The evolution, structure, and function of RubisCO and its homolog the RubisCO-like protein

F. Robert Tabita*[†‡¶], Thomas E. Hanson[ʃ], Sriram Satagopan[†], Brian H. Witte[†], and Nathan E. Kreel[¶]

Department of Microbiology[†], the Plant Molecular Biology Biology/Biotechnology Program[‡], and the OSU Biochemistry Program[¶]
The Ohio State University
484 West 12th Avenue
Columbus, Ohio 43210-1292;
Graduate College of Marine and Earth Studies[ʃ], Delaware Biotechnology Institute, University of Delaware, 127 DBI, 15 Innovation Way, Newark, DE 19711

***To whom correspondence should be addressed at:**
Department of Microbiology
The Ohio State University
484 West 12th Avenue
Columbus, Ohio 43210-1292
Telephone: 614-292-4297
Fax: 614-292-6337
E-mail: Tabita.1@osu.edu

Running Title: RubisCO and RLP

Key words: RubisCO, structure/function, evolution, different forms, CBB cycle

Abstract

Ribulose 1,5-bisphosphate (RuBP) carboxylase/oxygenase (RubisCO) catalyzes the key reaction by which inorganic carbon may be assimilated into organic carbon. Phylogenetic analyses indicate that there are three classes of bona fide RubisCO proteins, forms I, II, and III, that all catalyze the same reactions. In addition, there exists another form of RubisCO, form IV, which does not catalyze typical RubisCO reactions. Form IV is actually a homolog of RubisCO and is called the RubisCO-like protein (RLP). Both RubisCO and RLP appear to have evolved from a methanogenic archaeal ancestor protein and comprehensive analyses indicate that the various forms (I, II, III, and IV) contain various subgroups, with individual sequences derived from representatives of all three kingdoms of life. The diversity of RubisCO molecules, many of which function in distinct milieus, have provided convenient model systems to study the ways in which the active site of this protein has evolved to accommodate necessary molecular adaptations. Such studies have proven useful to help provide a framework for understanding the molecular basis for many important aspects of RubisCO catalysis, including the elucidation of factors or functional groups that impinge on RubisCO carbon dioxide/oxygen substrate discrimination.

Introduction

Ribulose 1,5-bisphosphate (RuBP) carboxylase/oxygenase (RubisCO) is a protein that arguably catalyzes the most important biochemical reaction in biology, as it is responsible for the vast majority of all the organic carbon found in the biosphere. The enzyme, found in phototrophic and chemoautotrophic organisms, basically functions to catalyze the removal and sequestration of carbon dioxide from the environment by reducing this oxidized gas to the level of organic carbon, in the process providing the organic building blocks needed to sustain life. However, RubisCO has an innate problem in that the enediolate intermediate of RuBP that acts as CO_2 acceptor may also be attacked by other ligands, including molecular oxygen, such that CO_2 and O_2 compete for the enediolate intermediate at the same active site. The promiscuity of the enediol intermediate thus enables RubisCO to function as an internal monooxygenase, as well as a carboxylase, with the unique product, 2-phosphoglycolate (2-PG), formed as a result of O_2 fixation (Fig. 1). Efficient RubisCO catalysis is thus dependent on the inherent ability of the enzyme to discriminate between CO_2 and O_2 (the substrate specificity or Ω or τ value) at the relative concentration of CO_2 and O_2 employed in a particular reaction. The rates of the two reactions (v_c and v_o) may be defined by $v_c/v_o = \Omega\ [CO_2]/[O_2]$. Thus, $\Omega = v_c\ [O_2]/v_o\ [CO_2]$ and $\Omega = V_cK_o/V_oK_c$ with V_c and V_o representing maximum velocities for carboxylation and oxygenation, respectively, and K_c and K_o the relative Michaelis constants for CO_2 and O_2, respectively.

Because of the great disparity in ambient O_2 concentrations relative to CO_2 levels, aerobic organisms produce considerable amounts of 2-PG. The resultant oxidative metabolism of 2-PG is inimical to maximal CO_2 fixation (and biomass productivity) because up to 50 % of the CO_2 that is fixed via RubisCO may be released as a result of 2-PG metabolism (Fig. 1). Clearly, a major key to solving the CO_2 sequestration issue and perhaps improving photosynthetic crop yields is to somehow improve RubisCO's ability to discriminate between CO_2 and O_2 (Long et al. 2006a,b). In this article we review likely scenarios as to how RubisCO

may have evolved from primordial ancestors, how the active site from different forms of RubisCO became adapted to diverse intracellular milieus, and how we might take advantage of these findings.

Different molecular forms of RubisCO for the same and different functions

Sixty years after the discovery of Fraction One Protein (i.e., RubisCO) from plant leaf extracts, studies with prokaryotic CO_2 assimilatory organisms have made it abundantly clear that nature has evolved different structural forms of this enzyme (Tabita 1999; 2007). In most instances RubisCO, as expected, catalyzes the typical carboxylation reaction so necessary for CO_2 reduction via the Calvin-Benson-Bassham (CBB) reductive pentose phosphate pathway (Fig. 1). However, as the different molecular forms were discovered and analyzed, it also soon became apparent that RubisCO may be used for purposes other than for primary carbon assimilation. Indeed, in certain archaea, RubisCO plays primarily an anaplerotic role in providing a way in which intermediates of purine/pyrimidine metabolism may be salvaged (reviewed in Tabita et al. 2007). Indeed, representative archaea have developed a novel means to synthesize RuBP (Finn & Tabita 2004; Sato et al. 2007) and require RubisCO to remove this metabolite, which has no known metabolic fate beyond serving as a substrate for RubisCO. Moreover, in some organisms, a type of RubisCO analog has evolved which does not even catalyze RuBP-dependent CO_2 or O_2 fixation.

What are these different forms of RubisCO and in what type organisms are they found? What distinguishes these enzymes? On the basis of amino acid sequences, four known forms of RubisCO, forms I, II, III, and IV, are found in nature. A summary of the different forms of RubisCO is described in Table 1. X-ray structures are available for representatives of each of these proteins, so there is considerable baseline information relative to how their function is related to their structure (Tabita et al. 2007). The most abundant form of RubisCO is the form I protein. This is the classic high molecular weight protein originally found in plants and once

called Fraction One Protein. Form I RubisCO is comprised of large, approximately 50 KDa (catalytic) subunits encoded by either the *rbcL* or *cbbL* genes. In addition, the form I holoenzyme is comprised of small, approximately 15 Kda polypeptides encoded by the *rbcS* or *cbbS* genes. [Note, in proteobacteria the *cbbL* and *cbbS* genes are often found associated with other structural genes of the CBB pathway (*cbb* genes) in discrete operons; thus these genes all have the same three letter (*cbb*) prefix]. Eight large subunits are arranged as four dimeric pairs to form a catalytic core that is decorated on the top and bottom by four small subunits to form basically an L_8S_8 structure. Form I is widespread amongst higher plants, eukaryotic algae, and bacteria. There appear to be four subclasses of form I RubisCO, termed IA, IB, IC, and ID, found in different organisms (Tabita 1999). In prokaryotes and in nongreen algae the form I genes are cotranscribed behind a single promoter, while in plants and green algae, the *rbcL* gene is chloroplast-encoded and the *rbcS* gene is nuclear-encoded.

Form II RubisCO is found in different types of proteobacteria, and in one group of eukaryotes, the dinoflagellates. It is comprised of a single large subunit, which shares only about 30 percent sequence identity with form I large subunits (Tabita 1999). Depending on the organism, different numbers of dimeric pairs comprise the quaternary structure of form II RubisCO. Indeed, the fundamental unit common to all forms of RubisCO is the large subunit dimer in which the active site is formed from the interface between the N-terminal domain of one subunit and the α/β barrel of the C-terminal domain of the second subunit. Form II proteins discriminate CO_2 from O_2 less well than form I RubisCO. Form II RubisCO is often found in organisms that also contain form I. In such instances, form II RubisCO appears not to be the major means to acquire carbon, but rather this RubisCO, along with other enzymes of the CBB pathway, functions to allow CO_2 to be used as an electron acceptor to balance the intracellular redox potential when organic carbon is oxidized (Dubbs & Tabita 2004). In such organisms, form I RubisCO appears to be selectively synthesized when carbon or CO_2 is limiting, consistent with the form I enzymes having a higher affinity for CO_2 than form II proteins (Tabita 1999).

Form III RubisCO is found (thus far) only in archaea. In these organisms, as mentioned previously, the enzyme basically serves as a means to remove RuBP, which is produced by the isomerization of ribose 1,5-bisphosphate during purine/pyrimidine metabolism (Finn &Tabita 2004, Sato et al. 2007). Many of the form III proteins thus far studied come from anaerobic extremophiles. Indeed, we have shown that several of these enzymes are highly oxygen sensitive (Finn & Tabita 2003), due largely to the extremely high affinity of these enzymes for O_2 (Kreel & Tabita 2007). Structurally, the proteins from methanogens and *Archaeoglobus fulgidus* (Finn & Tabita 2003; Watson et al. 1999) are found as dimers, although there are interesting exceptions such as the pentamer of dimers found in *Thermococcus kodakaraensis* (Ezaki et al. 1999).

Form IV is also called the RubisCO-like protein (RLP) because this protein does not catalyze RubisCO activity (Hanson & Tabita 2001). Yet, there are similarities in the primary sequence (Hanson & Tabita 2001) and tertiary structure (Imker et al. 2007; Li et al. 2005) of these proteins that clearly indicate that RLPs are homologs of RubisCO and are derived from some common ancestor (Tabita et al. 2007). RLPs cannot catalyze RubisCO activity because there are alterations in key active site residues. The *Bacillus subtilis* (Ashida et al. 2003) and *Geobacillus kaustophilus* (Imker et al. 2007) RLP (or YkrW/MtnW), the cyanobaceterial RLP from *Microcystis aeruginosa* (Carre-Mlouka et al. 2006), and RLPs from the photosynthetic bacteria *Rhodospirillum rubrum* and *Rhodopseudomonas palustris* (Singh and Tabita unpublished observations) participate in a methionine salvage pathway and catalyze the enolization of the RuBP analog, 2,3-diketo-5-methylthiopentyl-1-P in a reaction much akin to the enolase reaction catalyzed by RubisCO (Ashida et al. 2003; Imker et al. 2007). Physiological results indicate that RLP from the green sulfur bacterium *Chlorobium tepidum* is involved with some aspect of thiosulfate oxidation (Hanson & Tabita 2001; 2003), however the precise reaction catalyzed has not been identified as yet. Other clades of RLP molecules from different organisms have not yet been assigned a function; interestingly many of these latter RLP genes

do not complement RLP knockout strains from organisms with defined functions, indicative of different physiological roles (Singh & Tabita unpublished observations). Only one archaeon, *A. fulgidus*, and one eukaryote, the alga *Ostreococcus taurii*, have thus far been shown to contain an RLP gene (Tabita et al. 2007).

