous codons in spinach *rbcL* is 0.5 times and 0.67 times less in the gene *rbcL* as compared with genes of the chloroplast genome. The codon-usage odds ratio for each codon for all 132 species in the dataset for comparison of *rbcL* and the whole-chloroplast genome codon-usage was combined into the common odds ratio using the Mantel-Haenszel procedure.

4.2.4.4 Association between preferred codons and evolutionarily conserved sites, buried sites and structurally important sites

I defined a set of preferred codons for the *rbcL* gene (see example in Table 4.4). For each of 4944 species, separate 2×2 contingency tables were constructed for the 18 amino acids encoded by at least two codons for three properties, i.e. evolutionary conservation (see section 4.2.2.4), solvent accessibility (see section 4.2.2.6) and structural importance (see section 4.2.2.5), as shown in Table 4.6 for evolutionarily conserved sites. For each of these 18 amino acids, I calculated a joint odds ratio of the preferred codon-usage between category variables such as buried and exposed/conserved and variable/structurally important and unimportant sites using the Mantel-Haenszel procedure.

Table 4.6 Example of a 2×2 contingency table for amino acid Alanine to test the association between preferred codon [a] and conserved and variable [b] residue sites for spinach

Codon	Conserved	Variable
GCA	11	2
GCG	2	2
GCC	3	1
GCT	14	6

	Codon	Conserved	Variable	Odds ratio
Preferred	GCA+GCT	25	8	**1.87**
Non-preferred	GCG+GCC	5	3	

[a] Preferred codons are defined by the procedure demonstrated in Table 4.4. Only 453 codons (codons 21 to codon 473) were used for counting Alanine. [b] Conserved and variable sites are defined in section 4.2.2.4.

The odds ratio of preferred codons (GCA+GCT) usage between conserved-site codon-usage is then (25/8) /(5/3) = 1.87 for this contingency table. This shows the degree of association of preferred codons for Alanine with conserved sites. The probability of preferred codons being used at conserved sites is 1.87 times more than that of non-preferred codons.

4.2.4.5 Calculation of overall odds ratio between preferred codons and structural properties

For each of 4944 species in my *rbcL* dataset, separate 2×2 contingency tables for preferred and non-preferred codons for all amino acids and for each of the properties being considered i.e. evolutionarily conserved sites, buried sites and structurally important sites were constructed. Then overall odds ratios for all species for each of the properties were calculated using the Mantel-Haesenzel procedure.

4.3 Results

4.3.1 A+T % of *rbcL* gene and chloroplast genome

I calculated the overall A+T content and third-position A + T content of all the protein-coding genes of 132 chloroplast genomes and their corresponding *rbcL* gene. As depicted in Figure 4.3, the overall A + T content of protein-coding genes of the chloroplast genomes (59-67%) is marginally higher than that for the *rbcL* gene (53-58%) in my dataset. The third-position A + T content of chloroplast genomes (65-73 %) is very similar to that for the *rbcL* gene (63-75 %).

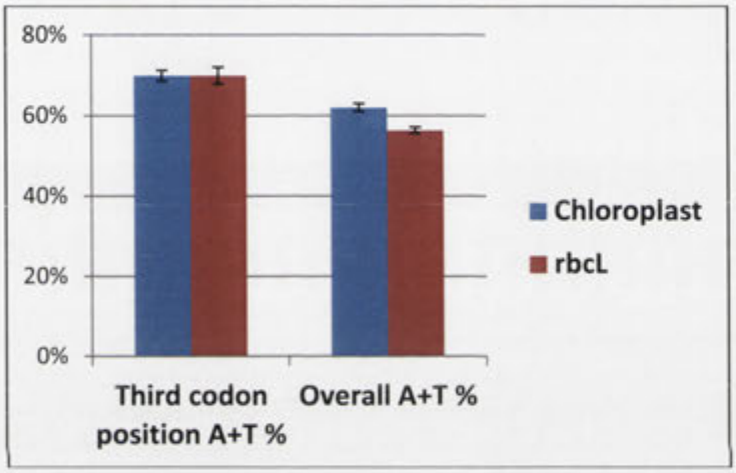


Figure 4.3 A+T content of the third-codon position and full codon of 132 chloroplast genomes and their corresponding *rbcL* genes. Error bars are plotted with standard deviation.

4.3.2 Preferred codons in *rbcL*

To identify preferred codons in *rbcL*, I compared the codon-usage pattern of *rbcL* to codon frequencies predicted by the degeneracy of the genetic code (see Methods 4.2.3.2). The results in Table 4.7 show that except for Ile (ATC), all amino acids show a clear preference for NNA (codon ending in A) and NNT codons (codon ending in T), consistent with the overall high A+T content of chloroplast genomes. Furthermore, 10 amino acids, i.e. Ala (GCT), Asp (GAT), Glu (GAA), Lys (AAA), Pro (CCT), Gln (CAA), Arg (CGT), Ser (TCT), Thr (ACT) and Val (GTA) have odds ratio > 2.5 for preferred codons.

Table 4.7 Odds ratio of *rbcL* codon-usage as compared to codon frequencies predicted by the degeneracy of the genetic code [a]

Amino Acid	Codon	Odds ratio	Amino Acid	Codon	Odds ratio	Amino Acid	Codon	Odds ratio
Ala	GCA	1.11	Lys	AAA	4.09	Ser	AGT	0.85
	GCG	0.32		AAG	0.24		AGC	0.92
	GCT	2.97	Leu	TTA	1.40		TCA	0.50
	GCC	0.48		TTG	1.56		TCG	0.40
Cys	TGT	1.91		CTA	1.23		TCT	3.09
	TGC	0.52		CTG	0.61		TCC	1.47
Asp	GAT	4.32		CTT	1.44	Thr	ACA	0.62
	GAC	0.23		CTC	0.16		ACG	0.17
Glu	GAA	2.80	Asn	AAT	2.09		ACT	4.23
	GAG	0.36		AAC	0.48		ACC	0.81
Phe	TTT	1.63	Pro	CCA	0.59	Val	GTA	2.71
	TTC	0.62		CCG	0.39		GTG	0.41
Gly	GGA	1.36		CCT	3.52		GTT	1.70
	GGG	0.64		CCC	0.69		GTC	0.16
	GGT	2.29	Gln	CAA	2.73	Tyr	TAT	2.15
	GGC	0.27		CAG	0.37		TAC	0.47
His	CAT	1.47	Arg	AGA	1.27			
	CAC	0.68		AGG	0.24			
Ile	ATA	0.16		CGA	1.23			
	ATT	1.87		CGG	0.28			
	ATC	1.68		CGT	3.49			
				CGC	0.75			

[a] **Preferred NNA and NNT codons highlighted in blue. Preferred NNC and NNG codons highlighted in red.**

4.3.3 Number of tRNA genes encoded by chloroplast

I compiled the data on tRNA genes encoded by chloroplast genomes for 123 Angiosperm chloroplast genomes available in the public domain (Appendix 4.2). The 123 genomes were selected on the basis of the availability of annotated tRNA genes in the public databases. The results in Table 4.8 demonstrate that most of these chloroplast genomes code for only 28 cognate tRNA genes. In addition, residues Leu, Val, Ser, Thr, Arg and Gly have tRNA genes encoded for more than one codon in chloroplast. Furthermore, chloroplast genomes only encode for NNC (codon ending in C) and NNG (codon ending in

G) tRNA genes for Cys, Phe, His, Tyr, Ile, Asn and Asp whereas only NNA and NNT tRNA genes are encoded in chloroplast for Lys, Pro, Arg, Ala, Gln and Glu.

Table 4.8 Number of cognate[a] tRNA genes encoded by 123 Angiosperm chloroplast genomes[b]

AA	tRNA	Codon	AA	tRNA	Codon	AA	tRNA	Codon	AA	tRNA	Codon
Phe	1	TTT	Ser	0	TCT	Pro	1	CCT	Stop	0	TAA
	125	TTC		121	TCC		4	CCC	Stop	0	TAG
Leu	124	TTA		121	TCA		123	CCA	Stop	0	TGA
	235	TTG		13	TCG		0	CCG	Trp	125	TGG
	0	CTT		0	AGT	His	4	CAT	Cys	0	TGT
	1	CTC		125	AGC		149	CAC		123	TGC
	123	CTA	Tyr	2	TAT	Gln	123	CAA	Arg	234	CGT
	0	CTG		125	TAC		0	CAG		0	CGC
Ile	1	ATT	Thr	0	ACT	Asn	2	AAT		0	CGA
	204	ATC		128	ACC		238	AAC		0	CGG
	9	ATA		123	ACA	Lys	115	AAA		123	AGA
Met	493	ATG		9	ACG		0	AAG		0	AGG
Val	0	GTT	Ala	0	GCT	Asp	0	GAT	Gly	0	GGT
	241	GTC		0	GCC		123	GAC		102	GGC
	127	GTA		231	GCA	Glu	139	GAA		126	GGA
	0	GTG		0	GCG		0	GAG		0	GGG

[a] Cognate tRNA is tRNA that recognises a codon during translation. [b] All **NNC** and **NNG** codons with cognate tRNA are highlighted in **red**. **NNA** and **NNT** codons with cognate tRNA are highlighted in **blue**.

4.3.4 Comparison of *rbcL* and chloroplast codon-usage

I generated and compared codon-usage statistics for the 132 chloroplast genomes from Angiosperms and their corresponding *rbcL* genes (Table 4.9). The 132 genomes were selected on the basis of the availability of annotated protein-coding genes in public databases. The results in Table 4.9 reveal that 9 amino acids, i.e. Cys, Glu, Phe, His, Ile, Asn, Gln, Tyr and Ser show a significant preference in *rbcL* for NNC and NNG codons, 6 amino acids, i.e. Ala, Asp, Gly, Lys, Pro and Val show a significant preference in *rbcL* for NNA and NNT codons, and 3 amino acids, i.e. Leu, Arg and Thr show no preference, when codon-usage of the *rbcL* gene is compared with that for all the protein-coding genes in the chloroplast genome.

Table 4.9 Odds ratio of *rbcL* codon-usage compared with that for chloroplast genome [a]

AA	Codon	Odds ratio	AA	Codon	Odds ratio	AA	Codon	Odds ratio
Ala	GCA	**1.03**	Lys	AAA	**1.36**	Ser	AGT	0.86
	GCG	0.80		AAG	0.73		AGC	**2.24**
	GCT	**1.24**	Leu	TTA	0.65		TCA	0.82
	GCC	0.73		TTG	**1.27**		TCG	0.82
Cys	TGT	0.64		CTA	**1.60**		TCT	**1.20**
	TGC	**1.56**		CTG	**1.76**		TCC	**1.27**
Asp	GAT	**1.16**		CTT	0.98	Thr	ACA	0.52
	GAC	0.86		CTC	0.44		ACG	0.34
Glu	GAA	0.91	Asn	AAT	0.55		ACT	**2.06**
	GAG	**1.10**		AAC	**1.82**		ACC	**1.29**
Phe	TTT	0.80	Pro	CCA	0.78	Val	GTA	**1.39**
	TTC	**1.25**		CCG	0.98		GTG	0.68
Gly	GGA	0.67		CCT	**1.64**		GTT	**1.23**
	GGG	0.99		CCC	0.67		GTC	0.31
	GGT	**1.73**	Gln	CAA	0.87	Tyr	TAT	0.47
	GGC	0.63		CAG	**1.15**		TAC	**2.11**
His	CAT	0.44	Arg	AGA	0.61			
	CAC	**2.25**		AGG	0.44			
Ile	ATA	0.22		CGA	0.77			
	ATT	0.92		CGG	0.65			
	ATC	**3.54**		CGT	**2.53**			
				CGC	**2.32**			

[a] **Preferred NNA and NNT codons highlighted in blue. Preferred NNC and NNG codons highlighted in red.**

Interestingly, 6 of 9 amino acids, i.e. 2-fold degenerate Cys, Phe, His, Asn and Tyr and 3-fold degenerate Ile showing preference for NNC codons in the *rbcL* gene have only cognate tRNA for NNC codons encoded in chloroplast as shown in Table 4.8. Considering the evident compositional bias both in *rbcL* and all the protein-coding genes in chloroplast genomes towards higher A+T content (as illustrated in Figure 4.3), this finding suggests a role for selection in adapting the codon-usage of these 2-fold (Cys, Phe, His, Asn and Tyr) and 3-fold (Ile) degenerate amino acids from a low to high NNC representation in *rbcL*. It is also interesting to note that the *rbcL* gene has the second highest synonymous substitution rate in the two-fold degenerate groups as well as the second highest C content of chloroplast genes after the gene *psbA* (Morton, 1994).

4.3.5 Codon-usage of catalytic residues

In 1994, Hiroshi Akashi developed an elegant hypothesis for selection of translational accuracy of coding sequences which postulated that usage of more-accurate synonymous codons (preferred codons in this case) will be favored at important (e.g., catalytic residues) amino-acid sites where translation errors could disrupt protein folding or function. At less-important (e.g., evolutionarily variable) amino-acid sites, errors are presumably more tolerated, and, therefore, less-accurate codons (non-preferred codons in this case) are more likely to be favored.

I tested Akashi's hypothesis on my *rbcL* dataset. First I delineated catalytic residues as summarized in the literature (Cleland et al., 1998, Kannappan and Gready, 2008), and then calculated codon-usage for each of the 4944 sequences in the *rbcL* dataset (Appendix 4.3) for each catalytic residue.

Table 4.10 Codon-usage for catalytic residues of *rbcL* dataset [a,b]

Residue	Codon	% of seq	Residue	Codon	% of seq
Glu60	GAA	98%	Asp203	GAT	98%
	GAG	2%		GAC	2%
Asn123	AAT	96%	Glu204	GAA	43%
	AAC	3%		GAG	57%
Lys175	AAA	99%	His294	CAC	60%
	AAG	1%		CAT	40%
Lys177	AAA	99%	Lys334	AAA	96%
	AAG	1%		AAG	4%
Lys201	AAA	79%			
	AAG	21%			

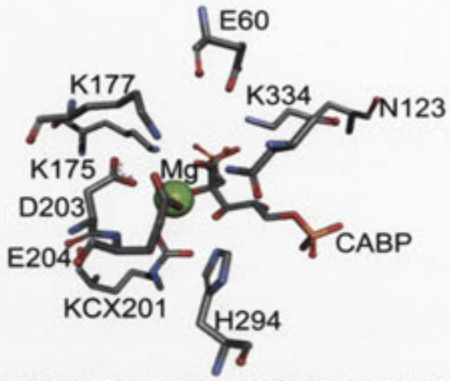


Figure 4.4 The active site of Spinach Rubisco with CABP bound (8RUC). Cartoon representation of the active-site residues, Mg and CABP. Side chains of active-site residues are shown in sticks. CABP stands for 2-carboxyarabinitol-1, 5-diphosphate, an inhibitor of Rubisco's catalytic reaction. Mg is Magnesium ion.

[a] All catalytic residues are totally conserved. [b] The favored preferred codons are highlighted in red.

As evident in Table 4.10, most of the catalytic residues, i.e. 7 of 9 residues (Glu60, Asn123, Lys175, Lys177, Lys201, Asp203 and Lys 334) in the *rbcL* dataset favor the preferred codon over the non-preferred codon; all of the catalytic residues have 2-fold degenerate codons. The exceptions, residues Glu204 and His294 show only moderate codon preference; Glu 204 with 57% and 43% respectively, for non-preferred (GAG) and preferred (GAA) codons and His 294 with 60% and 40% respectively, for non-preferred (CAC) and preferred (CAT) codons.

To compare these results with the general codon-usage pattern of residues present in the Rubisco active-site, I analyzed occurrences of these residues in other sequence positions of the Rubisco-LSU; considering only those positions that are more than 95% conserved.

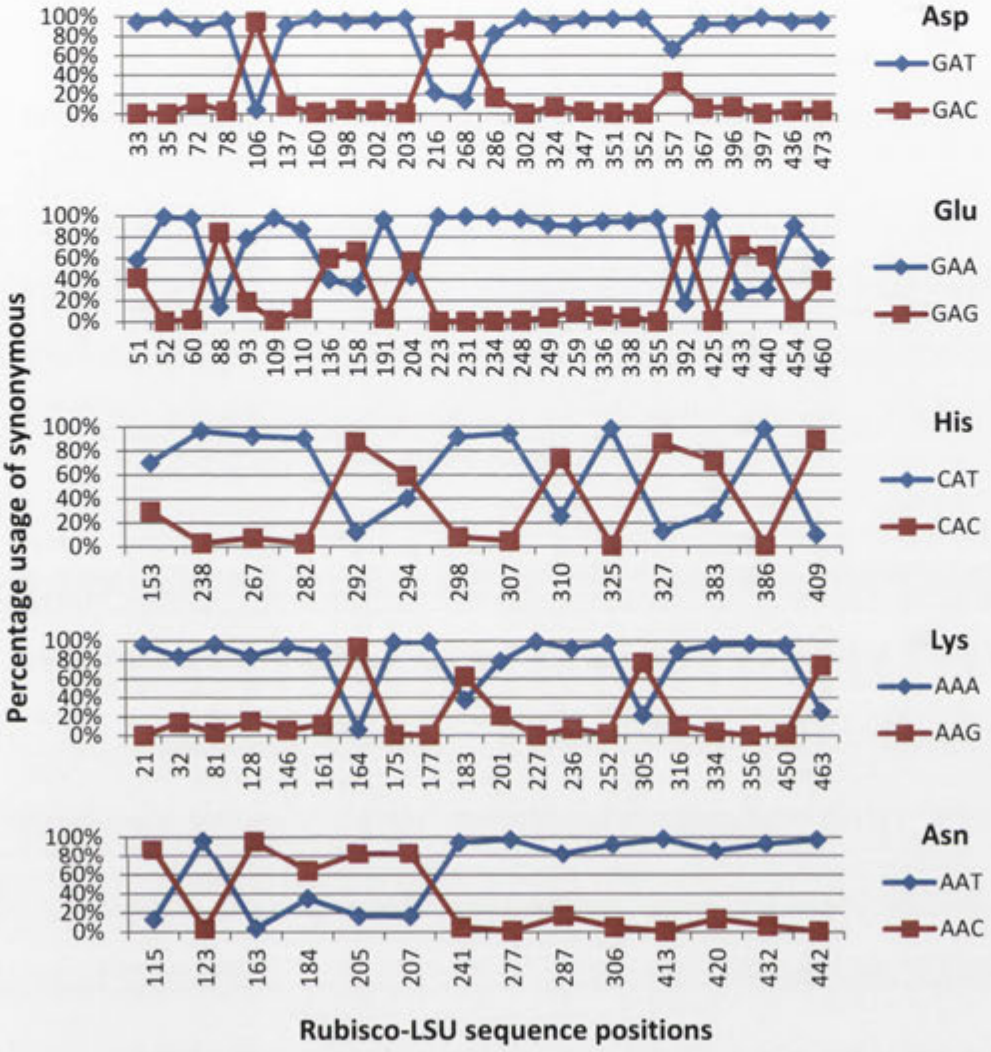


Figure 4.5 General codon usage patterns of residues forming the Rubisco active site at other sequence positions in Rubisco-LSU. The codon-based multiple sequence alignment of my *rbcL* dataset was used to prepare the plot.