Phylogenetic Relationships and the Evolution of RubisCO and the RubisCO-Like Protein:

Re-evaluating the Global Ocean Survey data set's contribution to RubisCO/RLP diversity

Recent phylogenetic and bioinformatic analyses of RubisCO and RLP amino acid sequences provide a useful framework to understand the relationship of the different forms and how they may have evolved from a common ancestor (Tabita et al. 2007). Moreover, structural and functional studies impinge on these analyses as a coherent picture begins to emerge as to how the active site of this protein might have evolved and became adapted to different intracellular milieus. The overall conclusion from these studies was that a form III RubisCO from a methanogenic archaeon ancestor was the most probable source of all RubisCO and RLP lineages. Clearly, this conclusion is directly tied to the data currently available; thus it is certainly possible that additional sequences, when and if they are discovered, might cause this hypothesis to be reexamined.

In the report describing the construction and analysis of protein families encoded by DNA sequenced via the Global Ocean Survey (GOS) program (Yooseph et al. 2007), the claim was made that the GOS data contained thousands of previously unrecognized protein families and remarkably expanded our understanding of the sequence diversity present in well known protein families, including the RubisCO/RLP super family. However, the phylogeny presented for RubisCO in that work was strikingly different than those previously published by ourselves and others (Ashida et al. 2005; Carre-Mlouka et al. 2006; Hanson & Tabita 2001; Hanson & Tabita 2003; Tabita et al. 2007). To rigorously examine these claims with respect to the

RubisCO superfamily in the context of our previous work, here we present our re-analysis of the GOS-derived RubisCO sequences.

GOS sequences for analysis were collected in two batches, which were treated independently. First, all sequences within cluster 3734 used by Yooseph et al. in constructing their RubisCO/RLP phylogeny (Yooseph et al. 2007) were retrieved (Huiying Li, personal communication). Second, sequences representing each major lineage of the RubisCO superfamily were used as queries in BLASTP searches of the NCBI environmental non-redundant protein sequence database (env-nr), retaining all sequences that matched the example RubisCO/RLP sequences with E-values of less than 1×10^{-10}. Each batch of GOS-derived sequences was added independently to a FASTA file containing the recently described non-redundant collection of RubisCO and RLP amino acid sequences (Tabita et al. 2007). Both collections were then clustered with CD-Hit software (Li & Godzik 2006) at a cut off value of 83% identity, which was chosen based on the ability of this value to discriminate and preserve sub-groups observed within major RubisCO/RLP lineages (i.e. form IA vs. IB). The longest sequence seeding each cluster was retained and all sequences representing clusters of less than 200 amino acid residues in length (<50% of the average 448 $\pm$ 40 residue length of well characterized RubisCO/RLP's) were discarded. Remaining sequences were aligned by ClustalW (Thompson et al. 1994) and phylogenetic trees constructed in MEGA 4.0 (Tamura et al. 2007) using the Minimum Evolution method with a p-distance model of amino acid substitution and a gamma parameter of 1.554 for rate distributions between lineages.

Together, the approaches above created two sequence sets, one of 44 sequences from the cluster 3734 data and one of 35 sequences from the BLASTP search of env-nr. The sets had 31 sequences in common, leaving 13 unique sequences in the env-nr data (Table 2). When each set was used in concert with known RubisCO/RLP's in phylogenetic reconstructions, the trees that we observed for each set collected were nearly identical to one another and to that which we recently reported (Fig. 2). Indeed, the trees for each set reveal

206 that identical sequences were recovered and placed in extremely similar positions within each
207 tree. The minor discrepancies between the trees are likely derived from unique sequences
208 recovered within each set causing slight differences during the CD-Hit clustering of the
209 independently collected sequence groups. Both trees indicate that there is a single,
210 monophyletic group of GOS-derived RubisCO/RLP sequences that do not have any
211 correspondent in currently analyzed genome sequences. As noted in our recent review (Tabita
212 et al. 2007), the active site of this clade (IV-GOS) does not appear to be functional, thus making
213 it an RLP lineage. The branching position of the IV-GOS clade within our trees is slightly
214 different, though both positions are consistent with the group being an RLP lineage. With data
215 from GOS cluster 3734 (Fig. 2A), the IV-GOS group emerges from between the IV-NonPhoto
216 and IV-YkrW/MtnW clades. However, with GOS sequences retrieved by BLASTP (Fig. 2B), the
217 IV-GOS clade is an early branch of the RLP's relative to the bona fide RubisCO lineage. The
218 bona fide RubisCO's still appear to be a monphyletic clade in either analysis and are likely to
219 have originated from an ancestral methanogen RubisCO as per our favored model for the
220 evolution of RubisCO/RLP clades (see Fig. 9 of Tabita et al. 2007).

221 The GOS data set contained sequences that tree within every major clade of RubisCO
222 and RLP with two exceptions. No GOS sequences were placed within the IV-YkrW/MtnW or IV-
223 EnvOnly clades. The IV-YkrW/MtnW clade is restricted to Gram-positive organisms related to
224 *Bacillus subtilis* while the IV-EnvOnly clade is derived exclusively from an acid mine drainage
225 community metagenome (Tyson et al. 2004). This observation likely reflects a limited
226 distribution of *Bacillus* spp. and acidophilic microbes in the surface ocean waters that provided
227 the bulk of the GOS sequence data released to date (Yooseph et al. 2007).

228 The phylogenies presented here do not agree well with that presented by Yooseph et al
229 (Yooseph et al. 2007). In particular, the distances between major clades within the
230 RubisCO/RLP superfamily appear to be significantly shorter in the Yooseph et al. tree with a
231 larger number of intervening branches, which were interpreted to mean that a large number of

232 additional RubisCO/RLP lineages were represented in the GOS data set (Yooseph et al. 2007).
233 In addition, branch lengths between relatively closely related form I subgroups appear to be
234 unusually long in the Yooseph *et al.* RubisCO tree. These incongruencies are likely due to three
235 major differences in the approaches used for phylogenetic reconstruction.

236 The first difference relates to ORF prediction and length. Our current, as well as
237 previous analyses (Tabita et al. 2007), utilized sequences that constitute a significant fraction of
238 the length of currently known RubisCO/RLP sequences obtained from genomic sequences. In
239 many of these instances, biochemical evidence existed for the start site and protein size. In
240 contrast, the GOS data set, retrieved by either BLASTP or from cluster 3734, contains
241 sequences as short as 61 amino acids with 35-45% of the data set having less than 200 amino
242 acids (Fig. 3A). All protein start sites in the GOS sequences are predicted computationally, with
243 no supporting biochemical evidence. The inclusion of short sequences presents a major
244 difficulty in the reconstruction of phylogenetic relationships as they limit the number of
245 informative positions in an alignment and increase the weight of each informative position
246 towards determining both branching order and branch length. This problem is particularly acute
247 if complete deletion of gaps is used in calculating distances from alignments for use in trees.
248 This was not specified in details of tree construction methods used by these authors (Yooseph
249 et al. 2007).

250 Given the uncertainties in start site prediction, it is not clear whether or not short
251 sequences within the GOS data set represent authentic short variants of RubisCO/RLP or
252 fragmentary sequences of longer genes where the remainder of the sequence had not been
253 determined. One indication of fragmentation in the GOS data sets is that the initiation residue
254 for the amino acid sequences was frequently not the canonical methionine (Fig. 3B). While
255 alternative translation initiation sites do occur in prokaryotes (Makita et al. 2007; Tech &
256 Meinicke 2006), this is usually a rare occurrence where leucine or valine specified by the
257 codons T/CTG and GTG replace methionine. Over half (59%) of the ORFs in cluster 3734 data

set and just under half (47%) of GOS sequences retrieved by BLASTP did not initiate with M, L, or V while over 97% of the RubisCO/RLP sequences in our prior data set did (Fig. 3B). Therefore, it seems likely that the GOS ORF set may contain a large proportion of partial and/or corrupted ORF sequences that carry irrelevant sequence attached to true ORFs which may have driven spurious localization in phylogenetic tree reconstruction.

The observed difference in percentages of non-canonical initiation between the two GOS derived sets was found to depend on the database from which the sequences were retrieved. A large percentage of sequences in the cluster 3734 data set that initiated with a non-canonical amino acid were found in the BLASTP data set (retrieved from GenBank) to initiate with the canonical methionine. This is most easily seen by comparing the relative percentages of sequences initiating with I and M. The percentage of I-initiating sequences in the cluster 3734 data is high relative to the BLASTP data while these two groups are reversed in M-initiating sequences (i.e. BLASTP percentage is greater than cluster 3734). Within the twenty nine sequences common between the final sets used to construct the trees, this I for M substitution was observed in ten sequences, or 34%. It is not clear how this substitution came about, but when sequences were aligned, this substitution was the only difference observed. Thus, users of the GOS data should be skeptical about the identity of the initiating residue when these sequences are retrieved from Genbank.

The second major difference between our analyses and those of Yooseph et al. (2007) has to do with the assumption of constant rates of sequence change across multiple lineages within protein superfamilies. As others have noted (DeBry 1992), the failure to account for rate variation can have profound effects on branch lengths observed in phylogenetic reconstructions and also has the tendency to confound certain types of phylogenetic reconstruction methods, notably UPGMA and Maximum Parsimony. In the specific case of the RubisCO/RLP lineages, analysis of the sequence data indicates that a moderate amount of rate variation between

lineages must be included to accurately reconstruct sequence relationships (Tabita et al. 2007). Rate variation was apparently not taken into account in the neighbor-joining phylogeny presented by Yooseph et al. nor was the data set evaluated for appropriate substitution model via a tool like ProtTest (Abascal et al. 2005).

The third and final difference has to do with clustering of sequences to reduce complexity. The GOS protein sequence data were clustered at values of 100% identity and 98% similarity to produce non-redundant sets of protein sequences for further analysis. While this level will clearly delineate unique individual sequences, it is likely much too inclusive for the identification of clearly defined families or sub-families. Within all the RubisCO clades, average in-group pair-wise amino acid sequence identity range from a low of 39% within the IV-DeepYkr clade to a high of 85% within the form IB sub-family (Tabita et al. 2007). This is far below the 100% identity criterion used to cluster the GOS data, likely leading to the inclusion of unnecessary sequences from a diversity analysis standpoint.

While the value of the GOS data is clear in terms of improving our understanding of the overall genetic diversity within marine microbial communities, the re-analysis we present here of the GOS RubisCO/RLP sequences suggests that a more rigorous analysis of the claim that "…the predicted GOS proteins also add a great deal of diversity to known protein families and shed light on their evolution" must be performed on a case by case basis. In the case of RubisCO/RLP, we conclude that only one truly novel lineage of RubisCO/RLP sequences was found within the GOS data and that the major contribution was to provide evidence for the existence of genes encoding previously defined major clades of RubisCO and RLP in marine environments.

In summary, currently available curated sequence data, as well as proper phylogenetic analyses, are compatible with there being three distinct lineages of bona fide RubisCO forms (forms I, II, and III); each group contains varying numbers of sub-groups. Within the form IV

(RLP) group of the RubisCO super family, six different clades of RLP molecules have thus far been identified (Tabita et al. 2007) and confirmed in the current study.

Exploiting RubisCO Diversity to Learn More About Function

Phylogenetic, bioinformatic, and evolutional considerations provide thought-provoking discussions, but how can these analyses help us with solving some of the "mysteries" of RubisCO catalysis, such as the molecular basis for CO_2/O_2 discrimination or why the affinities for CO_2 and O_2 vary? Our approach has always been to use what nature provides. Thus, the different forms and structural adaptations of RubisCO available, from organisms that assimilate and metabolize CO_2 under diverse and even extreme environments, may provide useful insights as to how all RubisCO molecules function.

The interesting case of *Methanococcoides burtonii*

The recently sequenced genome of *Methanococcoides burtonii* revealed the presence of an ORF apparently coding for a RubisCO uniquely intermediate between form II and III (Fig. 4). The predicted protein sequence places the *M. burtonii* RubisCO (MBR) slightly closer to form II (40% identity to *Rhodobacter capsulatus* CbbM via BLASTP) than form III (35% identity to *Methanocaldococcus janaschii* RbcL). A more detailed look at the sequence, however, reveals a number of catalytic-site residues that are more similar to form II enzymes than to form III.

Biochemically, the classification of MBR is ambiguous as the enzyme has not been purified or characterized. *M. burtonii* is an obligately anaerobic methanogenic archaeon which would suggest that MBR functions in a role closer to that of previously characterized form III RubisCOs (Finn & Tabita 2004; Sato et al. 2007), namely, in pathways involved in recycling key metabolites rather than primary carbon fixation. Further, *M. burtonii* is apparently unable to grow with CO_2 as the sole carbon source (Franzmann et al. 1992). Several lines of evidence support this supposition. First, although MBR has been detected in whole-cell proteomic

studies, expression appears to be at levels consistent with pathways of secondary importance. Secondly, there is no evidence of other CBB cycle enzymes, especially phosphoribulokinase, the other enzyme unique to the CBB cycle. Finally, homologs of DeoA and E2B2 are present in the genome and are expressed under normal growth conditions (Goodchild et al. 2004a; Goodchild et al. 2005; Goodchild et al. 2004b; Saunders et al. 2005) indicative of the potential presence of an AMP-recycling pathway in which form III RubisCO is hypothesized to participate (Sato et al. 2006)

Complicating the classification of MBR, however, is the presence of a novel structural motif. Alignments of MBR amino acid sequence with form II and form III sequences indicates a 26 residue sequence (EQTWSKIMDTDKDVINLVNEDLAHHVI) near the C-terminus with no homology to extant RubisCO sequences or, indeed, to any sequence currently deposited in GenBank. SwissProt models of the MBR structure, whether using form II (9rub) or form III (1geh) as a template, suggest the presence of a loop (Fig. 5), possibly forming an anti-parallel β-sheet motif. This predicted motif is distant from the active site and may play a role in maintaining holoenzyme structure without directly impacting catalysis; certainly this remains to be determined. Given the huge array of sequences already known, however, this novel motif indicates that, at the very least, the RubisCO family still possesses surprises.

Interactions of archaeal RubisCO with oxygen

As discussed above, the profusion of genomic sequencing projects has played a dramatic role in altering perceptions as to how RubisCO might have evolved. With respect to structure/function studies, the discovery of form III enzymes from anaerobic archaea is particularly cogent as such organisms obviously evolved in the complete absence of oxygen. In particular, the finding that several representative archaeal RubisCO enzymes exhibit unusually high sensitivity to molecular oxygen, even in the presence of high levels of CO_2 (Watson et al. 1999; Finn & Tabita 2003; Kreel & Tabita 2007), was deemed to be particularly interesting.

Thus, experiments were initiated to determine the basis for this sensitivity to O_2, with the thought that these findings would relate or provide clues as to how all forms of RubisCO interact with O_2. Using the RubisCO from *Archaeoglobus fulgidus* as a model, it was shown that O_2 sensitivity correlated with the fact that this enzyme showed an extremely high affinity for O_2, with a K_o of 5 μM (Kreel & Tabita), some two orders of magnitude lower than the values obtained for form I or form II enzymes K_o (which vary from 500 to 1000 μM). Two residues, Met-295 and Ser-363, of the *A. fulgidus* enzyme appeared to influence the K_o value. Changing Met-295 to an aspartate (M295D) enhanced the ability of the enzyme to recover from O_2 inactivation with a consequent increase in the K_o to 24 μM. This was accompanied by a three-fold increase in the substrate specificity factor. Similar results were obtained with Ser-363, a residue which sits in a hydrophobic pocket near the active site. Moreover, double mutants (I.e., M295D/S363I) showed an additive effect in the ability of such an enzyme to recover from oxygen inactivation (Kreel and Tabita 2007). The M295D/S363I enzyme has a K_O of about 400-500 μM (Kreel & Tabita, manuscript in preparation), a K_O that approximates the value for form I (aerobic and plant) and form II RubisCO (i.e., nearly 100-fold greater than the wild-type *A. fulgidus* enzyme). Despite a trade off in which the k_{cat} has been reduced, this double mutant might serve as the starting point to select for a high activity, high K_o enzyme.

Structurally, Met-295 of the *A. fulgidus* enzyme was found to be in close proximity to a highly conserved residue, Arg-279, found in all other forms of RubisCO and shown to be necessary for substrate (RuBP) binding (Zhang et al. 1994). The model structure suggests that a mutation to an aspartate residue at the Met-295 position would allow for an interaction between one of the hydroxyl side chains of the aspartate residue with one of the side chain nitrogen atoms of Arg-279. In wild-type *A. fulgidus* RubisCO, this hydrogen bond is absent. However, there is definite hydrogen-bonding to the equivalent Arg residue in all other form I and form II RubisCO structures. In addition, the model structure of *A. fulgidus* RubisCO shows an interaction of the side chain of Ser-363 with highly conserved and catalytically important

residues Gly-313 and Thr-314 of *A. fulgidus* RubisCO (Fig. 6). Gly-313 and Thr-314, found in all forms of RubisCO, show no ionic interactions with the amino acid residue equivalent to Ser-363 of *A. fulgidus* RubisCO in form I and form II enzymes. Thus, not only does a mutation in Ser-363 influence O_2-sensitivity of the *A. fulgidus* enzyme but the degree of conservation of residues in this region may also provide clues as to the importance of this region in all RubisCO enzymes. In the model structure of the S363I and S363V mutants, it appears as though the introduction of a bulky hydrophobic amino acid in the hydrophobic pocket not only eliminates the hydrogen bonding interaction with highly conserved Gly-313 and Thr-314 (Fig. 6), but these substitutions may also cause conformational changes that likely affect the folding of the enzyme either before or during catalysis.

Interactions of form I RubisCO with oxygen

Previous work showed that directed enzyme evolution procedures might be applied to bioselect useful mutant forms of prokaryotic RubisCO after random mutagenesis using a *Rhodobacter capsulatus* host with its host RubisCO genes deleted (Smith & Tabita 2003; 2004). This system (using *R. capsulatus* RubisCO deletion strain SBI/II$^-$) offers many advantages, including the ability to select mutant enzymes that may or may not allow growth under aerobic conditions due to specific changes in the enzyme. During the course of isolating suppressor mutations that overcame the negative effects of a previously isolated D103V cyanobacterial (*Synechococcus* sp. strain PCC6301) RubisCO mutant, a residue was identified, Ala-375, that proved to be extremely interesting in another context (Satagopan et al. submitted for publication). The original suppressor, a D103V/A375V double mutant, was able to support growth of *R. capsulatus* under anaerobic conditions where the D103V enzyme could not. Moreover, upon kinetic analysis of the D103V/A375V, as well as an A375V mutant, it was found that the K_o of both enzymes was considerably increased. Interestingly, structural analysis indicated that Ala-375 was situated in a hydrophobic pocket similar to that previously discussed

412 for the archaeal (*A. fulgidus*) enzyme. Indeed, Ala-375 of the form I cyanobacterial enzyme is
413 equivalent to residue Ser-363 of the form III archaeal enzyme and this hydrophobic pocket is
414 found in all forms of bona fide RubisCO (Fig. 7).

415 Since mutations of both Ser-363 of the *A. fulgidus* form III (archaeal) enzyme and Ala-
416 375 of the *Synechococcus* form I enzyme affected the K_o, it was surmised that the A375V
417 mutant enzyme might show a phenotypic response, especially under aerobic growth conditions.
418 Thus, the affect of the A375V mutation was determined using the *R. capsulatus* system; i.e. *R.*
419 *capsulatus* strain SBI/I$^-$ was complemented with the gene containing the A375V change (Fig. 8).
420 In this experiment it is clear that only enzymes containing the A375V mutation are able to
421 support luxuriant growth in strain SBI/II$^-$ (equivalent to wild-type *R. capsulatus* strain SB1003) in
422 a complex medium under aerobic conditions. Neither the wild-type nor the D103V mutant
423 *Synechococcus rbcL* gene supported luxuriant growth in strain SBI/II$^-$ using complex media (Fig.
424 8). Furthermore, under aerobic chemoautotrophic conditions, with CO_2 as the sole source of
425 carbon, only the A375V mutant supported substantial growth (Satagopan et al, submitted for
426 publication). These studies indicate that residue Ala-375 influences the ability of RubisCO to
427 interact with oxygen and its effect is manifested in a physiologically significant way.
428
429 **Conclusions**

430 From the analyses presented here it is clear that recent claims for new RubisCO families
431 that fall outside the three classes of bona fide RubisCO proteins (forms I, II, and III) and the six
432 clades of form IV or RLP molecules are not tenable. RubisCO from extremophile organisms that
433 live in the absence of oxygen also provide convenient systems to learn more about how the
434 active site of the enzyme has evolved to accommodate this special milieu. These studies also
435 provide clues as to how RubisCO may be modified to enhance oxygen insensitivity in form I
436 type enzymes.
437

Acknowledgments

FRT was supported by grant GM24497 from the National Institutes of Health and grants DE-FG02-01ER63241 and DE-FG02-91ER20033 from the offices of Biological & Environmental Research (Genomics: GTL Program) and Energy Biosciences, respectively, of the U. S. Department of Energy. TEH was supported by by NSF Career Award MCB-0447649.