The most obvious feature of this analysis, as shown in Figure 4.5, is preferential usage of only one synonymous codon over the other by these residues, at most of these sequence positions for Asp, Glu and Lys. These residues i.e. Asp, Glu and Lys, in general appear to be using preferred codons (> 80 %) throughout the Rubisco-LSU sequence except for positions 106, 216 and 268 for Asp, positions 88 and 392 for Glu, and positions 164, 305 and 463 for Lys, respectively. For the other two residues, Asn and His, do not show such a consistent usage pattern of preferred codons, especially His. However, note that these residues also show mixed codon-usage at some other positions.

4.3.6 Codon preferences of secondary structures in Rubisco-LSU

For each sequence in my *rbcL* dataset, I calculated codon propensities for the 59 codons (see Methods 4.2.3.1 for an example) for each secondary structure (helix, beta sheet, not helix or beta which will be called N_{HB}). For each of the 59 codons, I calculated the percentage of sequences that have propensity >1 and < 1 for each secondary structure (H/B/N_{HB}) and classified them over- or under-represented by method illustrated in Table 4.11. Table 4.12 shows consolidated data for all secondary structures and significant results for each secondary structure. In the following I discuss these data in detail.

As illustrated in Table 4.11, a codon/set of codons is/are defined to be under-represented in a secondary structure if the majority of sequences (> 80% of the sequences) for that codon/set of codons has propensity < 1, and for at least one other synonymous codon the propensity is >1. Similarly, if a codon/set of codons is over-represented in a secondary structure then the majority of sequences should have propensity >1 for that codon/set of codons and for at least one other synonymous codon the propensity is < 1.

Table 4.11 Criteria for defining whether a codon/ set of codons is/are under- or over-represented in a given secondary structure

A codon/ set of codons is/are under- or over- represented?	Propensity of a codon/ set of codons [a]	Propensity of other synonymous codon/codons [a, b]
Under-represented	< 1	> 1
Over-represented	>1	< 1

[a] Propensity of majority of sequences i.e. > 80% of the sequences. [b] For at least one other synonymous codon.

Table 4.12 Sequence percentages of codon propensities >1 / <1 of *rbcL* dataset (4944 sequences) for each secondary structure in the Rubisco-LSU [a]

		Codon Propensity								Codon Propensity								Codon Propensity					
AA	Codon	Helix		Beta		N_{HB}		AA	Codon	Helix		Beta		N_{HB}		AA	Codon	Helix		Beta		N_{HB}	
		>1	<1	>1	<1	>1	<1			>1	<1	>1	<1	>1	<1			>1	<1	>1	<1	>1	<1
Ala	GCA	99	1	47	53	0	100	Lys	AAA	0	100	0	100	100	0	Ser	AGT	6	94	3	97	99	1
	GCG	97	3	7	93	7	93		AAG	19	81	16	84	92	8		AGC	70	30	1	99	64	36
	GCT	100	0	55	45	0	100	Leu	TTA	94	6	6	94	31	69		TCA	16	84	59	41	18	82
	GCC	100	0	6	94	3	97		TTG	68	32	99	1	0	100		TCG	10	90	30	70	10	90
Cys	TGT	100	0	2	98	0	100		CTA	56	44	98	2	1	99		TCT	9	91	0	100	100	0
	TGC	97	3	99	1	0	100		CTG	38	62	96	4	6	94		TCC	25	75	7	93	96	4
Asp	GAT	0	100	0	100	100	0		CTT	0	100	97	3	98	2	Thr	ACA	1	99	0	100	100	0
	GAC	23	77	69	31	61	39		CTC	36	64	12	88	3	97		ACG	23	77	4	96	35	65
Glu	GAA	100	0	0	100	40	60	Asn	AAT	100	0	0	100	2	98		ACT	43	57	5	95	84	16
	GAG	94	6	49	51	6	94		AAC	67	33	7	93	91	9		ACC	12	88	35	65	84	16
Phe	TTT	91	9	41	59	10	90	Pro	CCA	2	98	9	91	99	1	Val	GTA	95	5	99	1	0	100
	TTC	10	90	3	97	97	3		CCG	15	85	9	91	94	6		GTG	15	85	62	38	81	19
Gly	GGA	0	100	3	97	100	0		CCT	7	93	21	79	98	2		GTT	95	5	0	100	79	21
	GGG	0	100	6	94	100	0		CCC	9	91	75	25	71	29		GTC	15	85	6	94	84	16
	GGT	0	100	0	100	100	0	Gln	CAA	4	96	44	56	89	11	Tyr	TAT	31	69	100	0	0	100
	GGC	18	82	8	92	93	7		CAG	43	57	58	42	50	50		TAC	1	99	99	1	27	73
His	CAT	0	100	100	0	26	74	Arg	AGA	100	0	0	100	0	100								
	CAC	0	100	99	1	51	49		AGG	77	23	0	100	5	95								
Ile	ATA	39	61	56	44	8	92		CGA	23	77	100	0	0	100								
	ATT	49	51	96	4	4	96		CGG	33	67	4	96	6	94								
	ATC	51	49	98	2	0	100		CGT	97	3	6	94	9	91								
									CGC	37	63	49	51	61	39								

[a] Sequence percentages of over-represented codons are highlighted in **red** and under-represented codons are highlighted in **blue**.

4.3.6.1 Codon preferences in α-helix

As shown in Table 4.12 and depicted in Figure 4.6, codons coding for residues Phe (TTT), Leu (TTA), and Val (GTA and GTT) show a strong preference for helix formation. In my *rbcL* dataset (4944 sequences), 91, 94 and 95% of sequences show codon propensity >1 for codons: TTT (Phe), TTA (Leu), and, GTA and GTT (Val), respectively, to be incorporated into a helix. Conversely 90%, 100% and 85% sequences show propensity < 1 for codons TTC (Phe), CTT (Leu), and GTG and GTC (Val), respectively, and are found to be under-represented in codons coding for helices (Figure 4.6). Two of these under-represented codons TTC (Phe) and CTT (Leu) are found to be over-represented in N_{HB} (Figure 4.8).

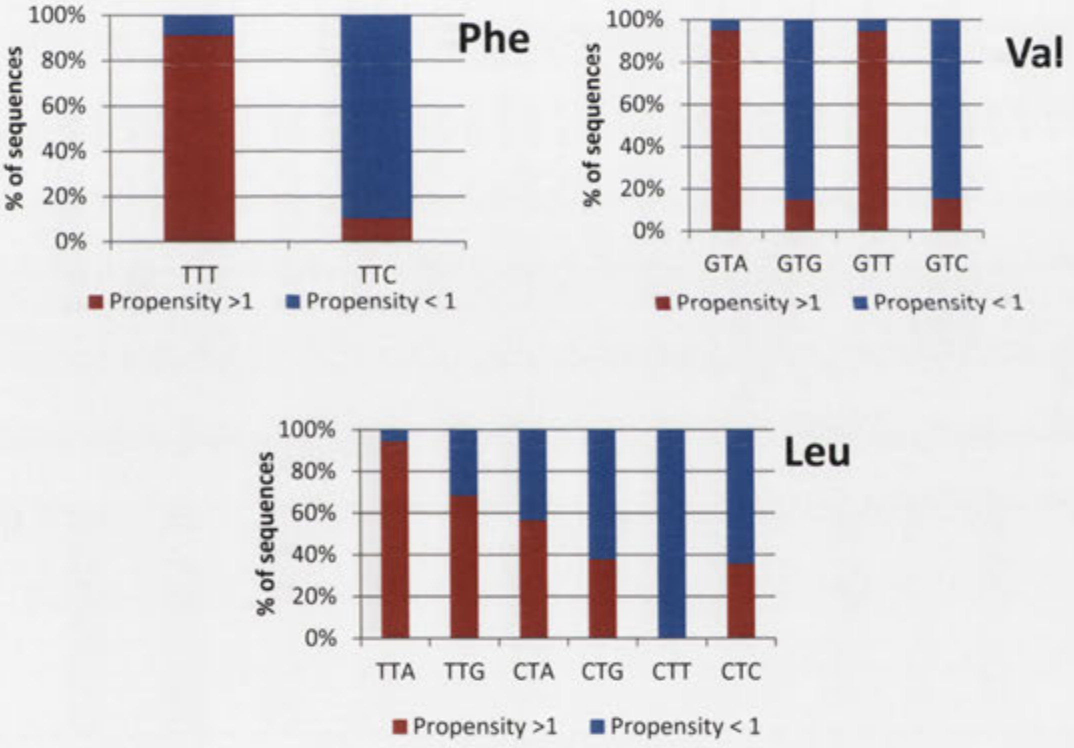


Figure 4.6 In Rubisco-LSU, 4 codons, TTC (Phe), CTT (Leu) and GTC and GTG (Val) are found to be under-represented in codons coding for α-helices. The synonymous codons from same amino acids TTT (Phe), TTA (Leu) and GTA & GTT (Val) are found to be over-represented.

4.3.6.2 Codon preferences in β-strand

As shown in Table 4.12 and depicted in Figure 4.7, codons coding for residues Cys (TGC), Leu (TTG, CTA, CTG and CTT), Arg (CGA) and Val (GTA) show a strong preference to form a β-unit. A high percentage of sequences in my *rbcL* dataset show codon propensity >1 for codons: TGC (Cys, 99% sequences), TTG, CTA, CTG and CTT (Leu, 99%, 98%, 96% and 98% sequences, respectively), CGA (Arg, 100% sequences) and GTA (Val, 99% sequences) to be included in β-strands when translated to protein. Synonymous codons for the same residues: TGT (Cys, 98% sequences), TTA and CTC (Leu, 94% and 88% sequences, respectively), AGA, AGG, CGG and CGT (Arg, 100%, 100%, 96% and 98% sequences, respectively) and GTT and GTC (Val, 100% and 94% sequences, respectively) are found to be under-represented in codons coding for β-sheets. Interestingly, codon CTT (Leu), which is found to be under-represented in helices, is over-represented in both β-strands and N_{HB}.

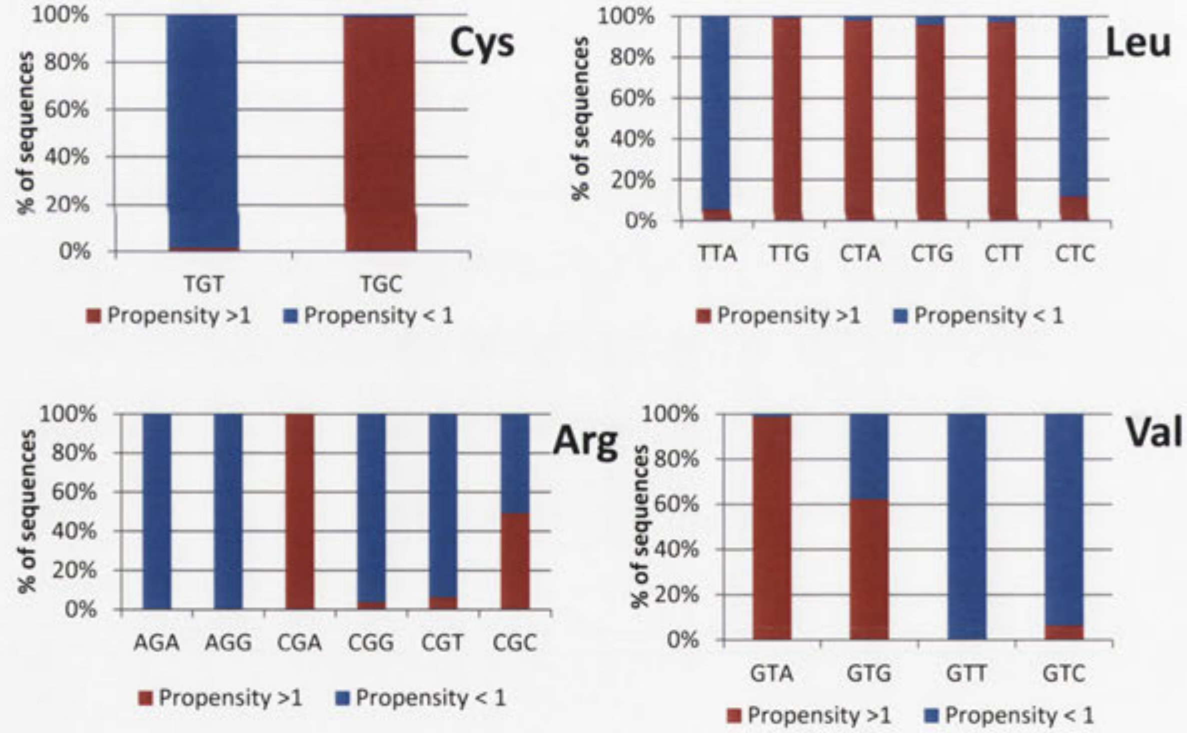


Figure 4.7 In Rubisco-LSU, 7 codons TGC (Cys), TTG, CTA, CTG and CTT (Leu), CGA (Arg) and GTA (Val) are found to be over-represented in codons coding for β-sheets. The synonymous codons from the same amino acids, TGT (Cys), TTA and CTC (Leu), AGA, AGG, CGG and CGT (Arg), and GTT and GTC (Val), are found to be under-represented.

4.3.6.3 Codon preferences in N_{HB}

As shown in Table 4.12 and depicted in Figure 4.8, codons in residues Phe (TTC), Leu (CTT), Ser (AGT, TCT and TCC), Asn (AAC) and Val (GTG and GTC) show a strong preference to code for N_{HB} regions in Rubisco-LSU. A large number of sequences in my *rbcL* dataset show codon propensity >1 for codons: TTC (Phe, 97% sequences), CTT (Leu, 98% sequences), AGT, TCT and TCC (Ser, 99%, 100% and 96% sequences, respectively), AAC (Asn, 91% sequences), and GTG and GTC (Val, 81% and 84% sequences, respectively) to be incorporated into N_{HB} regions. Conversely, synonymous codons from the same residues TTT (Phe, 90% sequences), TCA and TCG (Ser, 82% and 90% sequences, respectively), AAT (Asn, 98% sequences), TTG, CTA, CTG and CTC (Leu, 100%, 99%, 94% and 97% sequences, respectively) and GTA (Val, 100% sequences) are found to be under-represented in N_{HB} regions. Two of these under-represented codons TTT (Phe) and GTA (Val) are found to be over-represented in helices and β-strands, respectively (Figures 6 and 7).

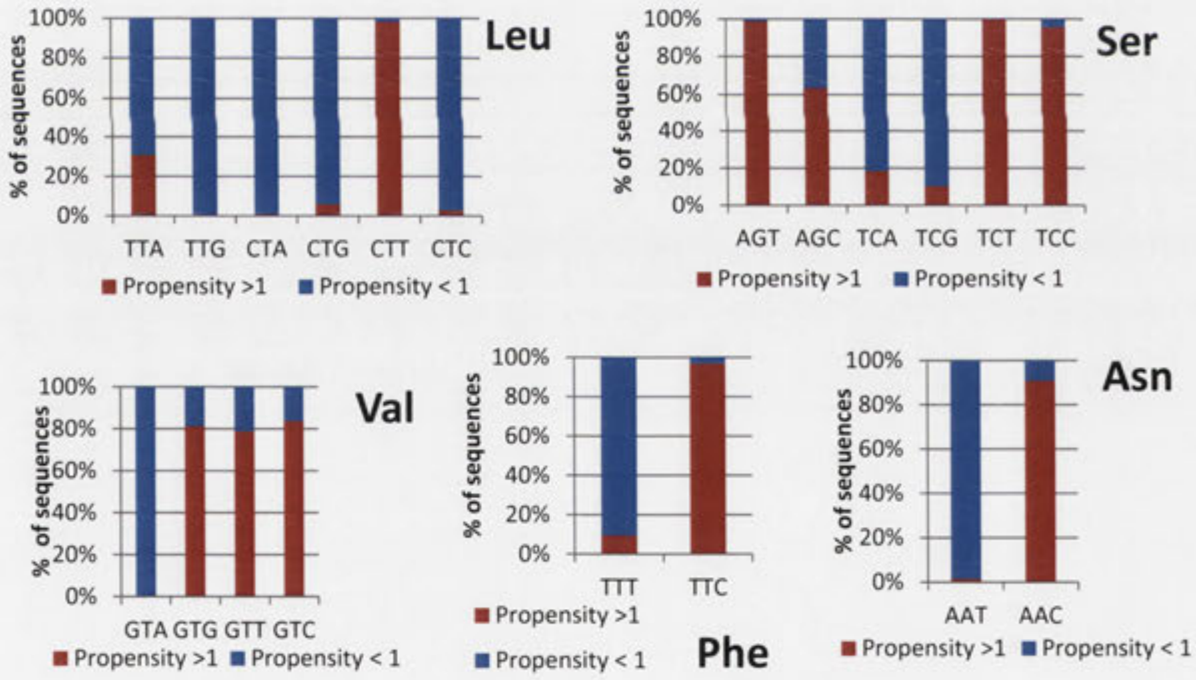


Figure 4.8 In Rubisco-LSU, 8 codons TTC (Phe), AGT, TCT and TCC (Ser), AAC (Asn), CTT (Leu) and Val (GTG and GTC) are found to be over-represented in codons coding for N_{HB} regions. The synonymous codons from same residues TTT (Phe), TCA and TCG (Ser), AAT (Asn), TTG, CTA, CTG and CTC (Leu) and GTA (Val) are found to be under-represented.

4.3.7 Association of preferred codons with conserved sites in the Rubisco-LSU

I applied Akashi's test on all conserved and variable sites in my *rbcL* dataset. For each sequence, I constructed separate 2×2 contingency tables for the 18 amino acids encoded (see Table 4.6 for an example) to test the association between preferred codons (as defined in Table 4.7) and conserved sites in *rbcL*. Then, for each of 18 amino acids, I calculated a joint odds ratio of preferred codon-usage between conserved and variable sites. I used the Mantel-Haenszel procedure to combine the odds ratio for all species for each amino acid. As mentioned in Methods (section 4.2.3.1), a value of the odds ratio greater than "1" signifies the association of preferred codons with conserved sites for the respective amino acid.