Literature Cited

Abascal, F., Zardoya, R. & Posada, D. 2005 ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105.

Ashida, H., Danchin, A. & Yokota, A. 2005 Was photosynthetic RubisCO recruited by acquisitive evolution from RubisCO-like proteins involved in sulfur metabolism? *Res. Microbiol.* **156**, 611-618.

Ashida, H., Saito, Y., Kojima, C., Kobayashi, K., Ogasawara, N. & Yokota, A. 2003 A functional link between RubisCO-like protein of Bacillus and photosynthetic RubisCO. *Science* **302**, 286-290.

Carre-Mlouka, A., Mejean, A., Quillardet, P., Ashida, H., Saito, Y., Yokota, A., Callebaut, I., Sekowska, A., Dittmann, E., Bouchier, C. & de Marsac, N. T. 2006 A new RubisCO-like protein coexists with a photosynthetic RubisCO in the planktonic cyanobacteria Microcystis. *J. Biol. Chem.* **281**, 24462-24471.

DeBry, R. W. 1992 The consistency of several phylogeny-inference methods
under varying evolutionary rates. *Mol. Biol. Evol.* **9**, 537-551.

Dubbs, J. M. & Tabita, F. R. 2004 Regulators of nonsulfur purple phototrophic bacteria and the interactive control of CO2 assimilation, nitrogen fixation, hydrogen metabolism and energy generation. *FEMS Microbiol. Rev.* **28**, 353-376.

Ezaki, S., Maeda, N., Kishimoto, T., Atomi, H. & Imanaka, T. 1999 Presence of a structurally novel type ribulose-bisphosphate carboxylase/oxygenase in the hyperthermophilic archaeon, Pyrococcus kodakaraensis KOD1. *J. Biol. Chem.* **274**, 5078-5082.

Finn, M. W. & Tabita, F. R. 2003 Synthesis of catalytically active form III ribulose 1,5-bisphosphate carboxylase/oxygenase in archaea. *J. Bacteriol.* **185**, 3049-3059.

Finn, M. W. & Tabita, F. R. 2004 Modified pathway to synthesize ribulose 1,5-bisphosphate in methanogenic archaea. *J Bacteriol* **186**, 6360-6366.

Franzmann, P. D., Springer, N., Ludwig, W., Demacario, E. C. & Rohde, M. 1992 A Methanogenic Archaeon from Ace Lake, Antarctica - *Methanococcoides-Burtonii* Sp-Nov. *System. & Applied Microbiol.* **15**, 573-581.

Goodchild, A., Raftery, M., Saunders, N. F. W., Guilhaus, M. & Cavicchioli, R. 2004a Biology of the cold adapted archaeon, Methanococcoides burtonii determined by proteomics using liquid chromatography-tandem mass spectrometry. *J.Proteome Res.* **3**, 1164-1176.

Goodchild, A., Raftery, M., Saunders, N. F. W., Guilhaus, M. & Cavicchioli, R. 2005 Cold adaptation of the Antarctic archaeon, Methanococcoides burtonii assessed by proteomics using ICAT. *J.Proteome Res.* **4**, 473-480.

Goodchild, A., Saunders, N. F. W., Ertan, H., Raftery, M., Guilhaus, M., Curmi, P. M. G. & Cavicchioli, R. 2004b A proteomic determination of cold adaptation in the Antarctic archaeon, Methanococcoides burtonii. *Mol.Microbiol.* **53**, 309-321.

505 Hanson, T. E. & Tabita, F. R. 2001 A ribulose-1,5-bisphosphate carboxylase/oxygenase
506 (RubisCO)-like protein from Chlorobium tepidum that is involved with sulfur metabolism and the
507 response to oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4397-402.
508
509 Hanson, T. E. & Tabita, F. R. 2003 Insights into the stress response and sulfur metabolism
510 revealed by proteome analysis of a Chlorobium tepidum mutant lacking the RubisCO-like
511 protein. *Photosynth. Res.* **78**, 231-248.
512
513 Imker, H. J., Fedorov, A. A., Fedorov, E. V., Almo, S. C. & Gerlt, J. A. 2007 Mechanistic
514 diversity in the RubisCO superfamily: the "enolase" in the methionine salvage pathway in
515 Geobacillus kaustophilus. *Biochemistry* **46**, 4077-4089.
516
517 Kreel, N. E. & Tabita, F. R. 2007 Substitutions at methionine 295 of *Archaeoglobus fulgidus*
518 ribulose-1,5-bisphosphate carboxylase/oxygenase affect oxygen binding and CO2/O2
519 specificity. *J. Biol. Chem.* **282**, 1341-1351.
520
521 Li, H., Sawaya, M. R., Tabita, F. R. & Eisenberg, D. 2005 Crystal structure of a RubisCO-like
522 protein from the green sulfur bacterium *Chlorobium tepidum*. *Structure* **13**, 779-789.
523
524 Li, W. & Godzik, A. 2006 Cd-hit: a fast program for clustering and comparing large sets of
525 protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.
526
527 Long, S. P., Ainsworth, E. A., Leakey, A. D. B., Nosberger, J. & Ort, D. R. 2006a. Food for
528 thought: lower-than-expected crop yield stimulation with rising CO_2 concentrations. *Science* **312**,
529 1918-1921.
530

Long, S. P., Zhu, X.-G., Naidu, S. L. & ort, D. R. 2006b. Can improvement in photosynthesis increase crop yields? *Plant Cell & Environ.* **29**, 315-330.

Makita, Y., de Hoon, M. J. & Danchin, A. 2007 Hon-yaku: a biology-driven Bayesian methodology for identifying translation initiation sites in prokaryotes. *BMC Bioinformatics* **8**, 47.

Sato, T., Atomi, H. & Imanaka, T. 2007 Archaeal type III RubisCOs function in a pathway for AMP metabolism. *Science* **315**, 1003-1006.

Saunders, N. F. W., Goodchild, A., Raftery, M., Guilhaus, M., Curmi, P. M. G. & Cavicchioli, R. 2005 Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon Methanococcoides burtonii. *J. Proteome Res.* **4**, 464-472.

Smith S. A. & Tabita F. R. 2003 Positive and negative bioselection of mutant forms of prokaryotic (cyanobacterial) ribulose-1, 5-bisphosphate carboxylase/oxygenase. *J.Mol.Biol.* **331,** 557-569.

Smith, S. A. & Tabita, F. R. 2004 Glycine 176 affects catalytic properties and stability of the *Synechococcus* sp. strain PCC 6301 ribulose 1,5-bisphosphate carboxylase/oxygenase. *J. Biol. Chem.* **279**, 25632-25637.

Tabita, F. R. 1999 Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: A different perspective. *Photosynth. Res.* **60**, 1-28.

556 Tabita, F. R., Hanson, T. E., Li, H., Satagopan, S., Singh, J. & Chan, S. 2007 Function,
557 structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol.*
558 *Mol. Biol. Rev.* **71 In Press**.
559
560 Tamura, K., Dudley, J., Nei, M. & Kumar, S. 2007 MEGA4: Molecular Evolutionary Genetics
561 Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596-1599.
562
563 Tech, M. & Meinicke, P. 2006 An unsupervised classification scheme for improving predictions
564 of prokaryotic TIS. *BMC Bioinformatics* **7**, 121.
565
566 Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994 CLUSTAL W: improving the sensitivity of
567 progressive multiple sequence alignment through sequence weighting, position-specific gap
568 penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
569
570 Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M.,
571 Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. & Banfield, J. F. 2004 Community structure and
572 metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-
573 43.
574
575 Watson, G. M., Yu, J. P. & Tabita, F. R. 1999 Unusual ribulose 1,5-bisphosphate
576 carboxylase/oxygenase of anoxic Archaea. *J. Bacteriol.* **181**, 1569-1575.
577
578 Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen,
579 J. A., Heidelberg, K. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S., Li, H.,
580 Mashiyama, S. T., Joachimiak, M. P., van Belle, C., Chandonia, J. M., Soergel, D. A., Zhai, Y.,
581 Natarajan, K., Lee, S., Raphael, B. J., Bafna, V., Friedman, R., Brenner, S. E., Godzik, A.,

582 Eisenberg, D., Dixon, J. E., Taylor, S. S., Strausberg, R. L., Frazier, M. & Venter, J. C. 2007 The
583 Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families.
584 *PLoS Biol.* **5**, e16.

585

586 Zhang, K. Y., Cascio, D., & Eisenberg, D. 1994. Crystal structure of the unactivated ribulose
587 1,5-bisphosphate carboxylase/oxygenase complexed with a transition state analog, 2-carboxy-
588 D-arabinitol 1,5-bisphosphate. *Protein Sci.* **3**, 64-69.

589

590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609

610 **Table I. Summary of the properties of different forms of RubisCO**

RubisCO form	Quaternary structure	Types of organisms
I	L_8S_8	Plants, algae, proteobacteria, cyanobacteria
II	$(L_2)_n$	Proteobacteria, dinoflagellates
III	$(L_2)_n$	Archaea
IV	L_2	Proteobacteria, cyanobacteria, archaea, algae

611 **Table 2. All sequences used in this study.**

Family	Organism or Environment	Accession	Label
IA	Bradyrhizobium sp. BTAi1	ZP_00862357	Bra.sp-BTAi1-1
	Hydrogenophilus thermoluteolus	Q51856	Hyd.the
	Nitrobacter vulgaris	Q59613	Nit.vul
IB	Anabaena variabilis ATCC 29413	ABA23512	Ana.var-ATCC 29413
	Nicotiana tabacum	P00876	Nic.tab
IC	Rubrivivax gelatinosus PM1	ZP_00243775	IC-1-Rub.gel
	Xanthobacter autotrophicus Py2	ZP_01200598	Xan.aut-Py2-1
	Chlorothrix halobium	D. Bryant Pers Comm	ChlhalRbc1
ID	Acidiphilium cryptum JF-5	ZP_01146719	Aci.cry-JF-5-1
	Aurantimonas sp. SI85-9A1	AAB41464	Mn-Ox-SI85-9A1
	Burkholderia xenovorans LB400	ZP_00283421	ID-Bur.xen-LB400
	Nitrosococcus oceani ATCC 19707	ABA56859	Nit.oce-ATCC 19707
	Nitrosospira multiformis ATCC 251966	YP_411385	Nit.mul-ATCC251966
	Odontella sinensis	NP_043654	Odo.sin
	Pleurochrysis carterae	Q08051	Ple.car
II	Hydrogenovibrio marinus	Q59462	II-Hyd.mar
	Lingulodinium polyedrum	AAA98748	Lin.pol
	Magnetospirillum magnetotacticum AMB-1	YP_422059	Mag.mag-AMB-1
	Rhodoferax ferrireducens T118	YP_522655	Rho.fer-T118
		CAA25080	II-Rho.rub
	Symbiodinium sp.	AAG37859	Sym.sp
	Thiomicrospira crunogena XCL-2	ABB41020	II-Thi.cru-XCL-2
	Methanococcoides burtonii DSM 6242	ZP_00563653	Met.bur-DSM6242
Unaffiliated archaea	Methanosaeta thermophila PT	ZP_01153096	Met.the-PT
	Methanospirillum hungatei JF-1	YP_503739	Met.hun-JF-1
III	Archaeoglobus fulgidus DSM 4304	NP_070466	III-Arc.ful-DSM 4304
	Haloferax volcanii	contig 3020 [1]	Hal.vol
	Hyperthermus butylicus DSM 5456	YP_001012710	Hyp.the-DSM5456
	Methanocaldococcus jannaschii	AAB99239	Met.jan
	Methanoculleus marisnigri JR1	ZP_01391373	Met.cul-JR1
	Methanosarcina acetivorans C2A	AAM07894	Met.ace-C2A
	Natronomonas pharaonis DSM 2160	CAI49476	Nat.pha-DSM 2160
	Pyrococcus horikoshii OT3	BAA30036	Pyr.hor-OT3
	Thermococcus kodakaraensis KOD1	BAD86479	The.kod-KOD1
	Thermofilum pendens Hrk 5	YP_920628	The.pen-Hrk5
IV-Non-photo	Acidiphilium cryptum JF-5	ZP_01146529	Aci.cry-JF-5-2
	Bordetella bronchiseptica RB50	CAE31534	Bor.bro-RB50
	Burkholderia xenovorans LB400	ZP_00284840	IV-Bur.xen-LB400
	Chromohalobacter salexigens DSM 3043	ZP_00471249	Chr.sal-DSM 3043
	Delftia acidovorans SPH1	ZP_01577127	Del.aci-SPH1
	Fulvimarina pelagi HTCC2506	ZP_01438569	Ful.mar-HTCC2506
	Jannaschia sp. CCS1	YP_511005	Jan.sp-CCS1
	Mesorhizobium loti	BAB53192	Mes.lot
	Polaromonas sp. JS666	ZP_00502320	IV-1-Pol.sp-JS666
	Polaromonas sp. JS666	ZP_00502381	IV-2-Pol.sp-JS666

	Pseudomonas putida F1	ZP_00900417	Pse.put-F1
	Rhizobium leguminosarum	pRL120396 [2]	Rhi.leg
	Roseobacter sp. MED193	ZP_01056409	Ros.sp-MED193
	Sinorhizobium meliloti 1021	CAC48779	IV-Sin.mel
	Thermomicrobium roseum	contig:2220 [3]	The.ros
	Xanthobacter autotrophicus Py2	ZP_01199940	Xan.aut-Py2-2
IV-DeepYkr	Alkalilimnicola ehrlichei MLHE-1	YP_742007	Alk.ehr-MLHE-1-2
	Halorhodospira halophila SL1	YP_001002057	Hal.hal-SL1-2
	Heliobacillus mobilis	ABH04879	Hel.mob
	Ostreococcus tauri	Ot07g01830 [4]	IV-1-Ost.tau
	Ostreococcus tauri	Ot08g02600 [4]	IV-2-Ost.tau
	Rhodopseudomonas palustris BisA53	YP_782588	IV-1-Rho.pal-BisA53
	Rhodopseudomonas palustris BisB18	YP_532057	IV-1-Rho.pal-BisB18
	Rhodopseudomonas palustris BisB5	YP_569369	IV-1-Rho.pal-BisB5
	Rhodopseudomonas palustris CGA009	CAE27610	IV-1-Rho.pal-CGA009
	Rhodospirillum rubrum	ABC22798	IV-Rho.rub
IV-EnvOnly	Acid Mine Consortium	AADL01000066	AMC-066
	Acid Mine Consortium	AADL01000179	AMC-179
	Acid Mine Consortium	AADL01000602	AMC602
IV-Photo	Allochromatium vinosum	BAB44150	IV-All.vin
	Chlorobium chlorochromatii CaD3	ABB28892	Chl.chl-CaD3
	Chlorobium phaeobacteroides DSM 266	ZP_00527577	Chl.pha-DSM 266
	Chlorobium tepidum TLS1	AAM72993	Chl.tep-TLS1
	Prosthecochloris aestuarii DSM 271	ZP_00590874	Pro.aes-DSM 271
	Rhodopseudomonas palustris BisB18	YP_530146	IV-2-Rho.pal-BisB18
IV-YkrW	Acid Mine Consortium	AADL01000541	AMC-541
	Bacillus cereus E33L	AAU16474	Bac.cer-E33L
	Bacillus clausii KSM-K16	BAD64310	Bac.cla-KSM-K16
	Bacillus licheniformis ATCC 14580	AAU23062	Bac.lic-ATCC-14580
	Bacillus subtilis subsp. subtilis str. 168	CAB13232	Bac.sub-168
	Exiguobacterium sibiricum 255-15	ZP_00539172	Exi.sib-255-15
	Geobacillus stearothermophilus	Contig321 [5]	Geo.ste
Unique GOS Sequences	GOS Cluster 3734	EB060441	EB060441
		EBH72800	EBH72800
		EBK38093	EBK38093
		ECC63254	ECC63254
		ECH76386	ECH76386
		ECI09606	ECI09606
		ECK47660	ECK47660
		ECM15273	ECM15273
		ECN61796	ECN61796
		ECV58360	ECV58360
		ECV93956	ECV93956
		ECW52658	ECW52658
		EDE24211	EDE24211
	BLASTP vs. env_nr	EBH72800	EBH72800
		ECE36765	ECE36765
		EDE26668	EDE26668

Common to 3734 and env_nr	EBE68134	EBE68134
	EBG47321	EBG47321
	EBH57905	EBH57905
	EBJ85584	EBJ85584
	EBK48460	EBK48460
	EBO94763	EBO94763
	EBO96920	EBO96920
	EBS14603	EBS14603
	EBV75076	EBV75076
	ECA82964	ECA82964
	ECD28674	ECD28674
	ECE21553	ECE21553
	ECO29831	ECO29831
	ECQ06045	ECQ06045
	ECQ25263	ECQ25263
	ECQ91781	ECQ91781
	ECR75886	ECR75886
	ECU18502	ECU18502
	ECU62179	ECU62179
	ECV95656	ECV95656
	ECW18074	ECW18074
	ECW49692	ECW49692
	EDD64574	EDD64574
	EDE27295	EDE27295
	EDG90708	EDG90708
	EDH70269	EDH70269
	EDH87249	EDH87249
	EDJ07695	EDJ07695
	EDJ25372	EDJ25372
	EDJ35923	EDJ35923
	EDJ36865	EDJ36865

1-Source: http://zdna2.umbi.umd.edu/Public/hvo/
2-Source: http://www.sanger.ac.uk/Projects/R_leguminosarum/
3-Source: http://tigrblast.tigr.org/ufmg/index.cgi?database=t_roseum%7Cseq
4-Source: http://bioinformatics.psb.ugent.be/blast/public/?project=ostreococcus
5-Source: http://www.genome.ou.edu/bstearo_blast.html

612

Figure Legends

Fig. 1. Reactions catalyzed by RubisCO showing the inherent problem that aerobic organisms face as a result of the O_2 fixation reaction catalyzed by the enzyme. CO_2/O_2 substrate discrimination may be determined experimentally as described in the text.

Fig. 2. Minimum evolution phylogenetic trees of the RubisCO and RLP protein superfamily. (A) Tree built with sequences derived from GOS cluster 3734. (B) Tree built with GOS sequences derived from BLASTP searches of the GenBank env_nr database. Major lineages are labeled and color coded identically in each panel. Bona fide RubisCO lineages were given individual colors (I-green and red, II-purple, III-gold), while all RLP's are colored Cyan except for the unique GOS RLP lineage, IV-GOS, which is colored red-brown. Abbreviations for non-GOS sequences are listed in Table X. GOS sequences are identified by their NCBI accession number.

Fig. 3. Comparison of starting data sets used to construct phylogenetic trees. (A) The percentage of total sequences falling into a given length category are shown for sequences in GOS cluster 3734 (dark grey bars, n = 195), GOS RubisCO/RLP sequences retrieved by BLASTP from Genbank (light grey bars, n = 377), and the set previously used for phylogenetic reconstruction (white bars, n = 191). B) Distribution of initiating residues for RubisCO and RLP sequences within data sets used to construct phylogenetic trees. The initiating residue is indicated on the horizontal axis. The bars for panel (B) denote the same data sets as in panel (A).

Fig. 4. Phylogenetic position of *Methanococcoides burtonii* (*M. bur* in figure)) RubisCO amongst known form II and form III RubisCO sequences. This is a subset of the full RubisCO phylogenetic tree (Fig. 2) that highlights the *M. burtonii* enzyme's unique position.

Fig. 5. Modeled structure of the *M. burtonii* RubisCO with the loop structure shown in gold. The red bar represents the approximate position of RuBP at the active site.

Fig. 6. The hydrophobic pocket of the *Archaeoglobus fulgidus* form III RubisCO showing interactions of Ser-363 with conserved residues Gly-313 and Thr-314.

Fig. 7. Hydrophobic pocket regions surrounding equivalent residues (in green) implicated in oxygen interactions. (A) Ala-375 of the form I *Synechococcus* PCC6301 RubisCO, (B) Ile-367 of the form II *Rhodospirillum rubrum* enzyme, and (C) Ser-366 of the archaeal *Pyrococcus kodakaraensis* form III enzyme. The adjacent active-site Ser residue is colored blue in all three structures. Ligands at the active-site, i.e. CABP in (A) and sulfate in (C) are colored black. Residue side-chains colored in yellow are within 4 Å of the Ala-375/Ile-367/Ser-366 of the form I/II/III proteins, respectively.

Fig. 8. Phenotypic response of mutant form I *Synechococcus* PCC6301 RubisCO constructs in *R. capsulatus* strain SBI/II$^-$. All strains were grown under aerobic conditions in a complex peptone-yeast extract media. Sectors contained (A) *R. capsulatus* wild-type strain SB1003, (B) *R. capsulatus* strain SBI/II$^-$ with no complementing RubisCO gene; (C) *R. capsulatus* strain SBI/II$^-$ complemented with the wild-type *Synechococus* PCC 6301 *rbcLS* genes; (D) *R. capsulatus* strain SBI/II$^-$ complemented with the D103V mutant *Synechococus* PCC 6301 *rbcL* gene and wild-type *rbcS*; (E) *R. capsulatus* strain SBI/II$^-$ complemented with the A375V mutant *Synechococus* PCC 6301 *rbcL* gene and wild-type *rbcS*; (F) *R. capsulatus* strain SBI/II$^-$

complemented with the D103V/A375V mutant *Synechococus* PCC 6301 *rbcL* gene and wild-type *rbcS*.