Table 4.13 Joint odds ratio of preferred codons usage in the Rubisco-LSU between conserved and variable sites

Residue	Odds ratio [a]
Ala	1.20
Cys	1.59
Asp	2.01
Glu	1.81
Phe	0.39
Gly	------
His	1.54
Ile	------
Lys	------
Leu	1.41
Asn	------
Pro	0.72
Gln	------
Arg	3.93
Ser	6.76
Thr	0.34
Val	------
Tyr	------

[a] Significant at P<0.001

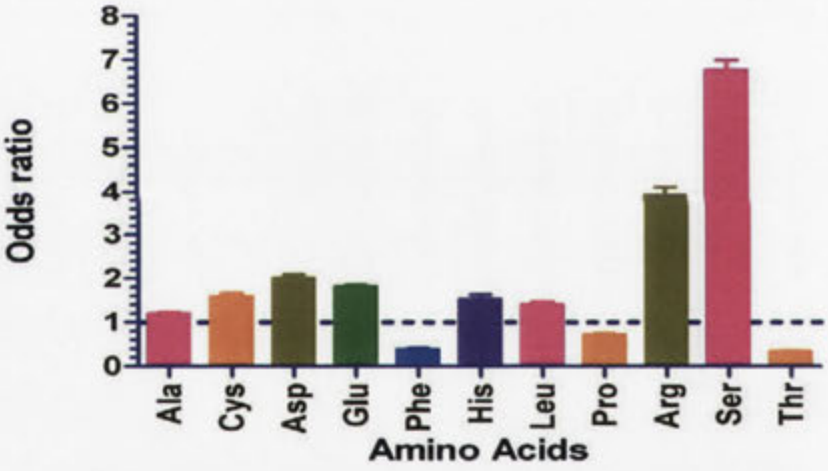


Figure 4.9 Joint odds ratio of preferred codon-usage in Rubisco-LSU between conserved and variable sites plotted with 95% confidence interval. Ala, Cys, Asp, Glu, His, Leu, Arg and Ser show significant association of conserved sites with the preferred codon at confidence P<0.001.

As shown in Table 4.13 and Figure 4.9, a statistically significant association between preferred codons (as defined in Table 4.7) and conserved sites for 8 of 18 amino acids (Ala, Cys, Asp, Glu, His, Leu, Arg and Ser) was found whereas 3 amino acids Phe, Pro and Thr show significant association with non-preferred codons (as defined in Table 4.7).

Non-significant results for Gly, Ile, Lys, Asn, Gln, Tyr and Val are due to absence of non-preferred codons in variable sites (values of "0") in most 2×2 contingency tables of these amino acids, thus making it meaningless to calculate a combined odds ratio for the *rbcL* dataset using the Mantel-Haesenzel procedure. Thus, these null results indicate a lack of statistical power.

4.3.8 Association of preferred codons with buried sites in the Rubisco-LSU

I extended use of Akashi's test to check the association of preferred codons with buried and exposed sites, in analogous fashion to above. This analysis was inspired by the work of Zhou et al. (2009), who discovered a statistically significant association between translationally optimal codons and buried sites.

Table 4.14 Joint odds ratio of preferred codons usage in Rubisco-LSU between buried and exposed sites

Residue	Odds ratio [a]
Ala	1.36
Cys	------
Asp	2.38
Glu	0.62
Phe	1.66
Gly	1.19
His	0.80
Ile	------
Lys	0.49
Leu	------
Asn	------
Pro	0.87
Gln	2.83
Arg	1.16
Ser	5.57
Thr	1.48
Val	2.62
Tyr	------

[a] Significant at P<0.001

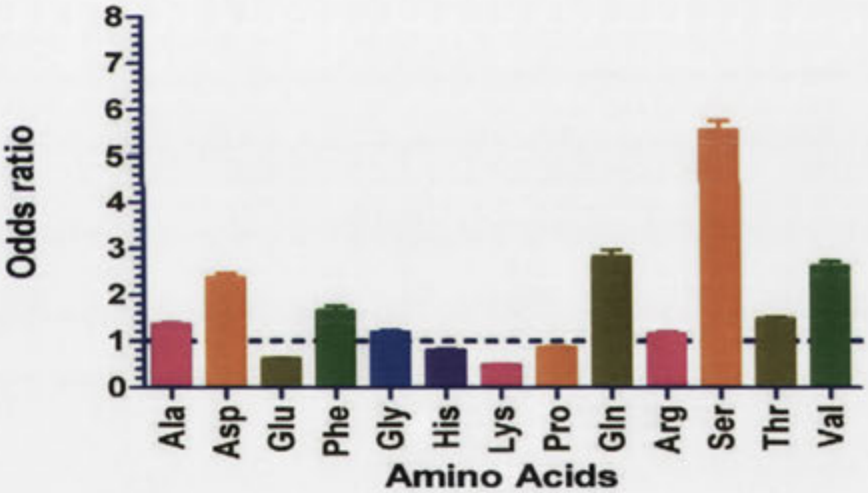


Figure 4.10 Joint odds ratio of preferred codon-usage in Rubisco-LSU between buried and exposed sites plotted with 95% confidence interval. Ala, Asp, Phe, Gly, Gln, Arg, Ser, Thr and Val show significant association of buried sites with preferred codon at confidence P<0.001

As shown in Table 4.14 and Figure 4.10, a statistically significant association between preferred codons (as defined in Table 4.7) and conserved sites for 9 of 18 amino acids (Ala, Asp, Phe, Gly, Gln, Arg, Ser, Thr and Val) was found, whereas 4 amino acids (Glu, His, Lys and Pro) show significant association with non-preferred codons. Non-significant results for Cys, Ile, Leu, Asn and Tyr are the result of absence of non-preferred codons in variable sites (values of "0") in most 2×2 contingency tables of these amino acids, making it meaningless to calculate a combined odds ratio for the *rbcL* dataset by Mantel-Haesenzel procedure. Thus, these null results indicate a lack of statistical power.

4.3.9 Association of preferred codons with structurally important sites in the Rubisco-LSU

I further extended use of Akashi's test to evaluate any association of preferred codons to structurally important and unimportant sites, in analogous fashion to above. This analysis was also inspired by the work of Zhou et al. (2009) which reported a statistically significant association between translationally optimal codons and structurally important sites.

Table 4.15 Joint odds ratio of preferred codons usage in *rbcL* between structurally important and unimportant sites

Residue	Odds ratio [a]
Ala	1.63
Cys	-----
Asp	2.68
Glu	0.72
Phe	0.83
Gly	1.49
His	-----
Ile	-----
Lys	1.24
Leu	0.88
Asn	0.78
Pro	1.83
Gln	-----
Arg	0.83
Ser	3.06
Thr	0.78
Val	-----
Tyr	0.47

[a] Significant at P<0.001 values

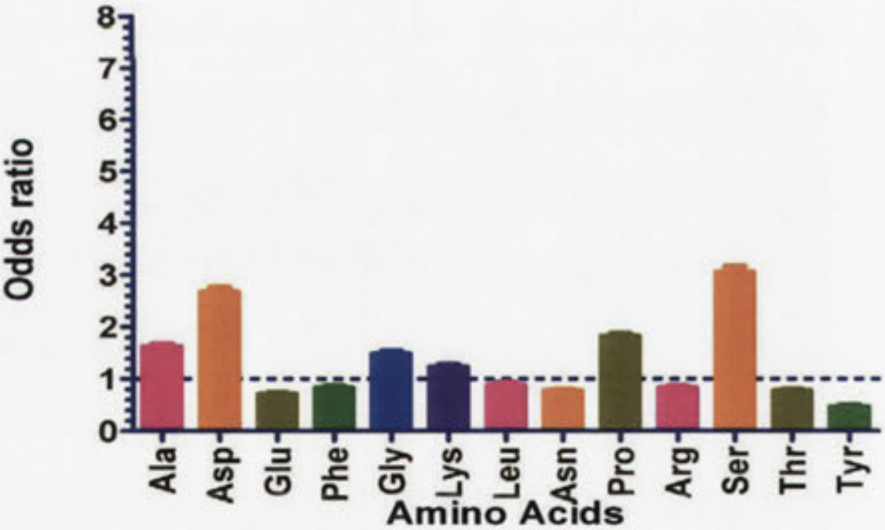


Figure 4.11 Joint odds ratio of preferred codon-usage in *rbcL* between structurally important and unimportant sites plotted with 95% confidence interval. Ala, Asp, Gly, Lys, Pro and Ser show significant association of structurally important sites with preferred codon at confidence P<0.001.

As shown in Table 4.15 and Figure 4.11, I found a statistically significant association between preferred codons (as defined in Table 4.7) and structurally important sites for 6 of 18 amino acids (Ala, Asp, Gly, Lys, Pro and Ser). The remaining seven amino acids (Glu, Phe, Leu, Asn, Arg, Thr and Tyr) show a statistically significant association between non-preferred codons and structurally important sites. The joint odds ratio for 5 amino acids (Cys, His, Ile, Gln and Val) are not statistically significant for this analysis.

4.3.10 Overall odds ratio for association of preferred codons with evolutionarily conserved, buried and structurally important sites in the Rubisco-LSU

For each property, I also used the Mantel–Haenszel procedure to combine all 2×2 contingency tables for all amino acids into a single overall odds ratio for my *rbcL* dataset (see Methods 4.2.3.5). This analysis corresponds to that reported by Drummond and Wilke (2008) and Zhou et al. (2009). I calculated the overall odds ratio separately for all three properties i.e. buried sites, evolutionarily conserved sites and structurally important sites. As evident in Figure 4.12, I found a statistically significant association between preferred codons and all three properties.

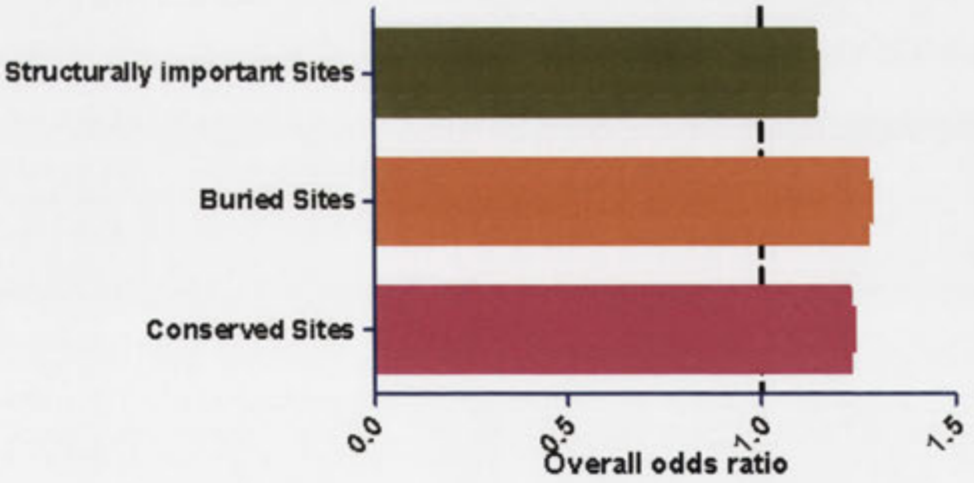


Figure 4.12 Overall odds ratio of preferred codon-usage in *rbcL* with conserved, buried and structurally important sites plotted with 95% confidence interval. All three properties show significant association with preferred codon at confidence P<0.001.

4.4 Discussion

4.4.1 Genome compositional bias towards NNA and NNT codons in *rbcL*

The high A + T content of *rbcL* (Figure 4.3) as well as the overall preference for NNA and NNT codons (Table 4.7) suggest that compositional bias is responsible for the codon bias in *rbcL*; this is consistent with previous reports that attribute codon-usage bias in *rbcL* to low G+C content of the chloroplast genome (Albert et al., 1994, Morton, 1994, Morton and Levin, 1997). In my study, as shown in Table 4.7, all amino acids except Ile (ATC) showed a clear preference for NNA and NNT codons. The mean overall A + T content of all protein-coding genes in 132 chloroplast genomes is 61.9% (Figure 4.3). This is higher than the mean overall A + T content of 56.3% (Figure 4.3) in their corresponding *rbcL* genes, as previously reported (Morton, 1994). The mean third-position A + T content of all protein-coding genes of chloroplast genomes and *rbcL* is the same at 69.8% (Figure 4.3). These results further support the conclusion that an overall bias toward NNA and NNT codons in *rbcL* is a consequence of a high A + T content in chloroplast genomes.

4.4.2 Adaptation to tRNAs encoded by chloroplast genomes to enhance translational efficiency

In the current study, two lines of evidence support the role of codon adaptation in *rbcL*. First, comparison of codon-usage of *rbcL* with that for the all protein-coding genes of chloroplast genome identified 9 of 18 amino acids showing a significant preference in *rbcL* for NNC and NNG codons (Table 4.9). Second, data compilation of tRNA genes on 123 Angiosperm chloroplast genomes revealed that 6 of these 9 amino acids, i.e. 2-fold degenerate Cys, Phe, His, Asn and Tyr and 3-fold degenerate Ile, have only cognate tRNA genes for NNC codons encoded in chloroplast (Table 4.8). Based on these observations, selection appears to be adapting codon-usage of these residues in *rbcL* to tRNA genes encoded in chloroplast, as proposed for unicellular organisms (Ikemura, 1985). Another likely explanation for the observed codon-usage pattern in *rbcL* could be that adaptation to tRNA genes encoded in chloroplast may be limited to those amino acids where a tRNA recognizing the C-terminated codon is the only one coded by the chloroplast genome. The observed codon-usage bias patterns in *rbcL* are analogous to codon-usage patterns of another chloroplast gene, *psbA*. Studies of codon-usage bias in *psbA*, have found adaptation to the chloroplast-encoded tRNA genes, a likely explanation for codon-usage bias of *psbA* (Morton, 1993, 1994, 1996, 1998, Morton and Levin, 1997). However, the extent of selection in *rbcL* is weak; it appears to be limited by genome compositional bias towards A+T content and overall codon-usage in *rbcL* for NNA and NNT codons remains high (Table 4.7).

Counter arguments questioning the premise favoring adaptation to tRNA abundance have been put forward by Wall & Herbeck (2003). They argued that without importing tRNAs, it will be difficult for the chloroplast-translation machinery to efficiently translate a highly expressed gene such as *rbcL* with a high percentage of NNA and NNT codons, even if theoretically chloroplast-encoded tRNAs are sufficient to recognize all codons by super-wobble mechanisms (Pfitzinger et al., 1990, Rogalski et al., 2008). Although tRNA importation may occur with concurrent adaptation to the tRNA pool encoded by the chloroplast genome, this argument needs validation with more research on tRNA import into chloroplasts of photosynthetic organisms.

4.4.3 Catalytic-site residues use high fidelity codons

In *rbcL*, patterns of codon-usage of catalytic residues show strong preferences for preferred codons, suggesting selection in favor of translational accuracy. As evident in Table 4.10, 7 of 9 catalytic residues in my *rbcL* dataset Glu60 (GAA), Asn123 (AAT), Lys175 (AAA), Lys177 (AAA), Lys201 (AAA), Asp203 (GAT) and Lys334 (AAA) have preferred codons. The high fidelity of protein synthesis required at these codons could be enhanced by greater fidelity in the initial discrimination step of protein synthesis at preferred codons, or by more effective proofreading in the subsequent step at these codons (Akashi, 1994).

The remaining 2 catalytic residues Glu204 (GAG) and His294 (CAC) show mixed preferences in codon-usage. The moderate bias of Glu204 towards GAG could suggest that selection is still acting to adapt its codon-usage to preferred codons. The tRNA gene for His in the chloroplast genome is complementary to codon CAC (Table 4.8), which could explain some preference of His294 for codon CAC.

Comparative analysis of general codon-usage patterns of residues present in the Rubisco active site showed that residues Asp, Glu and Lys predominantly use preferred codons throughout the Rubisco-LSU sequence (Figure 4.5). This finding is consistent with results obtained in Table 4.7 which show all amino acids except Ile prefer NNA and NNT codons. A few exceptions to this pattern have been identified, such as Rubisco-LSU sequence positions 106, 216 and 268 for Asp, positions 88 and 392 for Glu and positions 164, 305 and 463 for Lys, respectively, which use non-preferred codons. Further research is required to understand the structural importance of these positions in the Rubisco-LSU as it has been suggested that rarely used synonymous codons are translated more slowly and may have implications for protein folding and/or activity due to translational pause (Buchan and Stansfield, 2007, Tsai et al., 2008). The codon-usage pattern of the other two active-site residues, Asn and His, does not show consistent preferences for preferred codons at their positions in the protein outside the active-site (Figure 4.5); this could be the result of ongoing codon adaptation for these residues as both have cognate tRNA gene encoded in the chloroplast genome for non-preferred codons AAC (Asn) and CAC (His), respectively (Table 4.8).

4.4.4 Secondary structure codon preferences in the Rubisco-LSU

Calculated propensities for specific secondary structures of 59 codons in my *rbcL* dataset showed that propensities of synonymous codons used in regions of different protein secondary structure differ, as evident in Table 4.12. Most of the codons identified as significant for α-helices (TTT (Phe), TTA (Leu), GTA and GTT (Val); Figure 4.6), in β-strands (TGC (Cys), TTG, CTA, CTG and CTT (Leu), CGA (Arg) and GTA (Val); Figure 4.7), and in N_{HB} (TTC (Phe), CTT (Leu), AGT, TCT and TCC (Ser), AAC (Asn), and, GTG and GTC (Val); Figure 4.8) show clear preferences for their respective secondary structures. There is no evidence in my *rbcL* dataset that CGA (Arg) is over-represented at the termini of helices as found by Gu et al. (2003). Also, there is no support for the observation of Gupta et al. (2000) for Pro codons that CCC is over-represented in strand, or that CCA and CCT are most abundant in helices. This is in accordance with findings of Saunders and Deanne (2010), who noted that there is no universal set of significant codons, as structurally significant codons change between organisms.

The results discussed here clearly link codons of the *rbcL* gene with local secondary structure of the Rubisco-LSU. These codon effects have been explained as manifestations of changes in codon translation speed and as secondary structure signals at nucleotide level (Thanaraj and Argos, 1996, Makhoul and Trifonov, 2002, Brunak and Engelbrecht, 1996, Gu et al., 2003, Adzhubei et al., 1996).

4.4.5 Preferred codons in *rbcL* associate with conserved and buried sites in the Rubisco-LSU but show comparatively weak association with a structural sensitivity measure ΔΔG

My study has examined the relationship between *rbcL* codon-usage bias and Rubisco-LSU protein structure. According to Akashi's hypothesis (Akashi, 1994), if natural selection biases *rbcL* codon-usage to enhance the accuracy of Rubisco-LSU protein synthesis, then preferred codon-usage will be stronger at functionally constrained amino acid positions than at less constrained sites. Inspired by previous works of Akashi (1994) and Zhou et al. (2009), I analyzed three different kinds of Rubisco-LSU information that correlate with relative tolerance to amino acid changes at different peptide positions: i) sequence conservation (Table 4.13 and Figure 4.9), ii) solvent accessibility (Table 4.14 and Figure 4.10), and iii) a structural sensitivity measure (ΔΔG) as shown in Table 4.15 and Figure 4.11. I found that preferred codons in *rbcL* tend to be associated with conserved and buried sites (8/9 of 18 amino acid show association signals, in conserved/buried sites, respectively). Although in both cases, 6/5 (conserved/buried, respectively) of 18 amino acids, show non-significant results, this can be attributed to lack of statistical power rather than suggesting a biological effect, as noted in Results. However, for structurally important sites, only 6 of 18 amino acids show statistically significant association with preferred codons, whereas 7 of 18 amino acids show statistically significant association with non-preferred codons. This appears perplexing, as there is a reasonable signal (Figure 4.12) in the form of the overall odds ratio for the association of preferred codons with conserved, buried and structurally important sites in agreement with the results of Akashi (1994) and Zhou et al. (2009). An explanation for this anomaly could be, as noted by Zhou et al. (2009) in the context of their study, that a significant proportion of Rubisco-LSU sites under translational-accuracy selection are functionally important rather than structurally important and that the criterion of sequence conservation/solvent accessibility accurately identifies these sites. Indeed, there is good evidence that evolutionary sequence conservation/solvent accessibility in proteins reflect functional constraints (Lichtarge et al., 1996, Landgraf et al., 2001, Engelen et al., 2009, Goldman et al., 1998, Bustamante et al., 2000, Bloom et al., 2006).

4.5 Conclusion

The current study has thoroughly investigated codon-usage bias of *rbcL*. Several novel insights have been gained from analysis of the *rbcL* and chloroplast nucleotide sequence data. Based on sequence data from all available Angiosperm chloroplast genomes and their corresponding *rbcL* genes, it has been conclusively shown that both *rbcL* genes and chloroplast genomes have obvious A+T bias. The evidence presented here also supports a role for codon adaptation in *rbcL*, although it is limited to the two-fold codon degenerate amino acids Cys, Phe, His, Asn and Tyr and the three-fold codon degenerate amino acid Ile. For the first time, it has been shown that catalytic residues in the Rubisco-LSU utilize preferred codons, which could be to ensure greater fidelity in translation of these codons. The exploration of secondary structure preferences of codons in *rbcL* resulted in discovery of significant codon bias for different secondary structures of Rubisco-LSU. Importantly, findings of this study hint at translational accuracy selection in *rbcL*, as preferred codons in *rbcL* show statistically significant associations with conserved and buried sites in Rubisco-LSU, thus linking translational fidelity with synonymous codon-usage of *rbcL*.

Summary and Conclusion

The focus of studies carried out in this thesis was to systematically analyze the natural variation in Angiosperm Rubisco sequences in order to uncover the functional implications of these variations.

As a first step towards this objective, I created a relational database which archives data on Rubisco's sequence/kinetic/structure and taxonomy that can be accessed programmatically with a set of python modules. This local repository contains more than 11,000 unique Rubisco LSU protein/*rbcL* nucleotide sequence entries from Angiosperms; kinetic data information from 40 species, including 11 species from flowering plants; and structural information from 49 PDB structures including spinach, tobacco and rice from flowering plants.

This database facilitated the consolidation of available information on Rubisco from public domain resources and was very useful for my studies on Rubicso's coevolution and codon usage bias of *rbcL*.

The coevolution studies are based on the covarion hypothesis of molecular evolution. This proposes that selective pressures on a given amino acid site in any protein are dependent on the identity of other sites in the protein. Applying the covarion hypothesis to Rubisco, this implies that any mutation in Rubisco has to be optimized in the context of functional and/or structural constraints in Rubisco hexadecamer complex, as well as by its interactions with RA and RbcX. Thus, at any given point of time, the currently observed sequence variations in Rubisco LSU will have persisted through these optimizations and, in this process, may have influenced evolution of other sites in the Rubisco holoenzyme and in its interacting partners. To detect these correlated changes, Rubisco's coevolution has been studied using protein sequences from 5052 Rubisco large subunits, 44 Rubisco small subunits, and 14 and 23 sequences respectively, of its interacting partners chaperonins RbcX and RA.

The major findings were:

- Identification of a novel cluster of coevolving sites spatially proximal to loop 6 and in the C- terminal tail of Rubisco large subunit. This finding suggests that residues in loop 6 and the C-terminal tail are coevolving and any mutation in either region needs to be complemented, appropriately.
- Previous studies have shown residues 89 and 94 located in the loop between strands β-C and β-D of the N-terminal domain of the Rubisco LSU and residues 311 and 314 from RA to be involved in Rubisco-RA interaction. In addition to predicting these sites, my inter-molecular coevolution analysis of Rubisco LSU and RA has detected several new coevolving sites both in the Rubisco LSU and in RA. In the Rubisco LSU, these sites are located in the same β-C-β-D loop region, along with a network of polar/charged residues in the C-terminal domain of the Rubisco LSU. The surface, spatial locations of the predicted sites in the Rubisco LSU make them likely targets of RA interaction. These predictions could be experimentally tested, and together with the recently resolved structure of RA, can help in understanding the molecular basis of Rubisco-RA interaction.
- Consistent with the highly conserved interaction interfaces between Rubisco- LSU and RbcX and Rubisco-LSU and Rubisco-SSU, inter-molecular analyses resulted in identification of very few coevolving sites.

My last chapter dealt with studies on codon usage of *rbcL* using 4944 *rbcL* sequences, that covered ~96% of flowering plant orders and ~70% of families' *sensu* Angiosperm Phylogeny Group III, a large data resource not available to earlier researchers. Codon- usage studies on *rbcL* provide a handle to analyze synonymous variations in Rubisco sequences at nucleotide level; these variations do not change the amino acid at the protein level but may have an effect on its translational efficiency. With this background, the primary focus of my study was to investigate codon usage in *rbcL* within the context of the 3D structure of the Rubisco-LSU. To facilitate this analysis, I defined preferred codons in *rbcL* as those which occurred more frequently in *rbcL* than other synonymous codons for the same amino acid. In addition, I compiled codon-usage statistics for all protein-coding genes in all available Angiosperm chloroplast genomes in the public domain to find out if there are differential patterns of codon usage between *rbcL* and other genes in the chloroplast.

Inferences from the codon-usage study on *rbcL* are:

- In accordance with previous studies which were done on a smaller dataset, the consolidated data from 132 chloroplast genomes and their corresponding *rbcL* genes show conclusively that both the *rbcL* gene and chloroplast genomes have obvious A+T bias.
- As found earlier for another chloroplast gene, *psbA*, evidence found in my study also points to an important role of codon adaptation in *rbcL*, *albeit* it is limited to the two-fold degenerate amino acids Cys, Phe, His, Asn and Tyr and the three-fold degenerate Ile.
- For the first time, I have shown that the catalytic residues in the Rubisco-LSU utilize preferred codons. This could be to ensure greater fidelity in translation of these codons, as any errors in translation in these sites would likely compromise the activity of the Rubisco.
- The secondary structure preferences of codons in *rbcL* have been surveyed and codon preferences of different secondary structures in the Rubisco LSU have been discovered.
- Preferred codons in *rbcL* show statistically significant associations with conserved and buried sites in the Rubisco-LSU. These findings provide the link between translation fidelity and synonymous codon usage, thereby suggesting a role for translational-accuracy selection in *rbcL*.

In summary, *In- silico* analysis of sequence variations in the Rubisco-LSU have extended our knowledge of Rubisco's structure and function and also resulted in several experimentally testable predictions. I hope these predictions will be taken up by researchers in my supervisor's lab and elsewhere, and lead to extended knowledge on Rubisco.

List of references

ADZHUBEI, A. A., ADZHUBEI, I. A., KRASHENINNIKOV, I. A. & NEIDLE, S. 1996. Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett,* 399, 78-82.

AFONNIKOV, D. A., OSHCHEPKOV, D. Y. & KOLCHANOV, N. A. 2001. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. *Bioinformatics,* 17, 1035-46.

AKASHI, H. 1994. Synonymous Codon Usage in *Drosophila Melanogaster* - Natural-Selection and Translational Accuracy. *Genetics,* 136, 927-935.

AKASHI, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics,* 139, 1067-76.

AKASHI, H. & EYRE-WALKER, A. 1998a. Translational selection and molecular evolution. *Current Opinion in Genetics & Development,* 8, 688-693.

AKASHI, H. & EYRE-WALKER, A. 1998b. Translational selection and molecular evolution. *Curr Opin Genet Dev,* 8, 688-93.

ALBERT, V. A., BACKLUND, A., BREMER, K., CHASE, M. W., MANHART, J. R., MISHLER, B. D. & NIXON, K. C. 1994. Functional Constraints and *rbcL* Evidence for Land Plant Phylogeny. *Annals of the Missouri Botanical Garden,* 81, 534-567.

ALONSO, H., BLAYNEY, M. J., BECK, J. L. & WHITNEY, S. M. 2009. Substrate-induced Assembly of *Methanococcoides burtonii* D-Ribulose-1,5-bisphosphate Carboxylase/Oxygenase Dimers into Decamers. *J Biol Chem,* 284, 33876-33882.

ANDERSSON, I. 2008. Catalysis and regulation in Rubisco. *J Exp Bot,* 59, 1555-68.

ANDERSSON, I. & BACKLUND, A. 2008. Structure and function of Rubisco. *Plant Physiol Biochem,* 46, 275-91.

ANDERSSON, I., KNIGHT, S., SCHNEIDER, G., LINDQVIST, Y., LUNDQVIST, T., BRANDEN, C. I. & LORIMER, G. H. 1989. Crystal-Structure of the Active-Site of Ribulose-Bisphosphate Carboxylase. *Nature,* 337, 229-234.

ANDERSSON, I. & TAYLOR, T. C. 2003. Structural framework for catalysis and regulation in ribulose-1,5-bisphosphate carboxylase/oxygenase. *Arch Biochem Biophys,* 414, 130-40.

ANDERSSON, S. G. & KURLAND, C. G. 1990. Codon preferences in free-living microorganisms. *Microbiol Rev,* 54, 198-210.

ANDREWS, T. J. 1988. Catalysis by cyanobacterial ribulose-bisphosphate carboxylase large subunits in the complete absence of small subunits. *Journal of Biological Chemistry,* 263, 12213-9.

ANDREWS, T. J. & LORIMER, G. H. 1985. Catalytic properties of a hybrid between cyanobacterial large subunits and higher plant small subunits of ribulose bisphosphate carboxylase-oxygenase. *J Biol Chem,* 260, 4632-4636.

ATCHLEY, W. R., WOLLENBERG, K. R., FITCH, W. M., TERHALLE, W. & DRESS, A. W. 2000. Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Molecular Biology and Evolution,* 17, 164-178.

BADGER, M. R. 1980. Kinetic-properties of ribulose 1,5-bisphosphate carboxylase-oxygenase from *Anabaena variabilis*. *Arch Biochem Biophys,* 201, 247-254.

BADGER, M. R. & BEK, E. J. 2008. Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO_2 acquisition by the CBB cycle. *J Exp Bot,* 59, 1525-41.

BADGER, M. R. & COLLATZ, G. J. 1977. Studies on the kinetic mechanism of RuBP carboxylase and oxygenase reactions, with particular reference to the effect of temperature on kinetic papameters. *Carnegie YB,* 76, 355-361.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2011. GenBank. *Nucleic Acids Res,* 39, D32-7.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42.

BERNARDI, G. 1986. Compositional constraints and genome evolution. *J Mol Evol,* 24, 1-11.

BIRO, J. C. 2006. Indications that "codon boundaries" are physico-chemically defined and that protein-folding information is contained in the redundant exon bases. *Theor Biol Med Model,* 3, 28.

BLOOM, J. D., LABTHAVIKUL, S. T., OTEY, C. R. & ARNOLD, F. H. 2006. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A,* 103, 5869-74.

BRACHER, A., STARLING-WINDHOF, A., HARTL, F. U. & HAYER-HARTL, M. 2011. Crystal structure of a chaperone-bound assembly intermediate of form I Rubisco. *Nat Struct Mol Biol,* 18, 875-80.

BREMER, B., BREMER, K., CHASE, M. W., FAY, M. F., REVEAL, J. L., SOLTIS, D. E., SOLTIS, P. S., STEVENS, P. F., ANDERBERG, A. A., MOORE, M. J., OLMSTEAD, R. G., RUDALL, P. J., SYTSMA, K. J., TANK, D. C., WURDACK, K., XIANG, J. Q. Y., ZMARZTY, S. & GRP, A. P. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society,* 161, 105-121.

BRUNAK, S. & ENGELBRECHT, J. 1996. Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. *Proteins,* 25, 237-52.

BUCHAN, J. R. & STANSFIELD, I. 2007. Halting a cellular production line: responses to ribosomal pausing during translation. *Biol Cell,* 99, 475-87.

BULMER, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics,* 129, 897-907.

BUSTAMANTE, C. D., TOWNSEND, J. P. & HARTL, D. L. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol Biol Evol,* 17, 301-8.

CAPORASO, J. G., SMIT, S., EASTON, B. C., HUNTER, L., HUTTLEY, G. A. & KNIGHT, R. 2008. Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC Evol Biol,* 8, 327.

CHEN, Z. X. & SPREITZER, R. J. 1989. Chloroplast intragenic suppression enhances the low CO_2/O_2 specificity of mutant ribulose-bisphosphate carboxylase/oxygenase. *Journal of Biological Chemistry,* 264, 3051-3.

CHEN, Z. X., YU, W. Z., LEE, J. H., DIAO, R. & SPREITZER, R. J. 1991. Complementing amino acid substitutions within loop 6 of the alpha/beta-barrel active site influence the

CO_2/O_2 specificity of chloroplast ribulose-1,5-bisphosphate carboxylase/oxygenase. *Biochemistry,* 30**,** 8846-50.

CHRISTIN, P.-A., SALAMIN, N., MUASYA, A., ROALSON, E., RUSSIER, F. & BESNARD, G. 2008. Evolutionary switch and genetic convergence on rbcL following the evolution of C_4 photosynthesis. *Molecular biology and evolution,* 25**,** 2361-2369.

CLEGG, M. T. 1993. Chloroplast gene sequences and the study of plant evolution. *Proc Natl Acad Sci U S A,* 90**,** 363-7.

CLELAND, W. W., ANDREWS, T. J., GUTTERIDGE, S., HARTMAN, F. C. & LORIMER, G. H. 1998. Mechanism of Rubisco: The Carbamate as General Base. *Chem Rev,* 98**,** 549-562.

COCK, P. J., ANTAO, T., CHANG, J. T., CHAPMAN, B. A., COX, C. J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNSKI, B. & DE HOON, M. J. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* 25**,** 1422-3.

COMERON, J. M. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics,* 167**,** 1293-304.

CORTAZZO, P., CERVENANSKY, C., MARIN, M., REISS, C., EHRLICH, R. & DEANA, A. 2002. Silent mutations affect in vivo protein folding in Escherichia coli. *Biochem Biophys Res Commun,* 293**,** 537-41.

CURMI, P. M. G., CASCIO, D., SWEET, R. M., EISENBERG, D. & SCHREUDER, H. 1992. Crystal-Structure of the Unactivated Form of Ribulose-1,5-Bisphosphate Carboxylase Oxygenase from Tobacco Refined at 2.0-a-Circle Resolution. *Journal of Biological Chemistry,* 267**,** 16980-16989.

DARWIN, C. R. 1862. On the Various Contrivances by which British and Foreign Orchids are Fertilised by Insects, and on the Good Effects of Intercrossing. *London: John Murray*.

DEANE, C. M. & SAUNDERS, R. 2011. The imprint of codons on protein structure. *Biotechnol J,* 6**,** 641-9.

DRUMMOND, D. A. & WILKE, C. O. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell,* 134**,** 341-52.

DUFF, A. P., ANDREWS, T. J. & CURMI, P. M. 2000. The transition between the open and closed states of rubisco is triggered by the inter-phosphate distance of the bound bisphosphate. *J Mol Biol,* 298**,** 903-16.

DUNN, S. D., WAHL, L. M. & GLOOR, G. B. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics,* 24**,** 333-40.

DURET, L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev,* 12**,** 640-9.

DURET, L. & MOUCHIROUD, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci U S A,* 96**,** 4482-7.

DUTHEIL, J., PUPKO, T., JEAN-MARIE, A. & GALTIER, N. 2005. A model-based approach for detecting coevolving positions in a molecule. *Molecular Biology and Evolution,* 22**,** 1919-28.

DUTHEIL, J. Y. 2011. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform*.

EASTON, B. C. 2006. Novel techniques for detecting correlated evolution. *In PhD thesis Australian National University; 2006.*

EHRLICH, P. R. & RAVEN, P. H. 1964. Butterflies and Plants - a Study in Coevolution. *Evolution,* 18**,** 586-608.

ELLIS, R. J. 1979. Most Abundant Protein in the World. *Trends in Biochemical Sciences,* 4**,** 241-244.

ENGELEN, S., TROJAN, L. A., SACQUIN-MORA, S., LAVERY, R. & CARBONE, A. 2009. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput Biol,* 5**,** e1000267.

ESQUIVEL, M. G., ANWARUZZAMAN, M. & SPREITZER, R. J. 2002. Deletion of nine carboxy-terminal residues of the Rubisco small subunit decreases thermal stability but does not eliminate function. *FEBS Lett,* 520**,** 73-6.

FARES, M. A., RUIZ-GONZALEZ, M. X. & LABRADOR, J. P. 2011. Protein Coadaptation and the Design of Novel Approaches to Identify Protein-Protein Interactions. *Iubmb Life,* 63**,** 264-271.

FEDERHEN, S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res,* 40**,** D136-43.

FITCH, W. M. & MARKOWITZ, E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet,* 4**,** 579-93.

FODOR, A. A. & ALDRICH, R. W. 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins,* 56**,** 211-21.

FRACZKIEWICZ, R. & BRAUN, W. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry,* 19**,** 319-333.

FUKAMI-KOBAYASHI, K., SCHREIBER, D. R. & BENNER, S. A. 2002. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J Mol Biol,* 319**,** 729-43.

FUTUYMA, D. J. 1997. Evolutionary Biology. *Stamford, Connecticut: Sinauer Associates*.

GALMES, J., FLEXAS, J., KEYS, A. J., CIFRE, J., MITCHELL, R. A. C., MADGWICK, P. J., HASLAM, R. P., MEDRANO, H. & PARRY, M. A. J. 2005. Rubisco specificity factor tends to be larger in plant species from drier habitats and in species with persistent leaves. *Plant, Cell & Environment,* 28**,** 571-579.

GENKOV, T., MEYER, M., GRIFFITHS, H. & SPREITZER, R. J. 2010. Functional Hybrid Rubisco Enzymes with Plant Small Subunits and Algal Large Subunits engineered rbcS cDNA for expression in *Chlamydomonas*. *J Biol Chem,* 285**,** 19833-19841.

GLOOR, G. B., MARTIN, L. C., WAHL, L. M. & DUNN, S. D. 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry,* 44**,** 7156-65.

GOBEL, U., SANDER, C., SCHNEIDER, R. & VALENCIA, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins,* 18**,** 309-17.

GOLDMAN, N., THORNE, J. L. & JONES, D. T. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics,* 149**,** 445-58.

GOTO, N., PRINS, P., NAKAO, M., BONNAL, R., AERTS, J. & KATAYAMA, T. 2010. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics,* 26**,** 2617-9.

GRANTHAM, R., GAUTIER, C., GOUY, M., JACOBZONE, M. & MERCIER, R. 1981. Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity. *Nucleic Acids Research,* 9**,** R43-R74.