Fig. 1

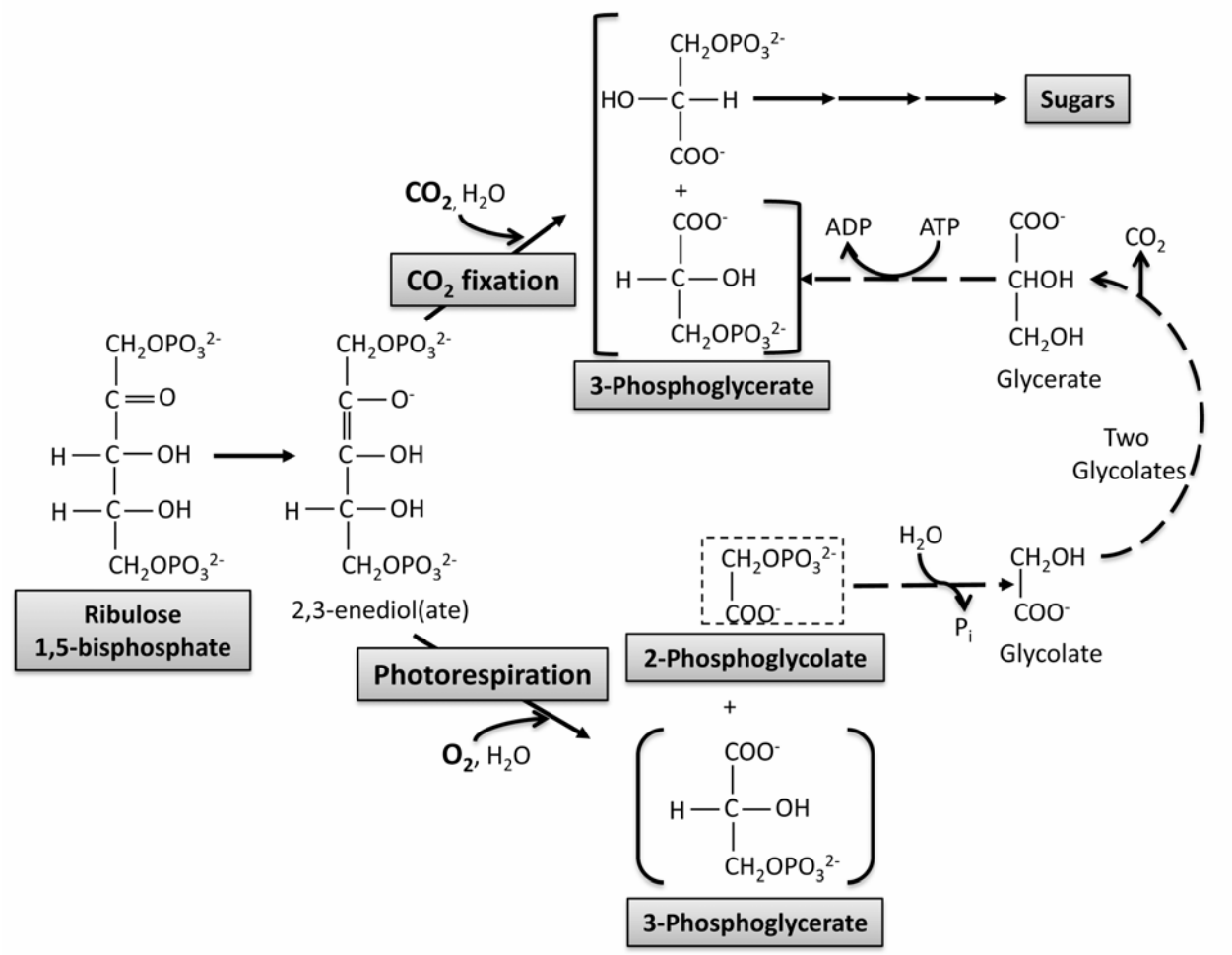

Sugars
CO_2, H_2O
CO_2 fixation
3-Phosphoglycerate
ADP
ATP
Glycerate
CO_2
Two Glycolates
Ribulose 1,5-bisphosphate
2,3-enediol(ate)
Photorespiration
O_2, H_2O
H_2O
P_i
2-Phosphoglycolate
Glycolate
3-Phosphoglycerate

Fig. 2A

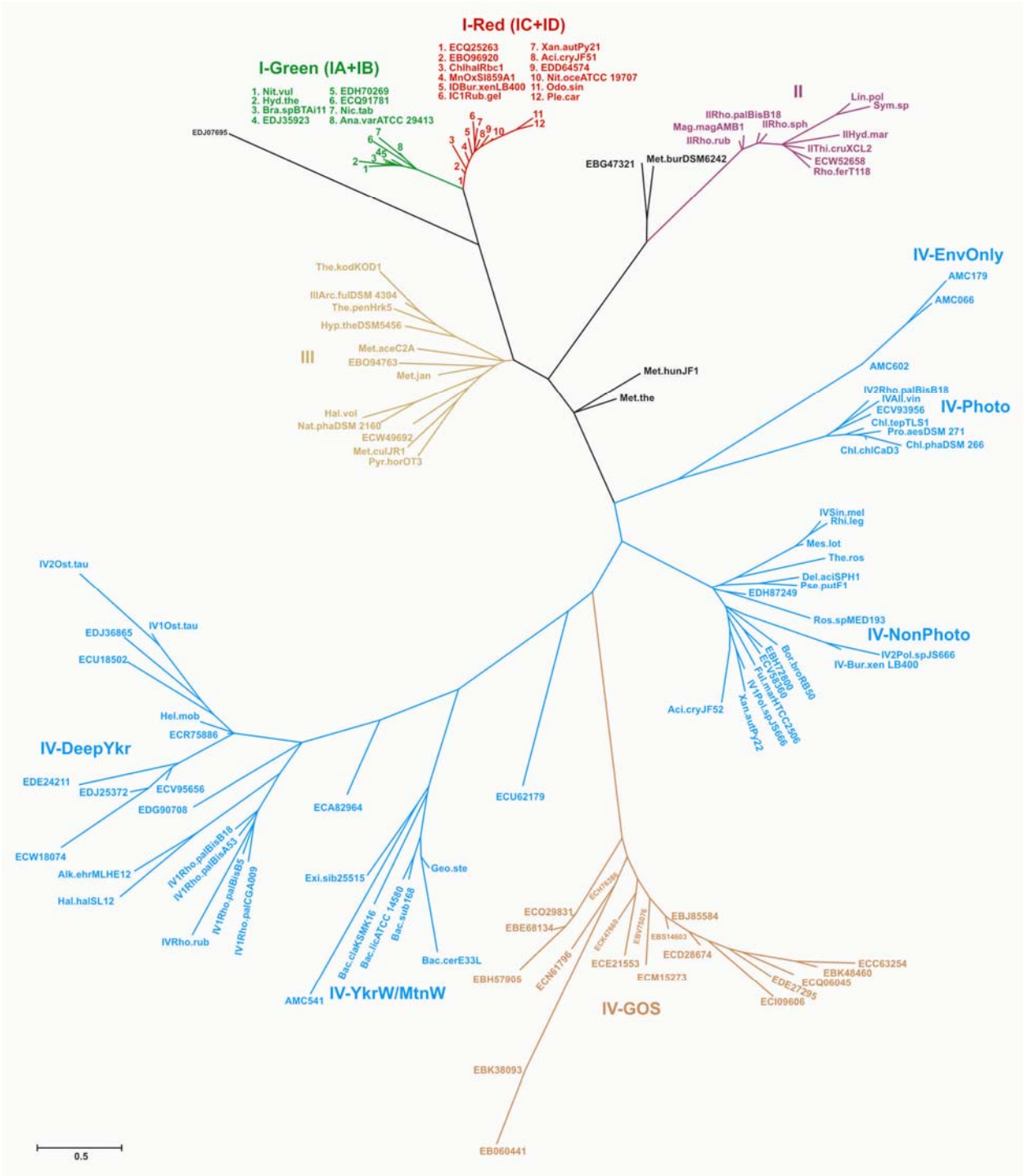

I-Red (IC+ID)
1. ECQ25263
2. EBO96920
3. ChlhalRbc1
4. MnOxSi859A1
5. IDBur.xenLB400
6. IC1Rub.gel
7. Xan.autPy21
8. Aci.cryJF51
9. EDD64574
10. Nit.oceATCC 19707
11. Odo.sin
12. Ple.car
I-Green (IA+IB)
1. Nit.vul
2. Hyd.the
3. Bra.spBTAi11
4. EDJ35923
5. EDH70269
6. ECQ91781
7. Nic.tab
8. Ana.varATCC 29413
II
Lin.pol
Sym.sp
IIRho.palBisB18
IIRho.sph
Mag.magAMB1
IIRho.rub
IIHyd.mar
IIThi.cruXCL2
ECW52658
Rho.ferT118
EDJ07695
EBG47321
Met.burDSM6242
IV-EnvOnly
AMC179
AMC066
AMC602
III
The.kodKOD1
IIIArc.fulDSM 4304
The.penHrk5
Hyp.theDSM5456
Met.aceC2A
EBO94763
Met.jan
Met.hunJF1
Met.the
Hal.vol
Nat.phaDSM 2160
ECW49692
Met.culJR1
Pyr.horOT3
IV2Rho.palBisB18
IVAll.vin
ECV93956
IV-Photo
Chl.tepTLS1
Pro.aesDSM 271
Chl.phaDSM 266
Chl.chlCaD3
IVSin.mel
Rhi.leg
Mes.lot
The.ros
Del.aciSPH1
Pse.putF1
EDH87249
Ros.spMED193
IV-NonPhoto
IV2Pol.spJS666
IV-Bur.xen LB400
Bor.broRB50
EBH72800
ECV92360
Ful.mahHTCC2506
IV1Pol.spJS666
Xan.autPy22
Aci.cryJF52
IV2Ost.tau
IV1Ost.tau
EDJ36865
ECU18502
Hel.mob
ECR75886
IV-DeepYkr
EDE24211
EDJ25372
ECV95656
EDG90708
ECW18074
Alk.ehrMLHE12
Hal.halSL12
IV1Rho.palBisB18
IV1Rho.palBisA53
IV1Rho.palBisB5
IV1Rho.palCGA009
IVRho.rub
ECA82964
ECU62179
Exi.sib25515
Geo.ste
Bac.claKSMK16
Bac.licATCC 14580
Bac.subT168
Bac.cerE33L
AMC541
IV-YkrW/MtnW
ECO29831
EBE68134
EBH57905
ECN61796
ECH78286
EBV75076
EBS14603
EBJ85584
ECD28674
ECE21553
ECM15273
ECC63254
EBK48460
ECQ06045
EDE27295
ECI09606
IV-GOS
EBK38093
EB060441
0.5