GRANTHAM, R., GAUTIER, C., GOUY, M., MERCIER, R. & PAVE, A. 1980. Codon Catalog Usage and the Genome Hypothesis. *Nucleic Acids Research,* 8**,** R49-R62.

GRANTHAM, R., PERRIN, P. & MOUCHIROUD, D. 1986. Patterns in codon usage of different kinds of species. *Oxford Surveys in Evolutionary Biology,* 3**,** 48-81.

GU, W. J., ZHOU, T., MA, J. M., SUN, X. & LU, Z. H. 2003. Folding type specific secondary structure propensities of synonymous codons. *Ieee Transactions on Nanobioscience,* 2**,** 150-157.

GUPTA, S. K., MAJUMDAR, S., BHATTACHARYA, T. K. & GHOSH, T. C. 2000. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem Biophys Res Commun,* 269**,** 692-6.

GUTTERIDGE, S. 1991. The relative catalytic specificities of the large subunit core of *Synechococcus* ribulose bisphosphate carboxylase/oxygenase. *Journal of Biological Chemistry,* 266**,** 7359-62.

GUTTERIDGE, S., RHOADES, D. F. & HERRMANN, C. 1993a. Site-specific mutations in a loop region of the C-terminal domain of the large subunit of ribulose bisphosphate carboxylase/oxygenase that influence substrate partitioning. *J Biol Chem,* 268**,** 7818-24.

GUTTERIDGE, S., RHOADES, D. F. & HERRMANN, C. 1993b. Site-specific mutations in a loop region of the C-terminal domain of the large subunit of ribulose bisphosphate carboxylase/oxygenase that influence substrate partitioning. *Journal of Biological Chemistry,* 268**,** 7818-24.

HAFNER, M. S. & NADLER, S. A. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature,* 332**,** 258-9.

HAMANO, T., MATSUO, K., HIBI, Y., VICTORIANO, A. F., TAKAHASHI, N., MABUCHI, Y., SOJI, T., IRIE, S., SAWANPANYALERT, P., YANAI, H., HARA, T., YAMAZAKI, S., YAMAMOTO, N. & OKAMOTO, T. 2007. A single-nucleotide synonymous mutation in the gag gene controlling human immunodeficiency virus type 1 virion production. *J Virol,* 81**,** 1528-33.

HANSEN, S., VOLLAN, V. B., HOUGH, E. & ANDERSEN, K. 1999. The crystal structure of rubisco from *Alcaligenes eutrophus* reveals a novel central eight-stranded beta-barrel formed by beta-strands from four subunits. *J Mol Biol,* 288**,** 609-21.

HANSON, P. I. & WHITEHEART, S. W. 2005. AAA+ proteins: have engine, will work. *Nat Rev Mol Cell Biol,* 6**,** 519-29.

HANSON, T. E. & TABITA, F. R. 2001. A ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco)-like protein from *Chlorobium tepidum* that is involved with sulfur metabolism and the response to oxidative stress. *Proc Natl Acad Sci U S A,* 98**,** 4397-402.

HARTMAN, F. C. & HARPEL, M. R. 1994. Structure, function, regulation, and assembly of D-ribulose-1,5-bisphosphate carboxylase/oxygenase. *Annu Rev Biochem,* 63, 197-234.

HOLLAND, R. C., DOWN, T. A., POCOCK, M., PRLIC, A., HUEN, D., JAMES, K., FOISY, S., DRAGER, A., YATES, A., HEUER, M. & SCHREIBER, M. J. 2008. BioJava: an open-source framework for bioinformatics. *Bioinformatics,* 24, 2096-7.

HORNER, D. S., PIROVANO, W. & PESOLE, G. 2008. Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief Bioinform,* 9, 46-56.

HOUTZ, R. L., MAGNANI, R., NAYAK, N. R. & DIRK, L. M. 2008. Co- and post-translational modifications in Rubisco: unanswered questions. *J Exp Bot,* 59, 1635-45.

HOUTZ, R. L. & PORTIS, A. R., JR. 2003. The life of ribulose 1,5-bisphosphate carboxylase/oxygenase--posttranslational facts and mysteries. *Arch Biochem Biophys,* 414, 150-8.

IKEMURA, T. 1981. Correlation between the Abundance of Escherichia-Coli Transfer-RNAs and the Occurrence of the Respective Codons in Its Protein Genes. *Journal of Molecular Biology,* 146, 1-21.

IKEMURA, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol,* 2, 13-34.

JANIN, J., MILLER, S. & CHOTHIA, C. 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol,* 204, 155-64.

JORDAN, D. B. & CHOLLET, R. 1985. Subunit dissociation and reconstitution of ribulose-1,5-bisphosphate carboxylase from *Chromatium vinosum*. *Arch Biochem Biophys,* 236, 487-496.

JORDAN, D. B. & OGREN, W. L. 1981. Species variation in the specificity of ribulose bisphosphate carboxylase/oxygenase. *Nature,* 291, 513-515.

KANE, H. J., VIIL, J., ENTSCH, B., PAUL, K., MORELL, M. K. & ANDREWS, T. J. 1994. An Improved Method for Measuring the CO_2/O_2 Specificity of Ribulose-bisphosphate Carboxylase-Oxygenase. *Australian Journal of Plant Physiology,* 21 449 - 461.

KANNAPPAN, B. & GREADY, J. E. 2008. Redefinition of rubisco carboxylase reaction reveals origin of water for hydration and new roles for active-site residues. *J Am Chem Soc,* 130, 15063-80.

KAPLAN, A. & REINHOLD, L. 1999. Co2 Concentrating Mechanisms in Photosynthetic Microorganisms. *Annu Rev Plant Physiol Plant Mol Biol,* 50, 539-570.

KAPRALOV, M. & FILATOV, D. 2006. Molecular adaptation during adaptive radiation in the Hawaiian endemic genus *Schiedea*. *PloS one,* 1.

KAPRALOV, M., KUBIEN, D., ANDERSSON, I. & FILATOV, D. 2011. Changes in Rubisco kinetics during the evolution of C_4 photosynthesis in *Flaveria* (*Asteraceae*) are associated with positive selection on genes encoding the enzyme. *Molecular biology and evolution,* 28, 1491-1994.

KAPRALOV, M. V. & FILATOV, D. A. 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol Biol,* 7, 73.

KASS, I. & HOROVITZ, A. 2002. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins,* 48, 611-7.

KELLOGG, E. & JULIANO, N. 1997. The structure and function of Rubisco and their implications for systematic studies. *Am J Bot,* 84, 413.

KEPES, F. 1996. The "+70 pause": hypothesis of a translational control of membrane protein assembly. *Journal of Molecular Biology,* 262**,** 77-86.

KESKIN, O., MA, B. & NUSSINOV, R. 2005. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol,* 345**,** 1281-94.

KIM, J., KLEIN, P. G. & MULLET, J. E. 1991. Ribosomes pause at specific sites during synthesis of membrane-bound chloroplast reaction center protein D1. *J Biol Chem,* 266**,** 14931-8.

KIMCHI-SARFATY, C., OH, J. M., KIM, I. W., SAUNA, Z. E., CALCAGNO, A. M., AMBUDKAR, S. V. & GOTTESMAN, M. M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science,* 315**,** 525-8.

KIMURA, M. 1983. The Neutral Theory of Molecular Evolution. *Cambridge University Press, New York.*

KING, W. A., GREADY, J. E. & ANDREWS, T. J. 1998. Quantum chemical analysis of the enolization of ribulose bisphosphate: the first hurdle in the fixation of CO_2 by Rubisco. *Biochemistry,* 37**,** 15414-22.

KITANO, K., MAEDA, N., FUKUI, T., ATOMI, H., IMANAKA, T. & MIKI, K. 2001. Crystal structure of a novel-type archaeal rubisco with pentagonal symmetry. *Structure,* 9**,** 473-81.

KNIGHT, R., MAXWELL, P., BIRMINGHAM, A., CARNES, J., CAPORASO, J. G., EASTON, B. C., EATON, M., HAMADY, M., LINDSAY, H., LIU, Z., LOZUPONE, C., MCDONALD, D., ROBESON, M., SAMMUT, R., SMIT, S., WAKEFIELD, M. J., WIDMANN, J., WIKMAN, S., WILSON, S., YING, H. & HUTTLEY, G. A. 2007. PyCogent: a toolkit for making sense from sequence. *Genome Biol,* 8**,** R171.

KNIGHT, S., ANDERSSON, I. & BRANDEN, C. I. 1990. Crystallographic analysis of ribulose 1,5-bisphosphate carboxylase from spinach at 2.4 A resolution. Subunit interactions and active site. *J Mol Biol,* 215**,** 113-60.

KOMAR, A. A. 2007. Genetics. SNPs, silent but not invisible. *Science,* 315**,** 466-7.

KOMAR, A. A., LESNIK, T. & REISS, C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett,* 462**,** 387-91.

KORTEMME, T. & BAKER, D. 2002. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A,* 99**,** 14116-21.

KORTEMME, T., KIM, D. E. & BAKER, D. 2004. Computational alanine scanning of protein-protein interfaces. *Sci STKE,* 2004**,** pl2.

KUBIEN, D. S., WHITNEY, S. M., MOORE, P. V. & JESSON, L. K. 2008. The biochemistry of Rubisco in *Flaveria*. *J Exp Bot,* 59**,** 1767-1777.

KULIKOVA, T., AKHTAR, R., ALDEBERT, P., ALTHORPE, N., ANDERSSON, M., BALDWIN, A., BATES, K., BHATTACHARYYA, S., BOWER, L., BROWNE, P., CASTRO, M., COCHRANE, G., DUGGAN, K., EBERHARDT, R., FARUQUE, N., HOAD, G., KANZ, C., LEE, C., LEINONEN, R., LIN, Q., LOMBARD, V., LOPEZ, R., LORENC, D., MCWILLIAM, H., MUKHERJEE, G., NARDONE, F., PASTOR, M. P., PLAISTER, S., SOBHANY, S., STOEHR, P., VAUGHAN, R., WU, D., ZHU, W. & APWEILER, R. 2007. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res,* 35**,** D16-20.

LAING, W. A., OGREN, W. L. & HAGEMAN, R. H. 1974. Regulation of soybean net photosynthetic CO_2 fixation by the interaction of CO_2, O_2, and Ribulose 1,5-Diphosphate Carboxylase. *Plant Physiol,* 54, 678-85.

LANDGRAF, R., XENARIOS, I. & EISENBERG, D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *Journal of Molecular Biology,* 307, 1487-502.

LARSON, E. M., O'BRIEN, C. M., ZHU, G., SPREITZER, R. J. & PORTIS, A. R., JR. 1997. Specificity for activase is changed by a Pro-89 to Arg substitution in the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase. *Journal of Biological Chemistry,* 272, 17033-7.

LARSON, S. M., DI NARDO, A. A. & DAVIDSON, A. R. 2000. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J Mol Biol,* 303, 433-46.

LAVNER, Y. & KOTLAR, D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene,* 345, 127-138.

LEE, B. G., READ, B. A. & TABITA, F. R. 1991. Catalytic properties of recombinant octameric, hexadecameric, and heterologous cyanobacterial/bacterial ribulose- 1,5-bisphosphate carboxylase/oxygenase. *Arch Biochem Biophys,* 291, 263-9.

LI, C., SALVUCCI, M. E. & PORTIS, A. R., JR. 2005a. Two residues of Rubisco activase involved in recognition of the Rubisco substrate. *J Biol Chem,* 280, 24864-9.

LI, H., SAWAYA, M. R., TABITA, F. R. & EISENBERG, D. 2005b. Crystal structure of a Rubisco-like protein from the green sulfur bacterium *Chlorobium tepidum*. *Structure,* 13, 779-89.

LI, K. B. 2003. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics,* 19, 1585-6.

LI, W. H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of Molecular Evolution,* 24, 337-345.

LICHTARGE, O., BOURNE, H. R. & COHEN, F. E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology,* 257, 342-58.

LIU, C., YOUNG, A. L., STARLING-WINDHOF, A., BRACHER, A., SASCHENBRECKER, S., RAO, B. V., RAO, K. V., BERNINGHAUSEN, O., MIELKE, T., HARTL, F. U., BECKMANN, R. & HAYER-HARTL, M. 2010. Coupled chaperone action in folding and assembly of hexadecameric Rubisco. *Nature,* 463, 197-202.

LOCKLESS, S. W. & RANGANATHAN, R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science,* 286, 295-9.

LORIMER, G. H. & MIZIORKO, H. M. 1980. Carbamate formation on the epsilon-amino group of a lysyl residue as the basis for the activation of ribulose-bisphosphate carboxylase by CO_2 and Mg^{2+}. *Biochemistry,* 19, 5321-8.

MAKHOUL, C. H. & TRIFONOV, E. N. 2002. Distribution of rare triplets along mRNA and their relation to protein folding. *J Biomol Struct Dyn,* 20, 413-20.

MANGALAM, H. 2002. The Bio* toolkits--a brief overview. *Brief Bioinform,* 3, 296-302.

MANTEL, N. 1963. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc.,* 58, 690-700.

MANTEL, N. & HAENSZEL, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst,* 22**,** 719-48.

MARTIN, L. C., GLOOR, G. B., DUNN, S. D. & WAHL, L. M. 2005. Using information theory to search for co-evolving residues in proteins. *Bioinformatics,* 21**,** 4116-24.

MATSUMURA, H., MIZOHATA, E., ISHIDA, H., KOGAMI, A., UENO, T., MAKINO, A., INOUE, T., YOKOTA, A., MAE, T. & KAI, Y. 2012. Crystal structure of rice Rubisco and implications for activation induced by positive effectors NADPH and 6-phosphogluconate. *J Mol Biol.*

MATSUOKA, M., FURBANK, R. T., FUKAYAMA, H. & MIYAO, M. 2001. Molecular Engineering of C_4 Photosynthesis. *Annu Rev Plant Physiol Plant Mol Biol,* 52**,** 297-314.

MAUSER, H., KING, W. A., GREADY, J. E. & ANDREWS, T. J. 2001. CO_2 fixation by Rubisco: computational dissection of the key steps of carboxylation, hydration, and C-C bond cleavage. *J Am Chem Soc,* 123**,** 10821-9.

MIZOHATA, E., MATSUMURA, H., OKANO, Y., KUMEI, M., TAKUMA, H., ONODERA, J., KATO, K., SHIBATA, N., INOUE, T., YOKOTA, A. & KAI, Y. 2002. Crystal structure of activated ribulose-1,5-bisphosphate carboxylase/oxygenase from green alga *Chlamydomonas reinhardtii* complexed with 2-carboxyarabinitol-1,5-bisphosphate. *J Mol Biol,* 316**,** 679-91.

MORELL, M. K., KANE, H. J. & ANDREWS, T. J. 1990. Carboxylterminal deletion mutants of ribulose-bisphosphate carboxylase from *Rhodospirillum rubrum*. *FEBS Lett,* 265**,** 41-45.

MORELL, M. K., WILKIN, J. M., KANE, H. J. & ANDREWS, T. J. 1997. Side reactions catalyzed by ribulose-bisphosphate carboxylase in the presence and absence of small subunits. *Journal of Biological Chemistry,* 272**,** 5445-51.

MORTON, B. R. 1993. Chloroplast DNA codon use: evidence for selection at the *psbA* locus based on tRNA availability. *J Mol Evol,* 37**,** 273-80.

MORTON, B. R. 1994. Codon use and the rate of divergence of land plant chloroplast genes. *Mol Biol Evol,* 11**,** 231-8.

MORTON, B. R. 1996. Selection on the codon bias of *Chlamydomonas reinhardtii* chloroplast genes and the plant *psbA* gene. *J Mol Evol,* 43**,** 28-31.

MORTON, B. R. 1998. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J Mol Evol,* 46**,** 449-59.

MORTON, B. R. 2000. Codon bias and the context dependency of nucleotide substitutions in the evolution of plastid DNA. *Evolutionary Biology, Vol 31,* 31**,** 55-103.

MORTON, B. R. 2001. Selection at the amino acid level can influence synonymous codon usage: implications for the study of codon adaptation in plastid genes. *Genetics,* 159**,** 347-58.

MORTON, B. R. & LEVIN, J. A. 1997. The atypical codon usage of the plant *psbA* gene may be the remnant of an ancestral bias. *Proc Natl Acad Sci U S A,* 94**,** 11434-8.

MOYA, A., PERETO, J., GIL, R. & LATORRE, A. 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat Rev Genet,* 9**,** 218-29.

MUELLER-CAJAR, O., STOTZ, M., WENDLER, P., HARTL, F. U., BRACHER, A. & HAYER-HARTL, M. 2011. Structure and function of the AAA+ protein CbbX, a red-type Rubisco activase. *Nature,* 479**,** 194-9.

MUELLER-CAJAR, O. & WHITNEY, S. M. 2008. Evolving improved *Synechococcus* Rubisco functional expression in *Escherichia coli. Biochem J,* 414**,** 205-14.

MURZIN, A. G., BRENNER, S. E., HUBBARD, T. & CHOTHIA, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology,* 247**,** 536-40.

NEWMAN, J., BRANDEN, C. I. & JONES, T. A. 1993. Structure determination and refinement of ribulose 1,5-bisphosphate carboxylase/oxygenase from *Synechococcus PCC6301*. *Acta Crystallogr D Biol Crystallogr,* 49**,** 548-60.

NEWMAN, J. & GUTTERIDGE, S. 1993. The X-ray structure of *Synechococcus* ribulose-bisphosphate carboxylase/oxygenase-activated quaternary complex at 2.2-A resolution. *Journal of Biological Chemistry,* 268**,** 25876-86.

NEWMAN, J. & GUTTERIDGE, S. 1994. Structure of an effector-induced inactivated state of ribulose 1,5-bisphosphate carboxylase/oxygenase: the binary complex between enzyme and xylulose 1,5-bisphosphate. *Structure,* 2**,** 495-502.

NOIVIRT, O., EISENSTEIN, M. & HOROVITZ, A. 2005. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng Des Sel,* 18**,** 247-53.

OLMEA, O. & VALENCIA, A. 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des,* 2**,** S25-32.

OTT, C. M., SMITH, B. D., PORTIS, A. R. & SPREITZER, R. J. 2000. Activase region on chloroplast ribulose-1,5-bisphosphate carboxylase/oxygenase - Nonconservative substitution in the large subunit alters species specificity of protein interaction. *Journal of Biological Chemistry,* 275**,** 26241-26244.

PARRY, M. A. J., MADGWICK, P., PARMAR, S., CORNELIUS, M. J. & KEYS, A. J. 1992. Mutations in loop six of the large subunit of ribulose-1,5-bisphosphate carboxylase affect substrate specificity. *Planta* 187**,** 109–12.

PARRY, M. A. J., MADGWICK, P. J., CARVALHO, J. F. C. & ANDRALOJC, P. J. 2007. Prospects for increasing photosynthesis by overcoming the limitations of Rubisco. *Journal of Agricultural Science,* 145**,** 31-43.

PAZOS, F., HELMER-CITTERICH, M., AUSIELLO, G. & VALENCIA, A. 1997. Correlated mutations contain information about protein-protein interaction. *J Mol Biol,* 271**,** 511-23.

PETERHANSEL, C., NIESSEN, M. & KEBEISH, R. M. 2008. Metabolic engineering towards the enhancement of photosynthesis. *Photochem Photobiol,* 84**,** 1317-23.

PFITZINGER, H., WEIL, J. H., PILLAY, D. T. N. & GUILLEMAUT, P. 1990. Codon Recognition Mechanisms in Plant Chloroplasts. *Plant Molecular Biology,* 14**,** 805-814.

POLLOCK, D. D. & TAYLOR, W. R. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng,* 10**,** 647-57.

POLLOCK, D. D., TAYLOR, W. R. & GOLDMAN, N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol,* 287**,** 187-98.

PORTIS, A. R. 1992. Regulation of ribulose 1,5-bisphosphate carboxylase oxygenase Activity. *Annual Review of Plant Physiology and Plant Molecular Biology,* 43**,** 415-437.

PORTIS, A. R. 1995. The Regulation of Rubisco by Rubisco activase. *Journal of Experimental Botany,* 46, 1285-1291.

PORTIS, A. R., JR. 2003. Rubisco activase - Rubisco's catalytic chaperone. *Photosynth Res,* 75, 11-27.

R DEVELOPMENT CORE TEAM 2008. R: a language and environment for statistical computing. . *Vienna (Austria): R Foundation for Statistical Computing.,* ISBN 3-900051-07-0.

RAINES, C. A. 2006. Transgenic approaches to manipulate the environmental responses of the C_3 carbon fixation cycle. *Plant Cell Environ,* 29, 331-9.

RANTY, B., LUNDQVIST, T., SCHNEIDER, G., MADDEN, M., HOWARD, R. & LORIMER, G. 1990. Truncation of ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) from *Rhodospirillum rubrum* affects the holoenzyme assembly and activity. *EMBO J,* 9, 1365-73.

RAUNSER, S., MAGNANI, R., HUANG, Z., HOUTZ, R. L., TRIEVEL, R. C., PENCZEK, P. A. & WALZ, T. 2009. Rubisco in complex with Rubisco large subunit methyltransferase. *Proc Natl Acad Sci U S A,* 106, 3160-5.

READ, B. A. & TABITA, F. R. 1994. High substrate specificity factor ribulose bisphosphate carboxylase oxygenase from eukaryotic marine algae and properties of recombinant cyanobacterial rubisco containing algal residue modifications. *Arch Biochem Biophys,* 312, 210-218.

ROBINSON, J. J., SCOTT, K. M., SWANSON, S. T., O'LEARY, M. H., HORKEN, K., TABITA, F. R. & CAVANAUGH, C. M. 2003. Kinetic isotope effect and characterization of form II Rubisco from the chemoautotrophic endosymbionts of the hydrothermal vent tubeworm *Riftia pachyptila*. *Limnol and Oceanog,* 48, 48-54.

ROGALSKI, M., KARCHER, D. & BOCK, R. 2008. Superwobbling facilitates translation with reduced tRNA sets. *Nature Structural & Molecular Biology,* 15, 192-198.

SAGE, R. F. 2002. Variation in the k_{cat} of Rubisco in C_3 and C_4 plants and some implications for photosynthetic performance at high and low temperature. *J Exp Bot,* 53, 609-20.

SAGE, R. F. & SEEMANN, J. R. 1993. Regulation of rribulose-1,5-bisphosphate carboxylase/oxygenase activity in response to reduced light intensity in C_4 plants. *Plant Physiol.,* 102, 21-28.

SARAF, M. C., MOORE, G. L. & MARANAS, C. D. 2003. Using multiple sequence correlation analysis to characterize functionally important protein regions. *Protein Eng,* 16, 397-406.

SASCHENBRECKER, S., BRACHER, A., RAO, K. V., RAO, B. V., HARTL, F. U. & HAYER-HARTL, M. 2007. Structure and function of RbcX, an assembly chaperone for hexadecameric rubisco. *Cell,* 129, 1189-1200.

SAUNDERS, R. & DEANE, C. M. 2010. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Research,* 38, 6719-28.

SAYERS, E. W., BARRETT, T., BENSON, D. A., BOLTON, E., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., FEDERHEN, S., FEOLO, M., FINGERMAN, I. M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., KRASNOV, S., LANDSMAN, D., LIPMAN, D. J., LU, Z., MADDEN, T. L., MADEJ, T., MAGLOTT, D. R., MARCHLER-BAUER, A., MILLER, V., KARSCH-MIZRACHI, I., OSTELL, J., PANCHENKO, A., PHAN, L., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T., SHUMWAY, M., SIROTKIN, K., SLOTTA, D., SOUVOROV, A., STARCHENKO, G., TATUSOVA, T. A., WAGNER, L., WANG, Y., WILBUR, W. J., YASCHENKO, E. & YE, J. 2012. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 40, D13-25.

SCHNEIDER, G., LINDQVIST, Y., BRANDEN, C. I. & LORIMER, G. 1986. Three-dimensional structure of ribulose-1,5-bisphosphate carboxylase/oxygenase from *Rhodospirillum rubrum* at 2.9 A resolution. *EMBO J,* 5, 3409-15.

SCHREUDER, H. A., KNIGHT, S., CURMI, P. M., ANDERSSON, I., CASCIO, D., BRANDEN, C. I. & EISENBERG, D. 1993a. Formation of the active site of ribulose-1,5-bisphosphate carboxylase/oxygenase by a disorder-order transition from the unactivated to the activated form. *Proc Natl Acad Sci U S A,* 90, 9968-72.

SCHREUDER, H. A., KNIGHT, S., CURMI, P. M., ANDERSSON, I., CASCIO, D., SWEET, R. M., BRANDEN, C. I. & EISENBERG, D. 1993b. Crystal structure of activated tobacco rubisco complexed with the reaction-intermediate analogue 2-carboxy-arabinitol 1,5-bisphosphate. *Protein Sci,* 2, 1136-46.

SCHULER, G. D., EPSTEIN, J. A., OHKAWA, H. & KANS, J. A. 1996. Entrez: molecular biology database and retrieval system. *Methods Enzymol,* 266, 141-62.

SEN, L., FARES, M. A., LIANG, B., GAO, L., WANG, B., WANG, T. & SU, Y. J. 2011. Molecular evolution of *rbcL* in three gymnosperm families: identifying adaptive and coevolutionary patterns. *Biol Direct,* 6, 29.

SHARP, P. M. 1991. Determinants of DNA-Sequence Divergence between *Escherichia-Coli* and *Salmonella-Typhimurium* - Codon Usage, Map Position, and Concerted Evolution. *Journal of Molecular Evolution,* 33, 23-33.

SHARP, P. M. & LI, W. H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol,* 24, 28-38.

SHARP, P. M., TUOHY, T. M. F. & MOSURSKI, K. R. 1986. Codon Usage in Yeast - Cluster-Analysis Clearly Differentiates Highly and Lowly Expressed Genes. *Nucleic Acids Research,* 14, 5125-5143.

SHARWOOD, R. E., VON CAEMMERER, S., MALIGA, P. & WHITNEY, S. M. 2008. The catalytic properties of hybrid Rubisco comprising tobacco small and sunflower large subunits mirror the kinetically equivalent source Rubiscos and can support tobacco growth. *Plant Physiol,* 146, 83-96.

SHEINERMAN, F. B., NOREL, R. & HONIG, B. 2000. Electrostatic aspects of protein-protein interactions. *Curr Opin Struct Biol,* 10, 153-9.

SHENKIN, P. S., ERMAN, B. & MASTRANDREA, L. D. 1991. Information-theoretical entropy as a measure of sequence variability. *Proteins,* 11, 297-313.

SHINDYALOV, I. N., KOLCHANOV, N. A. & SANDER, C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng,* 7, 349-58.

SINHA, N. & SMITH-GILL, S. J. 2002. Electrostatics in protein binding and function. *Curr Protein Pept Sci,* 3, 601-14.

SMITH, C. K., BAKER, T. A. & SAUER, R. T. 1999. Lon and Clp family proteases and chaperones share homologous substrate-recognition domains. *Proc Natl Acad Sci U S A,* 96, 6678-82.

SPREITZER, R. J. 2003. Role of the small subunit in ribulose-1,5-bisphosphate carboxylase/oxygenase. *Archives of Biochemistry and Biophysics,* 414**,** 141-149.

SPREITZER, R. J. & SALVUCCI, M. E. 2002. Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. *Annu Rev Plant Biol,* 53**,** 449-75.

STAJICH, J. E. 2007. An Introduction to BioPerl. *Methods Mol Biol,* 406**,** 535-48.

STEIN, L. 2002. Creating a bioinformatics nation. *Nature,* 417**,** 119-20.

STEIN, L. D., MUNGALL, C., SHU, S., CAUDY, M., MANGONE, M., DAY, A., NICKERSON, E., STAJICH, J. E., HARRIS, T. W., ARVA, A. & LEWIS, S. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res,* 12**,** 1599-610.

STOLETZKI, N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. *BMC Evol Biol,* 8**,** 224.

STONE, A. R., HAWKSWORTH, D. L. 1985. Coevolution and Systematics. *Oxford: Clarendon Press.*

STOTZ, M., MUELLER-CAJAR, O., CINIAWSKY, S., WENDLER, P., HARTL, F. U., BRACHER, A. & HAYER-HARTL, M. 2011. Structure of green-type Rubisco activase from tobacco. *Nat Struct Mol Biol.*

SUEL, G. M., LOCKLESS, S. W., WALL, M. A. & RANGANATHAN, R. 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol,* 10**,** 59-69.

SUGAWARA, H., OGASAWARA, O., OKUBO, K., GOJOBORI, T. & TATENO, Y. 2008. DDBJ with new system and face. *Nucleic Acids Res,* 36**,** D22-4.

SUGAWARA, H., YAMAMOTO, H., SHIBATA, N., INOUE, T., OKADA, S., MIYAKE, C., YOKOTA, A. & KAI, Y. 1999. Crystal structure of carboxylase reaction-oriented ribulose 1, 5-bisphosphate carboxylase/oxygenase from a thermophilic red alga, *Galdieria partita. Journal of Biological Chemistry,* 274**,** 15655-61.

TABITA, F. R. 1999. Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: A different perspective. *Photosynthesis Research,* 60**,** 1-28.

TABITA, F. R., HANSON, T. E., LI, H., SATAGOPAN, S., SINGH, J. & CHAN, S. 2007. Function, structure, and evolution of the Rubisco-like proteins and their Rubisco homologs. *Microbiol Mol Biol Rev,* 71**,** 576-99.

TABITA, F. R., HANSON, T. E., SATAGOPAN, S., WITTE, B. H. & KREEL, N. E. 2008a. Phylogenetic and evolutionary relationships of Rubisco and the Rubisco-like proteins and the functional lessons provided by diverse molecular forms. *Philos Trans R Soc Lond B Biol Sci,* 363**,** 2629-40.

TABITA, F. R., SATAGOPAN, S., HANSON, T. E., KREEL, N. E. & SCOTT, S. S. 2008b. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J Exp Bot,* 59**,** 1515-24.

TANAKA, S., SAWAYA, M. R., KERFELD, C. A. & YEATES, T. O. 2007. Structure of the Rubisco chaperone RbcX from *Synechocystis* sp. PCC6803. *Acta Crystallogr D Biol Crystallogr,* 63**,** 1109-12.

TARNAWSKI, M., GUBERNATOR, B., KOLESINSKI, P. & SZCZEPANIAK, A. 2008. Heterologous expression and initial characterization of recombinant RbcX protein from

Thermosynechococcus elongatus BP-1 and the role of RbcX in Rubisco assembly. *Acta Biochim Pol,* 55, 777-85.

TAYLOR, T. C. & ANDERSSON, I. 1996. Structural transitions during activation and ligand binding in hexadecameric Rubisco inferred from the crystal structure of the activated unliganded spinach enzyme. *Nat Struct Biol,* 3, 95-101.

TAYLOR, T. C., BACKLUND, A., BJORHALL, K., SPREITZER, R. J. & ANDERSSON, I. 2001. First crystal structure of Rubisco from a green alga, *Chlamydomonas reinhardtii*. *Journal of Biological Chemistry,* 276, 48159-64.

TCHERKEZ, G. G., FARQUHAR, G. D. & ANDREWS, T. J. 2006. Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proc Natl Acad Sci U S A,* 103, 7246-51.

TERACHI, T., OGIHARA, Y. & TSUNEWAKI, K. 1987. The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. VI. Complete nucleotide sequences of the rbcL genes encoding H- and L-type Rubisco large subunits in common wheat and *Ae. crassa* 4x. *Japanese Journal of Genetics* 62, 375-387.

THANARAJ, T. A. & ARGOS, P. 1996. Ribosome-mediated translational pause and protein domain organization. *Protein Sci,* 5, 1594-612.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22, 4673-80.

THOMPSON, J. N. 1994. The Coevolutionary Process. *Chicago: University of Chicago Press.*

TSAI, C. J., SAUNA, Z. E., KIMCHI-SARFATY, C., AMBUDKAR, S. V., GOTTESMAN, M. M. & NUSSINOV, R. 2008. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *Journal of Molecular Biology,* 383, 281-91.

TUFF, P. & DARLU, P. 2000. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Molecular Biology and Evolution,* 17, 1753-9.

URRUTIA, A. O. & HURST, L. D. 2003. The signature of selection mediated by expression on human genes. *Genome Res,* 13, 2260-4.

VAN VALEN, L. 1977. The Red Queen. *The American Naturalist,* 11, 809-810.

VICATOS, S., REDDY, B. V. B. & KAZNESSIS, Y. 2005. Prediction of distant residue contacts with the use of evolutionary information. *Proteins-Structure Function and Bioinformatics,* 58, 935-949.

WADA, K., AOTA, S., TSUCHIYA, R., ISHIBASHI, F., GOJOBORI, T. & IKEMURA, T. 1990. Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Research,* 18 Suppl, 2367-411.

WALL, D. P. & HERBECK, J. T. 2003. Evolutionary patterns of codon usage in the chloroplast gene rbcL. *J Mol Evol,* 56, 673-88; discussion 689-90.

WANG, M., KAPRALOV, M. & ANISIMOVA, M. 2011. Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco. *BMC evolutionary biology,* 11, 266.

WANG, Z. O. & POLLOCK, D. D. 2007. Coevolutionary patterns in cytochrome c oxidase subunit I depend on structural and functional context. *J Mol Evol,* 65, 485-95.

WANG, Z. Y., SNYDER, G. W., ESAU, B. D., PORTIS, A. R. & OGREN, W. L. 1992. Species-dependent variation in the interaction of substrate-bound ribulose-1,5-

bisphosphate carboxylase/oxygenase (rubisco) and Rubisco activase. *Plant Physiol,* 100, 1858-62.

WATSON, G. M., YU, J. P. & TABITA, F. R. 1999. Unusual ribulose 1,5-bisphosphate carboxylase/oxygenase of anoxic Archaea. *J Bacteriol,* 181, 1569-75.

WHITNEY, S. M., BALDET, P., HUDSON, G. S. & ANDREWS, T. J. 2001. Form I Rubiscos from non-green algae are expressed abundantly but not assembled in tobacco chloroplasts. *Plant J,* 26, 535-47.

WHITNEY, S. M., HOUTZ, R. L. & ALONSO, H. 2011a. Advancing our understanding and capacity to engineer nature's CO_2-sequestering enzyme, Rubisco. *Plant Physiol,* 155, 27-35.

WHITNEY, S. M., SHARWOOD, R. E., ORR, D., WHITE, S. J., ALONSO, H. & GALMES, J. 2011b. Isoleucine 309 acts as a C_4 catalytic switch that increases ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in *Flaveria. Proc Natl Acad Sci U S A,* 108, 14688-93.

WHITNEY, S. M., VON CAEMMERER, S., HUDSON, G. S. & ANDREWS, T. J. 1999. Directed mutation of the Rubisco large subunit of tobacco influences photorespiration and growth. *Plant Physiol,* 121, 579-588.

WICKNER, S. & MAURIZI, M. R. 1999. Here's the hook: similar substrate binding sites in the chaperone domains of Clp and Lon. *Proc Natl Acad Sci U S A,* 96, 8318-20.

WOLFE, K. H. & SHARP, P. M. 1988. Identification of functional open reading frames in chloroplast genomes. *Gene,* 66, 215-22.

WRIGHT, S. I., YAU, C. B., LOOSELEY, M. & MEYERS, B. C. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol Biol Evol,* 21, 1719-26.

XIE, T. & DING, D. 1998. The relationship between synonymous codon usage and protein structure. *FEBS Lett,* 434, 93-6.

YEANG, C. H., DAROT, J. F. J., NOLLER, H. F. & HAUSSLER, D. 2007. Detecting the coevolution of biosequences - An example of RNA interaction prediction (vol 24, pg 1592, 2007). *Molecular Biology and Evolution,* 24, 2354-2354.

YEANG, C. H. & HAUSSLER, D. 2007. Detecting coevolution in and among protein domains. *Plos Computational Biology,* 3, 2122-2134.

YOSHIDA, S., ATOMI, H. & IMANAKA, T. 2007. Engineering of a type III rubisco from a hyperthermophilic archaeon in order to enhance catalytic performance in mesophilic host cells. *Appl Environ Microbiol,* 73, 6254-61.

ZAMA, M. 1995. Discontinuous translation and mRNA secondary structure. *Nucleic Acids Symp Ser*, 97-8.

ZHANG, G., HUBALEWSKA, M. & IGNATOVA, Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol,* 16, 274-80.

ZHOU, T., WEEMS, M. & WILKE, C. O. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol,* 26, 1571-80.

ZHU, G. & SPREITZER, R. J. 1996. Directed mutagenesis of chloroplast ribulose-1,5-bisphosphate carboxylase/oxygenase. Loop 6 substitutions complement for structural stability but decrease catalytic efficiency. *J Biol Chem,* 271, 18494-8.

ZHU, G. H., JENSEN, R. G., BOHNERT, H. J., WILDNER, G. F. & SCHLITTER, J. 1998. Dependence of catalysis and CO_2/O_2 specificity of Rubisco on the carboxy-terminus of the large subunit at different temperatures. *Photosynthesis Research,* 57**,** 71-79.