Fig. 2B

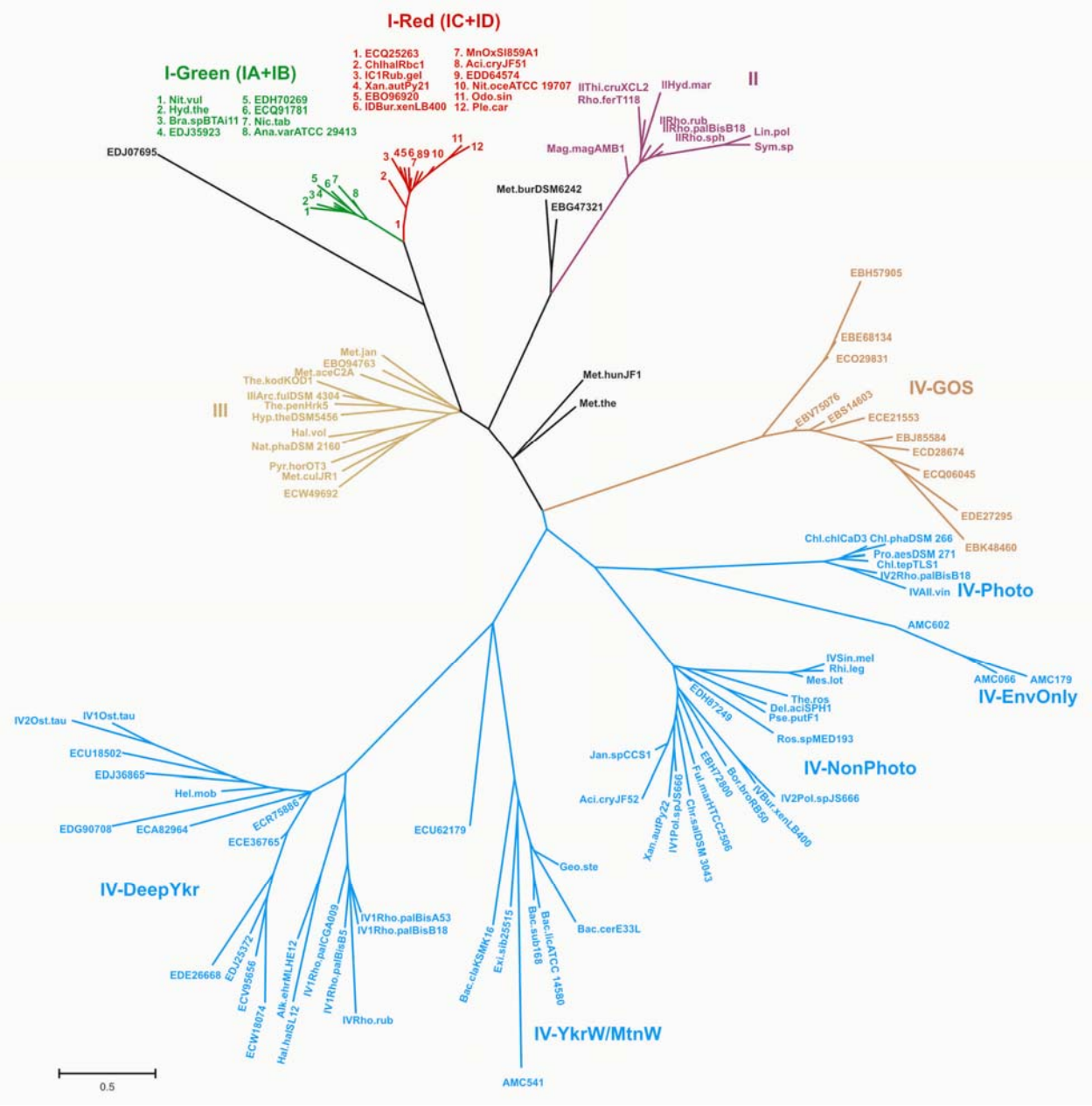

I-Red (IC+ID)
1. ECQ25263
2. ChlhalRbc1
3. IC1Rub.gel
4. Xan.autPy21
5. EBO96920
6. IDBur.xenLB400
7. MnOxSI859A1
8. Aci.cryJF51
9. EDD64574
10. Nit.oceATCC 19707
11. Odo.sin
12. Ple.car
I-Green (IA+IB)
1. Nit.vul
2. Hyd.the
3. Bra.spBTAi11
4. EDJ35923
5. EDH70269
6. ECQ91781
7. Nic.tab
8. Ana.varATCC 29413
II
IIThi.cruXCL2
Rho.ferT118
IIHyd.mar
IIRho.rub
IIRho.palBisB18
IIRho.sph
Lin.pol
Sym.sp
Mag.magAMB1
EDJ07695
Met.burDSM6242
EBG47321
Met.hunJF1
Met.the
III
Met.jan
EBO94763
Met.aceC2A
The.kodKOD1
IIIArc.fulDSM 4304
The.penHrk5
Hyp.theDSM5456
Hal.vol
Nat.phaDSM 2160
Pyr.horOT3
Met.culJR1
ECW49692
IV-GOS
EBH57905
EBE68134
ECO29831
EBV75076
EBS14603
ECE21553
EBJ85584
ECD28674
ECQ06045
EDE27295
EBK48460
IV-Photo
Chl.chlCaD3
Chl.phaDSM 266
Pro.aesDSM 271
Chl.tepTLS1
IV2Rho.palBisB18
IVAll.vin
AMC602
IV-EnvOnly
AMC066
AMC179
IVSin.mel
Rhi.leg
Mes.lot
EDH87249
The.ros
Del.aciSPH1
Pse.putF1
Ros.spMED193
IV-NonPhoto
Jan.spCCS1
Aci.cryJF52
Xan.autPy22
IV1Pol.spJS666
Chr.salDSM 3043
Ful.marHTCC2506
EBH72800
Bor.borRB50
IVBur.xenLB400
IV2Pol.spJS666
IV2Ost.tau
IV1Ost.tau
ECU18502
EDJ36865
Hel.mob
EDG90708
ECA82964
ECR75886
ECE36765
IV-DeepYkr
ECU62179
Geo.ste
Bac.cerE33L
IV1Rho.palBisA53
IV1Rho.palBisB18
EDE26668
EDJ25372
ECV95656
ECW18074
Alk.ehrMLHE12
Hal.halSL12
IV1Rho.palCGA009
IV1Rho.palBisB5
IVRho.rub
Bac.claKSMK16
Exi.sib25515
Bac.sub168
Bac.licATCC 14580
IV-YkrW/MtnW
AMC541
0.5

Fig. 3A

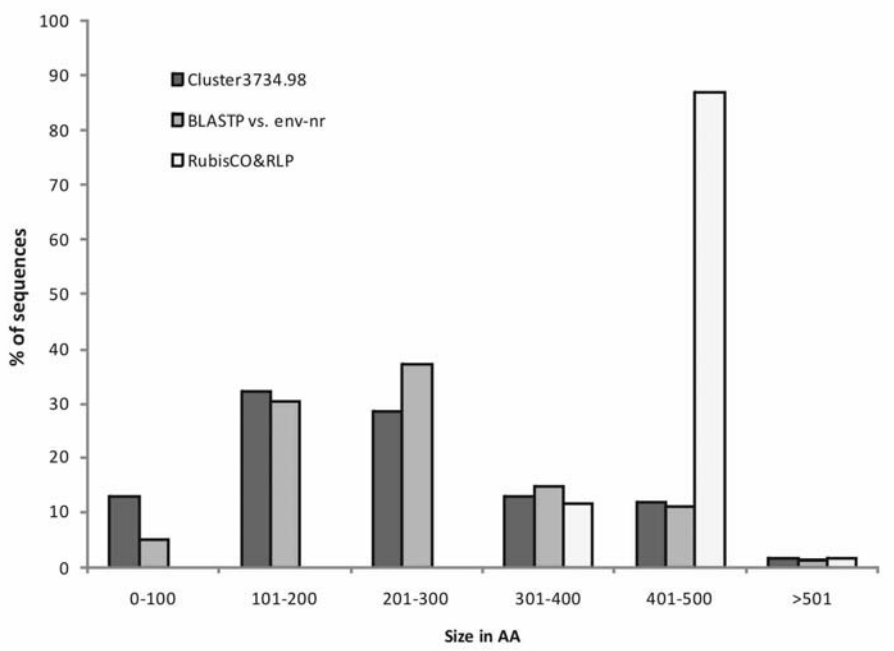

Cluster3734.98
BLASTP vs. env-nr
RubisCO&RLP
% of sequences
0
10
20
30
40
50
60
70
80
90
100
0-100
101-200
201-300
301-400
401-500
>501
Size in AA

Fig. 3B

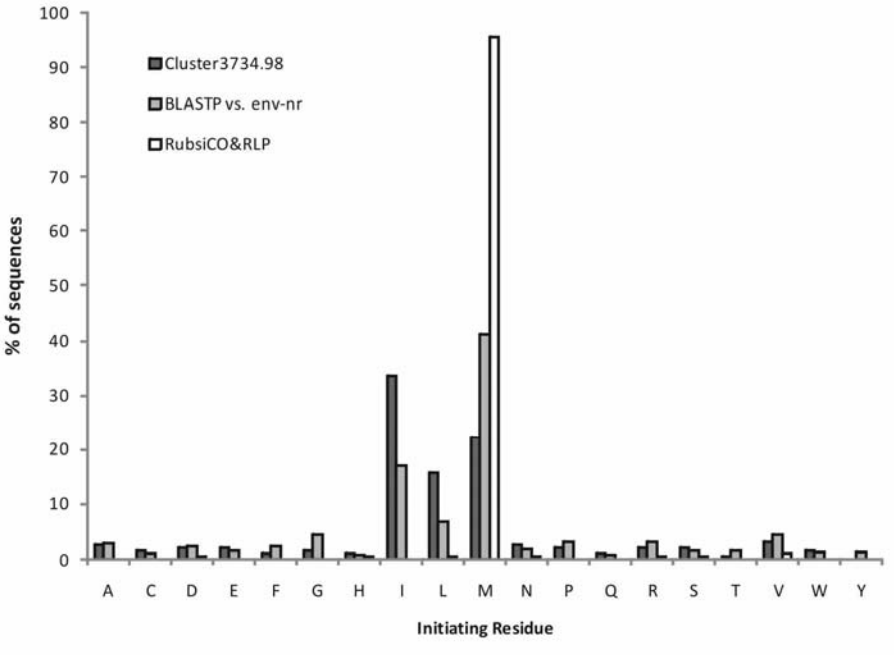

Cluster3734.98
BLASTP vs. env-nr
RubsiCO&RLP
% of sequences
0
10
20
30
40
50
60
70
80
90
100
A
C
D
E
F
G
H
I
L
M
N
P
Q
R
S
T
V
W
Y
Initiating Residue