Appendix

2.1 Rubisco structures in Rubisco database

PDB ID	Resolution	*Source*	Reference
1AA1	2.20	*Spinacia oleracea*	Taylor and Andersson (1997a)
1AUS	2.20	*Spinacia oleracea*	Taylor and Andersson (1997a)
1BWV	2.40	*Galdieria partita*	Sugawara et al.(1999)
1BXN	2.70	*Cupriavidus necator*	Hansen et al.(1999)
1EJ7	2.45	*Nicotiana tabacum*	Duff et al.(2000)
1GEH	2.80	*Thermococcus kodakarensis*	Kitano et al.(Kitano et al., 2001)
1GK8	1.40	*Chlamydomonas reinhardtii*	Taylor et al.(Taylor et al., 2001)
1IR1	1.80	*Spinacia oleracea*	Mizohata et al.(2002)
1IR2	1.84	*Chlamydomonas reinhardtii*	Mizohata et al.(2002)
1IWA	2.60	*Galdieria partita*	Okano et al.(2002)
1RBA	2.60	*Rhodospirillum rubrum*	Soderlind et al.(1992)
1RBL	2.20	*Synechococcus elongatus*	Newman et al.(1993)
1RBO	2.30	*Spinacia oleracea*	Taylor et al.(1996)
1RCO	2.30	*Spinacia oleracea*	Taylor et al.(1996)
1RCX	2.40	*Spinacia oleracea*	Taylor and Andersson (1997b)
1RLC	2.70	*Nicotiana tabacum*	Zhang et al.(1994)
1RLD	2.50	*Nicotiana tabacum*	Zhang and Eisenberg (1994)
1RSC	2.30	*Synechococcus elongatus*	Newman and Gutteridge (1994)
1RUS	2.90	*Rhodospirillum rubrum*	Lundqvist and Schneider (1989)
1RXO	2.20	*Spinacia oleracea*	Taylor and Andersson (1997b)
1SVD	1.80	*Halothiobacillus neapolitanus*	Kerfeld et al.(2004)
1UPM	2.30	Spinacia oleracea	Karkehabadi et al.(2003)
1UPP	2.30	*Spinacia oleracea*	Karkehabadi et al.(2003)
1UW9	2.05	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2005b)
1UWA	2.30	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2005b)
1UZD	2.40	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2005a)
1UZH	2.20	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2005a)
1WDD	1.35	*Oryza sativa Japonica Group*	Matsumura et al.(2012)
2CWX	2.00	*Pyrococcus horikoshii*	Mizohata et al.(2005)
2CXE	3.00	*Pyrococcus horikoshii*	Mizohata et al.(2005)
2D69	1.90	*Pyrococcus horikoshii*	Mizohata et al.(2005)
2RUS	2.30	*Rhodospirillum rubrum*	Lundqvist and Schneider (1991b)
2V63	1.80	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2007)
2V67	2.00	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2007)
2V68	2.30	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2007)
2V69	2.80	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2007)
2V6A	1.50	*Chlamydomonas reinhardtii*	Karkehabadi et al.(2007)
2VDH	2.30	*Chlamydomonas reinhardtii*	Garcia-Murria et al.(2008)
2VDI	2.65	*Chlamydomonas reinhardtii*	Garcia-Murria et al.(2008)
2WVW	9.00	*Synechococcus elongatus*	Liu et al.(2010)
2YBV	2.30	*Thermosynechococcus elongatus*	Terlecka et al.(2011)
3A12	2.30	*Pyrococcus kodakaraensis*	Nishitani et al.(2010)
3A13	2.34	*Pyrococcus kodakaraensis*	Nishitani et al.(2009)
3AXK	1.90	*Oryza sativa Japonica Group*	Matsumura et al.(2012)

PDB ID	Resolution	*Source*	Reference
3AXM	1.65	*Oryza sativa Japonica Group*	Matsumura et al.(2012)
3KDN	2.09	*Thermococcus kodakaraensis*	Nishitani et al.(2010)
3KDO	2.36	*Thermococcus kodakaraensis*	Nishitani et al.(2010)
3QFW	1.79	*Rhodopseudomonas palustris*	Fedorov et al.(2011)
3RG6	3.20	*Synechococcus elongatus*	Bracher et al.(2011)
5RUB	1.70	*Rhodospirillum rubrum*	Schneider et al.(1990)
8RUC	1.60	*Spinacia oleracea*	Andersson (1996)
9RUB	2.60	*Rhodospirillum rubrum*	Lundqvist and Schneider (1991a)

2.2 Scripts and sequences

Scripts used in Chapter 2 and all the sequences (~11,400 *rbcL* and Rubisco-LSU sequences) in Rubisco database are in RUBISCO_DB directory of accompanying compact disk.

3.1 Sequence ids of sequences used in Rubisco-LSU-RA coevolution analysis

Sequence ids of Rubisco-LSU sequences

Accession No.	Name of species (Rubisco-LSU sequences) [a]
NP_051067.1	*Arabidopsis thaliana*
AAO38781.1	*Brassica rapa subsp. campestris*
AAA18385.1	*Capsicum baccatum*
YP_538747.1	*Glycine max*
YP_538943.1	*Gossypium hirsutum*
YP_874661.1	*Hordeum vulgare subsp. vulgare*
AAX38267.1	*Ipomoea batatas*
CAA75253.1	*Larrea tridentata*
NP_054507.1	*Nicotiana tabacum*
YP_654221.1	*Oryza sativa Indica Group*
NP_039391.1	*Oryza sativa Japonica Group*
YP_001122790.1	*Phaseolus vulgaris*
NP_904194.1	*Physcomitrella patens subsp. patens*
AAX59144.1	*Ricinus communis*
YP_003097495.1	*Selaginella moellendorffii*
YP_514860.1	*Solanum lycopersicum*
ACR19808.1	*Solanum pennellii*
YP_899415.1	*Sorghum bicolor*
NP_054944.1	*Spinacia oleracea*
NP_114267.1	*Triticum aestivum*
YP_567084.1	*Vitis vinifera*
AAC97876.1	*Zantedeschia aethiopica*
NP_043033.1	*Zea mays*

[a] Please note that *Capsicum baccatum* Rubisco-LSU sequence was utilized in this analysis due to non-availability of *Capsicum annum* Rubisco-LSU sequence.

Sequence ids of RA sequences

Accession No.	Name of species (RA sequences)
NP_565913.1	*Arabidopsis thaliana*
AC189306.2	*Brassica rapa(Annotated cDNA)*
ACB05667.1	*Capsicum Annum*
ADD60242.1	*Glycine max*
AAG61120.1	*Gossypium hirsutum*
Q40073.1	*Hordeum vulgare*
ABX84141.1	*Ipomoea batatas*
AAP83929.1	*Larrea tridentata*
Q40460.1	*Nicotiana tabacum*
CT830274.1	*Oryza sativa indica(Annotated cDNA)*
BAA97583.1	*Oryza sativa japonica*
AAC12868.1	*Phaseolus vulgaris*
XP_001776035.1	*Physcomitrella patens subsp. patens*
XP_002524206.1	*Ricinus communis*
XP_002982838.1	*Selaginella moellendorffii*
AK325923.1	*Solanum lycopersicum(Annotated cDNA)*
AAC15236.1	*Solanum pennellii*
XP_002451328.1	*Sorghum bicolor*
AAA34038.1	*Spinacia oleracea*
AK330616.1	*Triticum aestivum(Annotated cDNA)*
XP_002282979.1	*Vitis vinifera*
AAK25798.1	*Zantedeschia aethiopica*
NP_001104921.1	*Zea mays*

3.2 Sequence ids of sequences used in Rubisco-LSU-Rubisco-SSU coevolution analysis

Sequence ids of Rubisco-LSU sequences

Accession No.	Name of species (Rubisco-LSU sequences)
AAX44989.1	*Aegilops speltoides*
AAX44974.1	*Aegilops tauschii*
P16306.1	*Amaranthus hypochondriacus*
NP_051067.1	*Arabidopsis thaliana*
AAB67895.1	*Arachis hypogaea*
AAA84028.1	*Avena sativa*
AAO38782.1	*Brassica juncea*
AAF78948.1	*Brassica napus*
YP_002149717.1	*Cicer arietinum*
YP_817490.1	*Coffea arabica*
BAB70581.1	*Fagus crenata*
CAA39356.1	*Flaveria bidentis*
CAA39355.1	*Flaveria pringlei*
YP_538747.1	*Glycine max*
YP_538943.1	*Gossypium hirsutum*
YP_588125.1	*Helianthus annuus*
YP_004327670.1	*Hevea brasiliensis*
P05698.2	*Hordeum vulgare*
YP_002720120.1	*Jatropha curcas*
YP_398337.1	*Lactuca sativa*
YP_001718445.1	*Manihot esculenta*
CAA28648.1	*Medicago sativa*
YP_001381744.1	*Medicago truncatula*
ABU85466.1	*Musa acuminata*
YP_358684.1	*Nicotiana sylvestris*
NP_054507.1	*Nicotiana tabacum*
YP_086974.1	*Panax ginseng*
CAA28649.1	*Petunia x hybrida*
YP_001122790.1	*Phaseolus vulgaris*
YP_003587524.1	*Pisum sativum*
YP_001109509.1	*Populus trichocarpa*
AEJ82563.1	*Ricinus communis*
AAN71851.1	*Rumex obtusifolius*
YP_054639.1	*Saccharum officinarum*
YP_514860.1	*Solanum lycopersicum*
YP_635647.1	*Solanum tuberosum*
YP_899415.1	*Sorghum bicolor*
NP_054944.1	*Spinacia oleracea*

Accession No.	Name of species (Rubisco-LSU sequences)
NP_114267.1	*Triticum aestivum*
YP_003434328.1	*Vigna radiata*
YP_002608342.1	*Vitis vinifera*
YP_004769955.1	*Wolffia australiana*
AAC97876.1	*Zantedeschia aethiopica*
NP_043033.1	*Zea mays*

Sequence ids of Rubisco-SSU sequences

Accession No.	Name of species (Rubisco-SSU sequences)
BAA35167.1	*Aegilops speltoides*
Q38793.1	*Aegilops tauschii*
Q9XGX5.1	*Amaranthus hypochondriacus*
AED94313.1	*Arabidopsis thaliana*
1211236B	*Arachis hypogaea*
BAA35164.1	*Avena sativa*
AEB00556.1	*Brassica juncea*
P05346.2	*Brassica napus*
CAA10290.1	*Cicer arietinum*
CAD11991.1	*Coffea arabica*
O22077.1	*Fagus crenata*
AAP31054.1	*Flaveria bidentis*
Q39746.1	*Flaveria pringlei*
P12468.1	*Glycine max*
CAA38026.1	*Gossypium hirsutum*
P08705.1	*Helianthus annuus*
ACA42439.1	*Hevea brasiliensis*
Q40004.1	*Hordeum vulgare*
ADB85091.1	*Jatropha curcas*
AAF19793.1	*Lactuca sativa*
AAF06101.1	*Manihot esculenta*
O65194.1	*Medicago sativa*
ACJ85905.1	*Medicago truncatula*
O24045.1	*Musa acuminata*
P22433.1	*Nicotiana sylvestris*
P69249.1	*Nicotiana tabacum*
BAE46384.1	*Panax ginseng*
CAA27445.1	*Petunia x hybrida*
CAA40339.1	*Phaseolus vulgaris*
CAA25390.1	*Pisum sativum*
XP_002305162.1	*Populus trichocarpa*
XP_002532149.1	*Ricinus communis*

Accession No.	Name of species (Rubisco-SSU sequences)
CAD21856.1	*Rumex obtusifolius*
S33613	*Saccharum officinarum*
P08706.2	*Solanum lycopersicum*
ABY21255.1	*Solanum tuberosum*
BAJ40065.1	*Sorghum bicolor*
AAB81105.1	*Spinacia oleracea*
BAB19814.1	*Triticum aestivum*
AAD27881.1	*Vigna radiata*
XP_002276991.1	*Vitis vinifera*
AEJ33935.1	*Wolffia australiana*
AAC18406.1	*Zantedeschia aethiopica*
NP_001105294.1	*Zea mays*

3.3 Sequence ids of sequences used in Rubisco-LSU-RbcX coevolution analysis

Sequence ids of Rubisco-LSU sequences

Accession No.	Name of species (Rubisco-LSU sequences)
AAO19427.1	*Arabidopsis lyrata subsp. lyrata*
NP_051067.1	*Arabidopsis thaliana*
YP_538747.1	*Glycine max*
YP_874661.1	*Hordeum vulgare subsp. vulgare*
CAG34174.1	*Oryza sativa*
YP_654221.1	*Oryza sativa Indica Group*
NP_039391.1	*Oryza sativa Japonica Group*
NP_904194.1	*Physcomitrella patens subsp. patens*
YP_002905095.1	*Picea sitchensis*
YP_001109509.1	*Populus trichocarpa*
AAX59144.1	*Ricinus communis*
YP_899415.1	*Sorghum bicolor*
YP_567084.1	*Vitis vinifera*
NP_043033.1	*Zea mays*

Sequence ids of RbcX sequences

Accession No.	Name of species (RbcX sequences)
XP_002873967.1	*Arabidopsis lyrata subsp. lyrata*
NP_568382.1	*Arabidopsis thaliana*

Accession No.	Name of species (RbcX sequences)
ACU20354.1	*Glycine max*
BAJ99949.1	*Hordeum vulgare subsp. vulgare*
CAV28344.1	*Oryza sativa*
EAY92274.1	*Oryza sativa Indica Group*
XP_001770683.1	*Physcomitrella patens subsp. patens*
NP_001060039.1	*Oryza sativa Japonica Group*
ABK23924.1	*Picea sitchensis*
XP_002314074.1	*Populus trichocarpa*
XP_002513502.1	*Ricinus communis*
XP_002466243.1	*Sorghum bicolor*
XP_002285429.1	*Vitis vinifera*
NP_001144731.1	*Zea mays*

3.4 Sequence ids for Solanales, Caryophyllales, Poales, Gentinales and Angiosperm dataset

Sequence ids for Solanales (141 sequences), Caryophyllales (207 sequences), Poales (213 sequences), Gentinales (440 sequences) and Angiosperm dataset (50552 sequences) are in Coevolution_Appendix.xlsx in accompanying compact disk.

4.1 Dataset for comparison of *rbcL* and whole chloroplast genome codon usage

	Accession No.	Name of species
1	NC_015820.1	*Acidosasa purpurea*
2	NC_010093.1	*Acorus americanus*
3	NC_007407.1	*Acorus calamus*
4	NC_009265.1	*Aethionema cordifolium*
5	NC_009266.1	*Aethionema grandiflorum*
6	NC_015621.1	*Ageratina adenophora*
7	NC_008591.1	*Agrostis stolonifera*
8	NC_014062.1	*Anomochloa marantoidea*
9	NC_015113.1	*Anthriscus cerefolium*
10	NC_000932.1	*Arabidopsis thaliana*
11	NC_009268.1	*Arabis hirsuta*
12	NC_004561.1	*Atropa belladonna*
13	NC_015830.1	*Bambusa emeiensis*
14	NC_012927.1	*Bambusa oldhamii*
15	NC_009269.1	*Barbarea verna*
16	NC_011032.1	*Brachypodium distachyon*
17	NC_009599.1	*Buxus microphylla*

	Accession No.	Name of species
18	NC_009270.1	*Capsella bursa-pastoris*
19	NC_010323.1	*Carica papaya*
20	NC_014674.1	*Castanea mollissima*
21	NC_011163.1	*Cicer arietinum*
22	NC_008334.1	*Citrus sinensis*
23	NC_008535.1	*Coffea arabica*
24	NC_013273.1	*Coix lacryma-jobi*
25	NC_014807.1	*Corynocarpus laevigata*
26	NC_015804.1	*Crithmum maritimum*
27	NC_009271.1	*Crucihimalaya wallichii*
28	NC_015983.1	*Cucumis melo subsp melo*
29	NC_007144.1	*Cucumis sativus*
30	NC_009963.1	*Cuscuta exaltata*
31	NC_009765.1	*Cuscuta gronovii*
32	NC_009949.1	*Cuscuta obtusiflora*
33	NC_009766.1	*Cuscuta reflexa*
34	NC_008325.1	*Daucus carota*
35	NC_013088.1	*Dendrocalamus latiflorus*
36	NC_009601.1	*Dioscorea elephantipes*
37	NC_009272.1	*Draba nemorosa*
38	NC_016430.1	*Eleutherococcus senticosus*
39	NC_015083.1	*Erodium carvifolium*
40	NC_014569.1	*Erodium texanum*
41	NC_008115.1	*Eucalyptus globulus subsp globulus*
42	NC_014570.1	*Eucalyptus grandis*
43	NC_010776.1	*Fagopyrum esculentum subsp ancestrale*
44	NC_015831.1	*Ferrocalamus rimosivaginus*
45	NC_011713.2	*Festuca arundinacea*
46	NC_015206.1	*Fragaria vesca subsp vesca*
47	NC_014573.1	*Geranium palmatum*
48	NC_007942.1	*Glycine max*
49	NC_008641.1	*Gossypium barbadense*
50	NC_007944.1	*Gossypium hirsutum*
51	NC_015204.1	*Gossypium thurberi*
52	NC_010601.1	*Guizotia abyssinica*
53	NC_007977.1	*Helianthus annuus*
54	NC_015308.1	*Hevea brasiliensis*
55	NC_008590.1	*Hordeum vulgare subsp vulgare*
56	NC_015818.1	*Hydrocotyle sp SRD-2010*
57	NC_015803.1	*Indocalamus longiauritus*
58	NC_009808.1	*Ipomoea purpurea*
59	NC_015543.1	*Jacobaea vulgaris*

	Accession No.	Name of species
60	NC_008407.1	*Jasminum nudiflorum*
61	NC_012224.1	*Jatropha curcas*
62	NC_007578.1	*Lactuca sativa*
63	NC_014063.1	*Lathyrus sativus*
64	NC_010109.1	*Lemna minor*
65	NC_009273.1	*Lepidium virginicum*
66	NC_009274.1	*Lobularia maritima*
67	NC_009950.1	*Lolium perenne*
68	NC_002694.1	*Lotus japonicus*
69	NC_010433.1	*Manihot esculenta*
70	NC_003119.6	*Medicago truncatula*
71	NC_012615.1	*Megaleranthis saniculifolia*
72	NC_014582.1	*Monsonia speciosa*
73	NC_008359.1	*Morus indica*
74	NC_008336.1	*Nandina domestica*
75	NC_009275.1	*Nasturtium officinale*
76	NC_015605.1	*Nelumbo lutea*
77	NC_015610.1	*Nelumbo nucifera*
78	NC_007500.1	*Nicotiana sylvestris*
79	NC_001879.2	*Nicotiana tabacum*
80	NC_007602.1	*Nicotiana tomentosiformis*
81	NC_016068.1	*Nicotiana undulata*
82	NC_010358.1	*Oenothera argillicola*
83	NC_010361.1	*Oenothera biennis*
84	NC_002693.2	*Oenothera elata subsp hookeri*
85	NC_010360.1	*Oenothera glazioviana*
86	NC_010362.1	*Oenothera parviflora*
87	NC_013707.2	*Olea europaea*
88	NC_015604.1	*Olea europaea subsp cuspidata*
89	NC_015401.1	*Olea europaea subsp europaea*
90	NC_015623.1	*Olea europaea subsp maroccana*
91	NC_015608.1	*Olea woodiana subsp woodiana*
92	NC_009267.1	*Olimarabidopsis pumila*
93	NC_014056.1	*Oncidium Gower Ramsey*
94	NC_005973.1	*Oryza nivara*
95	NC_008155.1	*Oryza sativa Indica Group*
96	NC_001320.1	*Oryza sativa Japonica Group*
97	NC_015832.1	*Oxypolis greenmanii*
98	NC_006290.1	*Panax ginseng*
99	NC_015990.1	*Panicum virgatum*
100	NC_013553.1	*Parthenium argentatum*
101	NC_015821.1	*Petroselinum crispum*

	Accession No.	Name of species
102	NC_007499.1	*Phalaenopsis aphrodite subsp formosana*
103	NC_009259.1	*Phaseolus vulgaris*
104	NC_013991.2	*Phoenix dactylifera*
105	NC_015817.1	*Phyllostachys edulis*
106	NC_015826.1	*Phyllostachys nigra var henonis*
107	NC_014057.1	*Pisum sativum*
108	NC_008335.1	*Platanus occidentalis*
109	NC_008235.1	*Populus alba*
110	NC_009143.1	*Populus trichocarpa*
111	NC_014697.1	*Prunus persica*
112	NC_015996.1	*Pyrus pyrifolia*
113	NC_008796.1	*Ranunculus macranthus*
114	NC_006084.1	*Saccharum hybrid cultivar NCo 310*
115	NC_005878.2	*Saccharum hybrid cultivar SP-80-3280*
116	NC_016433.2	*Sesamum indicum*
117	NC_007943.1	*Solanum bulbocastanum*
118	NC_007898.2	*Solanum lycopersicum*
119	NC_008096.2	*Solanum tuberosum*
120	NC_008602.1	*Sorghum bicolor*
121	NC_002202.1	*Spinacia oleracea*
122	NC_015891.1	*Spirodela polyrhiza*
123	NC_014676.2	*Theobroma cacao*
124	NC_010442.1	*Trachelium caeruleum*
125	NC_011828.1	*Trifolium subterraneum*
126	NC_002762.1	*Triticum aestivum*
127	NC_013823.1	*Typha latifolia*
128	NC_013843.1	*Vigna radiata*
129	NC_007957.1	*Vitis vinifera*
130	NC_015899.1	*Wolffia australiana*
131	NC_015894.1	*Wolffiella lingulata*
132	NC_001666.2	*Zea mays*

4.2 Dataset for compilation of tRNA genes

	Accession No.	Name of species
1	NC_015820.1	*Acidosasa purpurea*
2	NC_010093.1	*Acorus americanus*
3	NC_009265.1	*Aethionema cordifolium*
4	NC_009266.1	*Aethionema grandiflorum*
5	NC_015621.1	*Ageratina adenophora*
6	NC_008591.1	*Agrostis stolonifera*
7	NC_014062.1	*Anomochloa marantoidea*
8	NC_015113.1	*Anthriscus cerefolium*
9	NC_000932.1	*Arabidopsis thaliana*
10	NC_009268.1	*Arabis hirsuta*
11	NC_004561.1	*Atropa belladonna*
12	NC_015830.1	*Bambusa emeiensis*
13	NC_009269.1	*Barbarea verna*
14	NC_011032.1	*Brachypodium distachyon*
15	NC_009599.1	*Buxus microphylla*
16	NC_009270.1	*Capsella bursa-pastoris*
17	NC_010323.1	*Carica papaya*
18	NC_014674.1	*Castanea mollissima*
19	NC_011163.1	*Cicer arietinum*
20	NC_008334.1	*Citrus sinensis*
21	NC_008535.1	*Coffea arabica*
22	NC_013273.1	*Coix lacryma-jobi*
23	NC_015804.1	*Crithmum maritimum*
24	NC_009271.1	*Crucihimalaya wallichii*
25	NC_015983.1	*Cucumis melo subsp melo*
26	NC_007144.1	*Cucumis sativus*
27	NC_009963.1	*Cuscuta exaltata*
28	NC_009765.1	*Cuscuta gronovii*
29	NC_009949.1	*Cuscuta obtusiflora*
30	NC_009766.1	*Cuscuta reflexa*
31	NC_008325.1	*Daucus carota*
32	NC_013088.1	*Dendrocalamus latiflorus*
33	NC_009601.1	*Dioscorea elephantipes*
34	NC_009272.1	*Draba nemorosa*
35	NC_016430.1	*Eleutherococcus senticosus*
36	NC_015083.1	*Erodium carvifolium*
37	NC_014569.1	*Erodium texanum*
38	NC_008115.1	*Eucalyptus globulus subsp globulus*
39	NC_014570.1	*Eucalyptus grandis*
40	NC_010776.1	*Fagopyrum esculentum subsp ancestrale*

	Accession No.	Name of species
41	NC_015831.1	*Ferrocalamus rimosivaginus*
42	NC_011713.2	*Festuca arundinacea*
43	NC_015206.1	*Fragaria vesca subsp vesca*
44	NC_014573.1	*Geranium palmatum*
45	NC_007942.1	*Glycine max*
46	NC_008641.1	*Gossypium barbadense*
47	NC_007944.1	*Gossypium hirsutum*
48	NC_015204.1	*Gossypium thurberi*
49	NC_010601.1	*Guizotia abyssinica*
50	NC_007977.1	*Helianthus annuus*
51	NC_015308.1	*Hevea brasiliensis*
52	NC_008590.1	*Hordeum vulgare subsp vulgare*
53	NC_015818.1	*Hydrocotyle sp SRD-2010*
54	NC_015803.1	*Indocalamus longiauritus*
55	NC_009808.1	*Ipomoea purpurea*
56	NC_015543.1	*Jacobaea vulgaris*
57	NC_008407.1	*Jasminum nudiflorum*
58	NC_012224.1	*Jatropha curcas*
59	NC_007578.1	*Lactuca sativa*
60	NC_014063.1	*Lathyrus sativus*
61	NC_010109.1	*Lemna minor*
62	NC_009273.1	*Lepidium virginicum*
63	NC_009274.1	*Lobularia maritima*
64	NC_009950.1	*Lolium perenne*
65	NC_002694.1	*Lotus japonicus*
66	NC_010433.1	*Manihot esculenta*
67	NC_003119.6	*Medicago truncatula*
68	NC_012615.1	*Megaleranthis saniculifolia*
69	NC_014582.1	*Monsonia speciosa*
70	NC_008359.1	*Morus indica*
71	NC_008336.1	*Nandina domestica*
72	NC_009275.1	*Nasturtium officinale*
73	NC_015605.1	*Nelumbo lutea*
74	NC_015610.1	*Nelumbo nucifera*
75	NC_007500.1	*Nicotiana sylvestris*
76	NC_001879.2	*Nicotiana tabacum*
77	NC_007602.1	*Nicotiana tomentosiformis*
78	NC_016068.1	*Nicotiana undulata*
79	NC_010358.1	*Oenothera argillicola*
80	NC_010361.1	*Oenothera biennis*
81	NC_002693.2	*Oenothera elata subsp hookeri*
82	NC_010360.1	*Oenothera glazioviana*

	Accession No.	Name of species
83	NC_010362.1	*Oenothera parviflora*
84	NC_009267.1	*Olimarabidopsis pumila*
85	NC_014056.1	*Oncidium Gower Ramsey*
86	NC_005973.1	*Oryza nivara*
87	NC_008155.1	*Oryza sativa Indica Group*
88	NC_001320.1	*Oryza sativa Japonica Group*
89	NC_015832.1	*Oxypolis greenmanii*
90	NC_006290.1	*Panax ginseng*
91	NC_015990.1	*Panicum virgatum*
92	NC_015821.1	*Petroselinum crispum*
93	NC_007499.1	*Phalaenopsis aphrodite subsp formosana*
94	NC_009259.1	*Phaseolus vulgaris*
95	NC_013991.2	*Phoenix dactylifera*
96	NC_015817.1	*Phyllostachys edulis*
97	NC_015826.1	*Phyllostachys nigra var henonis*
98	NC_014057.1	*Pisum sativum*
99	NC_008335.1	*Platanus occidentalis*
100	NC_008235.1	*Populus alba*
101	NC_009143.1	*Populus trichocarpa*
102	NC_014697.1	*Prunus persica*
103	NC_015996.1	*Pyrus pyrifolia*
104	NC_008796.1	*Ranunculus macranthus*
105	NC_006084.1	*Saccharum hybrid cultivar NCo 310*
106	NC_005878.2	*Saccharum hybrid cultivar SP-80-3280*
107	NC_016433.2	*Sesamum indicum*
108	NC_007943.1	*Solanum bulbocastanum*
109	NC_007898.2	*Solanum lycopersicum*
110	NC_008096.2	*Solanum tuberosum*
111	NC_008602.1	*Sorghum bicolor*
112	NC_002202.1	*Spinacia oleracea*
113	NC_015891.1	*Spirodela polyrhiza*
114	NC_014676.2	*Theobroma cacao*
115	NC_010442.1	*Trachelium caeruleum*
116	NC_011828.1	*Trifolium subterraneum*
117	NC_002762.1	*Triticum aestivum*
118	NC_013823.1	*Typha latifolia*
119	NC_013843.1	*Vigna radiata*
120	NC_007957.1	*Vitis vinifera*
121	NC_015899.1	*Wolffia australiana*
122	NC_015894.1	*Wolffiella lingulata*
123	NC_001666.2	*Zea mays*

4.3 Sequences for *rbcL* Dataset

Sequence ids for *rbcL* dataset (4944 sequences) are in rbcL_dataset.xlsx in accompanying compact disk.

List of references for Appendix

ANDERSSON, I. 1996. Large structures at high resolution: the 1.6 A crystal structure of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase complexed with 2-carboxyarabinitol bisphosphate. *J Mol Biol,* 259, 160-74.

BRACHER, A., STARLING-WINDHOF, A., HARTL, F. U. & HAYER-HARTL, M. 2011. Crystal structure of a chaperone-bound assembly intermediate of form I Rubisco. *Nat Struct Mol Biol,* 18, 875-80.

DUFF, A. P., ANDREWS, T. J. & CURMI, P. M. 2000. The transition between the open and closed states of rubisco is triggered by the inter-phosphate distance of the bound bisphosphate. *J Mol Biol,* 298, 903-16.

FEDOROV, A. A., FEDOROV, E. V., GERLT, J. A., BURLEY, S. K. & AL, S. C. 2011. Crystal Structure Of Rubisco-Like Protein From Rhodopseudomonas
Palustris. *unpublished*.

GARCIA-MURRIA, M. J., KARKEHABADI, S., MARIN-NAVARRO, J., SATAGOPAN, S., ANDERSSON, I., SPREITZER, R. J. & MORENO, J. 2008. Structural and functional consequences of the replacement of proximal residues Cys(172) and Cys(192) in the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from Chlamydomonas reinhardtii. *Biochem J,* 411, 241-7.

HANSEN, S., VOLLAN, V. B., HOUGH, E. & ANDERSEN, K. 1999. The crystal structure of rubisco from Alcaligenes eutrophus reveals a novel central eight-stranded beta-barrel formed by beta-strands from four subunits. *J Mol Biol,* 288, 609-21.

KARKEHABADI, S., PEDDI, S. R., ANWARUZZAMAN, M., TAYLOR, T. C., CEDERLUND, A., GENKOV, T., ANDERSSON, I. & SPREITZER, R. J. 2005a. Chimeric small subunits influence catalysis without causing global conformational changes in the crystal structure of ribulose-1,5-bisphosphate carboxylase/oxygenase. *Biochemistry,* 44, 9851-61.

KARKEHABADI, S., SATAGOPAN, S., TAYLOR, T. C., SPREITZER, R. J. & ANDERSSON, I. 2007. Structural analysis of altered large-subunit loop-6/carboxy-terminus interactions that influence catalytic efficiency and CO2/O2 specificity of ribulose-1,5-bisphosphate carboxylase/oxygenase. *Biochemistry,* 46, 11080-9.

KARKEHABADI, S., TAYLOR, T. C. & ANDERSSON, I. 2003. Calcium supports loop closure but not catalysis in Rubisco. *J Mol Biol,* 334, 65-73.

KARKEHABADI, S., TAYLOR, T. C., SPREITZER, R. J. & ANDERSSON, I. 2005b. Altered intersubunit interactions in crystal structures of catalytically compromised ribulose-1,5-bisphosphate carboxylase/oxygenase. *Biochemistry,* 44, 113-20.

KERFELD, C. A., SAWAYA, M. R., PASHKOV, I., CANNON, G., WILLIAMS, E., TRAN, K. & YEATES, T. O. 2004. The Structure Of Halothiobacillus Neapolitanus Rubisco. *unpublished.*

KITANO, K., MAEDA, N., FUKUI, T., ATOMI, H., IMANAKA, T. & MIKI, K. 2001. Crystal structure of a novel-type archaeal rubisco with pentagonal symmetry. *Structure,* 9, 473-81.

LIU, C., YOUNG, A. L., STARLING-WINDHOF, A., BRACHER, A., SASCHENBRECKER, S., RAO, B. V., RAO, K. V., BERNINGHAUSEN, O., MIELKE, T., HARTL, F. U., BECKMANN, R. & HAYER-HARTL, M. 2010. Coupled chaperone action in folding and assembly of hexadecameric Rubisco. *Nature,* 463, 197-202.

LUNDQVIST, T. & SCHNEIDER, G. 1989. Crystal structure of the binary complex of ribulose-1,5-bisphosphate carboxylase and its product, 3-phospho-D-glycerate. *J Biol Chem,* 264, 3643-6.

LUNDQVIST, T. & SCHNEIDER, G. 1991a. Crystal structure of activated ribulose-1,5-bisphosphate carboxylase complexed with its substrate, ribulose-1,5-bisphosphate. *J Biol Chem,* 266, 12604-11.

LUNDQVIST, T. & SCHNEIDER, G. 1991b. Crystal structure of the ternary complex of ribulose-1,5-bisphosphate carboxylase, Mg(II), and activator CO2 at 2.3-A resolution. *Biochemistry,* 30, 904-8.

MATSUMURA, H., MIZOHATA, E., ISHIDA, H., KOGAMI, A., UENO, T., MAKINO, A., INOUE, T., YOKOTA, A., MAE, T. & KAI, Y. 2012. Crystal Structure of Rice Rubisco and Implications for Activation Induced by Positive Effectors NADPH and 6-Phosphogluconate. *J Mol Biol.*

MIZOHATA, E., MATSUMURA, H., OKANO, Y., KUMEI, M., TAKUMA, H., ONODERA, J., KATO, K., SHIBATA, N., INOUE, T., YOKOTA, A. & KAI, Y. 2002. Crystal structure of activated ribulose-1,5-bisphosphate carboxylase/oxygenase from green alga Chlamydomonas reinhardtii complexed with 2-carboxyarabinitol-1,5-bisphosphate. *J Mol Biol,* 316, 679-91.

MIZOHATA, E., MISHIMA, C., AKASAKA, R., UDA, H., TERADA, T., SHIROUZU, M. & YOKOYAMA, S. 2005. Crystal Structure Of Octameric Ribulose-1,5-Bisphosphate CarboxylaseOXYGENASE (Rubisco) From Pyrococcus Horikoshii Ot3 (Form-1 Crystal). *unpublished*.

NEWMAN, J., BRANDEN, C. I. & JONES, T. A. 1993. Structure determination and refinement of ribulose 1,5-bisphosphate carboxylase/oxygenase from Synechococcus PCC6301. *Acta Crystallogr D Biol Crystallogr,* 49, 548-60.

NEWMAN, J. & GUTTERIDGE, S. 1994. Structure of an effector-induced inactivated state of ribulose 1,5-bisphosphate carboxylase/oxygenase: the binary complex between enzyme and xylulose 1,5-bisphosphate. *Structure,* 2, 495-502.

NISHITANI, Y., FUJIHASHI, M., DOI, T., YOSHIDA, S., ATOMI, H., IMANAKA, T. & MIKI, K. 2009. Sturcture-Based Optimization Of A Type Iii Rubisco From A Hyperthermophile. *unpublished*.

NISHITANI, Y., YOSHIDA, S., FUJIHASHI, M., KITAGAWA, K., DOI, T., ATOMI, H., IMANAKA, T. & MIKI, K. 2010. Structure-based catalytic optimization of a type III Rubisco from a hyperthermophile. *J Biol Chem,* 285, 39339-47.

OKANO, Y., MIZOHATA, E., XIE, Y., MATSUMURA, H., SUGAWARA, H., INOUE, T., YOKOTA, A. & KAI, Y. 2002. X-ray structure of Galdieria Rubisco complexed with one sulfate ion per active site. *FEBS Lett,* 527, 33-6.

SCHNEIDER, G., LINDQVIST, Y. & LUNDQVIST, T. 1990. Crystallographic refinement and structure of ribulose-1,5-bisphosphate carboxylase from Rhodospirillum rubrum at 1.7 A resolution. *J Mol Biol,* 211, 989-1008.

SODERLIND, E., SCHNEIDER, G. & GUTTERIDGE, S. 1992. Substitution of ASP193 to ASN at the active site of ribulose-1,5-bisphosphate carboxylase results in conformational changes. *Eur J Biochem,* 206, 729-35.

SUGAWARA, H., YAMAMOTO, H., SHIBATA, N., INOUE, T., OKADA, S., MIYAKE, C., YOKOTA, A. & KAI, Y. 1999. Crystal structure of carboxylase reaction-oriented ribulose 1, 5-bisphosphate carboxylase/oxygenase from a thermophilic red alga, Galdieria partita. *J Biol Chem,* 274, 15655-61.

TAYLOR, T. C. & ANDERSSON, I. 1997a. Structure of a product complex of spinach ribulose-1,5-bisphosphate carboxylase/oxygenase. *Biochemistry,* 36, 4041-6.

TAYLOR, T. C. & ANDERSSON, I. 1997b. The structure of the complex between rubisco and its natural substrate ribulose 1,5-bisphosphate. *J Mol Biol,* 265, 432-44.

TAYLOR, T. C., BACKLUND, A., BJORHALL, K., SPREITZER, R. J. & ANDERSSON, I. 2001. First crystal structure of Rubisco from a green alga, Chlamydomonas reinhardtii. *J Biol Chem,* 276, 48159-64.

TAYLOR, T. C., FOTHERGILL, M. D. & ANDERSSON, I. 1996. A common structural basis for the inhibition of ribulose 1,5-bisphosphate carboxylase by 4-carboxyarabinitol 1,5-bisphosphate and xylulose 1,5-bisphosphate. *J Biol Chem,* 271, 32894-9.

TERLECKA, B., WILHELMI, V., BIALEK, W., GUBERNATOR, B., SZCZEPANIAK, A. & E, E. 2011. Structure Of Ribulose-1,5-Bisphosphate Carboxylase Oxygenase From Thermosynechococcus Elongatus. *unpublished*.

ZHANG, K. Y., CASCIO, D. & EISENBERG, D. 1994. Crystal structure of the unactivated ribulose 1,5-bisphosphate carboxylase/oxygenase complexed with a transition state analog, 2-carboxy-D-arabinitol 1,5-bisphosphate. *Protein Sci,* 3, 64-9.

ZHANG, K. Y. & EISENBERG, D. 1994. Solid-state phase transition in the crystal structure of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Acta Crystallogr D Biol Crystallogr,* 50, 258-62.