Fig. 4

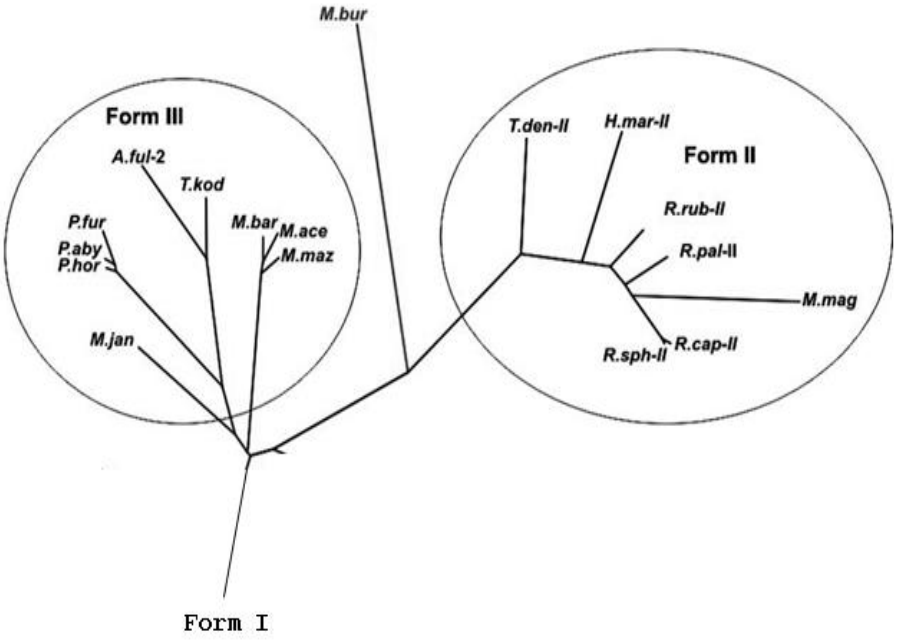

M.bur
Form III
A.ful-2
T.kod
P.fur
P.aby
P.hor
M.bar
M.ace
M.maz
M.jan
T.den-II
H.mar-II
Form II
R.rub-II
R.pal-II
M.mag
R.sph-II
R.cap-II
Form I

Fig. 5

Fig. 6

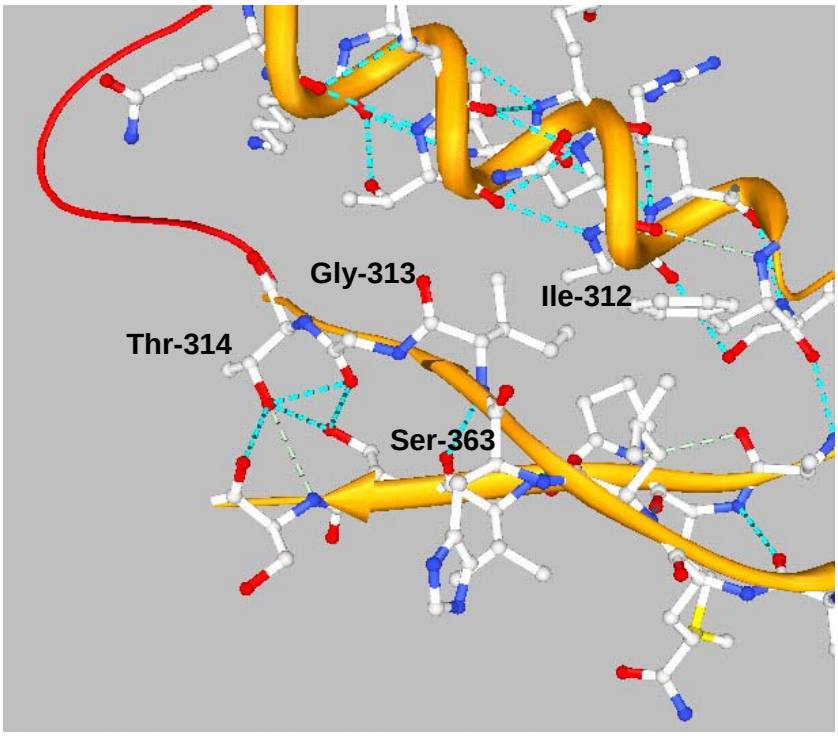


Fig. 7

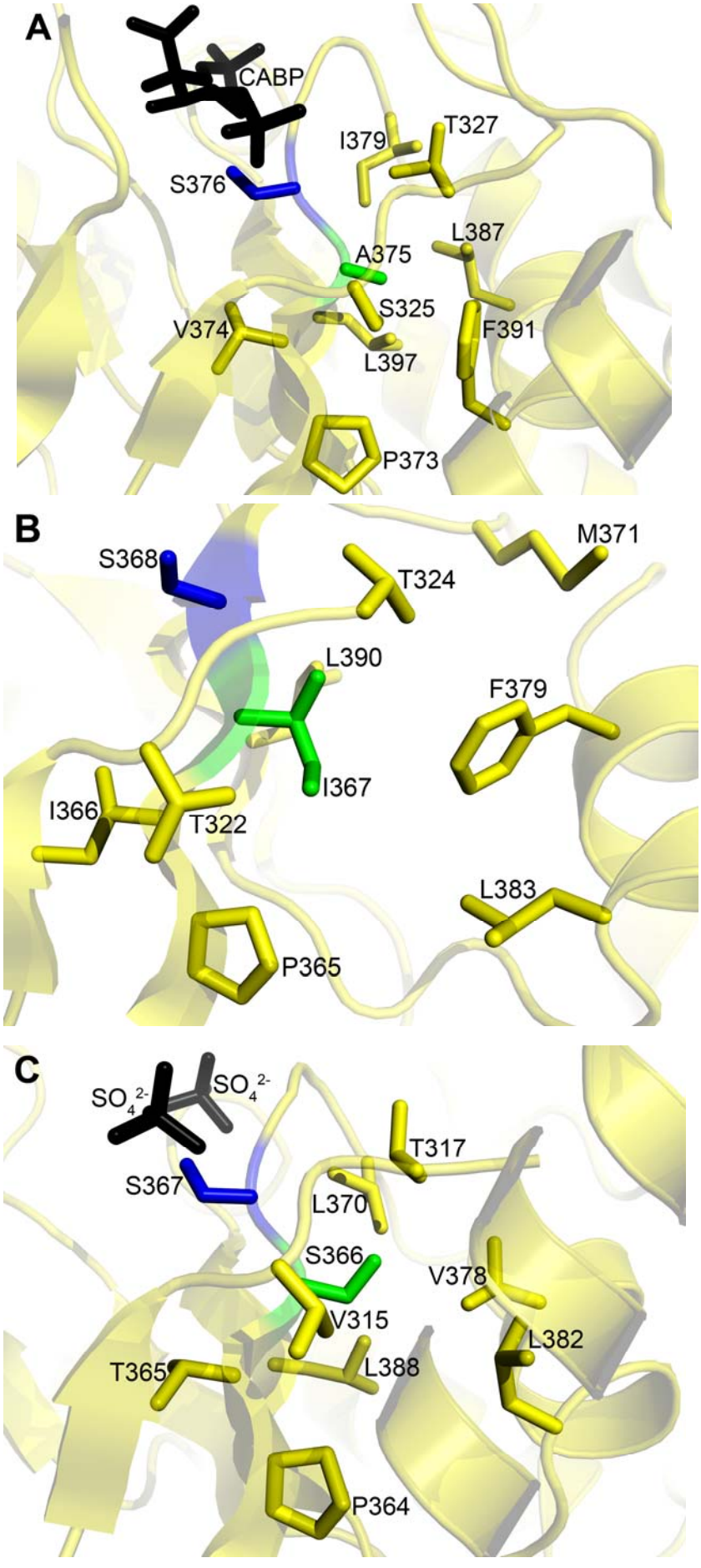
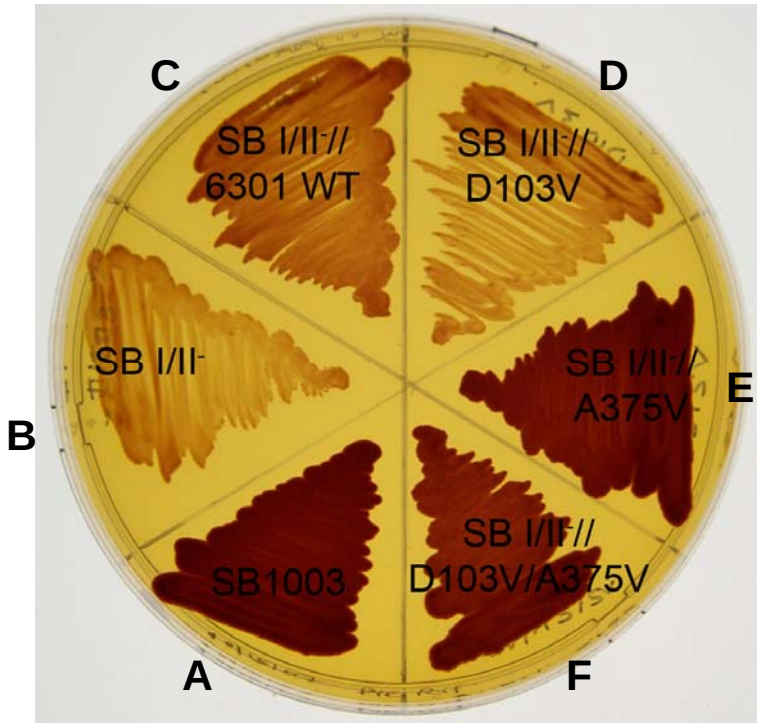

A
CABP
I379
T327
S376
L387
A375
S325
V374
F391
L397
P373
B
S368
M371
T324
L390
F379
I367
I366
T322
L383
P365
C
SO_4^{2-}
SO_4^{2-}
T317
S367
L370
S366
V378
V315
L382
T365
L388
P364

Fig. 8


C
D
SB I/II-//
6301 WT
SB I/II-//
D103V
SB I/II-
B
SB I/II-//
A375V
E
SB I/II-//
D103V/A375V
SB1003
A
F