PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 2629–2640
doi:10.1098/rstb.2008.0023
Published online 16 May 2008

Review

Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms

F. Robert Tabita[1,2,3,*], Thomas E. Hanson[4], Sriram Satagopan[1], Brian H. Witte[1] and Nathan E. Kreel[3]

[1]*Department of Microbiology,* [2]*The Plant Molecular Biology Biology/Biotechnology Program, and* [3]*The OSU Biochemistry Program, The Ohio State University, 484 West 12th Avenue, Columbus, OH 43210-1292, USA*

[4]*College of Marine and Earth Studies, Delaware Biotechnology Institute, University of Delaware, 127 DBI, 15 Innovation Way, Newark, DE 19711, USA*

Ribulose 1,5-bisphosphate (RuBP) carboxylase/oxygenase (RubisCO) catalyses 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, which all catalyse the same reactions. In addition, there exists another form of RubisCO, form IV, which does not catalyse RuBP carboxylation or oxygenation. Form IV is actually a homologue of RubisCO and is called the RubisCO-like protein (RLP). Both RubisCO and RLP appear to have evolved from an ancestor protein in a methanogenic archaeon, and comprehensive analyses indicate that the different 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, has 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.

Keywords: RubisCO; structure/function; evolution; different forms; Calvin–Benson–Bassham cycle

1. INTRODUCTION

Ribulose 1,5-bisphosphate (RuBP) carboxylase/oxygenase (RubisCO) arguably catalyses the most important biochemical reaction in biology and 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 catalyse 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 which the enediolate intermediate of RuBP that acts as a 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 (figure 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.

Owing to 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

*Author and address for correspondence: Department of Microbiology, The Ohio State University, 484 West 12th Avenue, Columbus, OH 43210-1292, USA (tabita.1@osu.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2008.0023 or via http://journals.royalsociety.org.

One contribution of 15 to a Discussion Meeting Issue 'Photosynthetic and atmospheric evolution'.

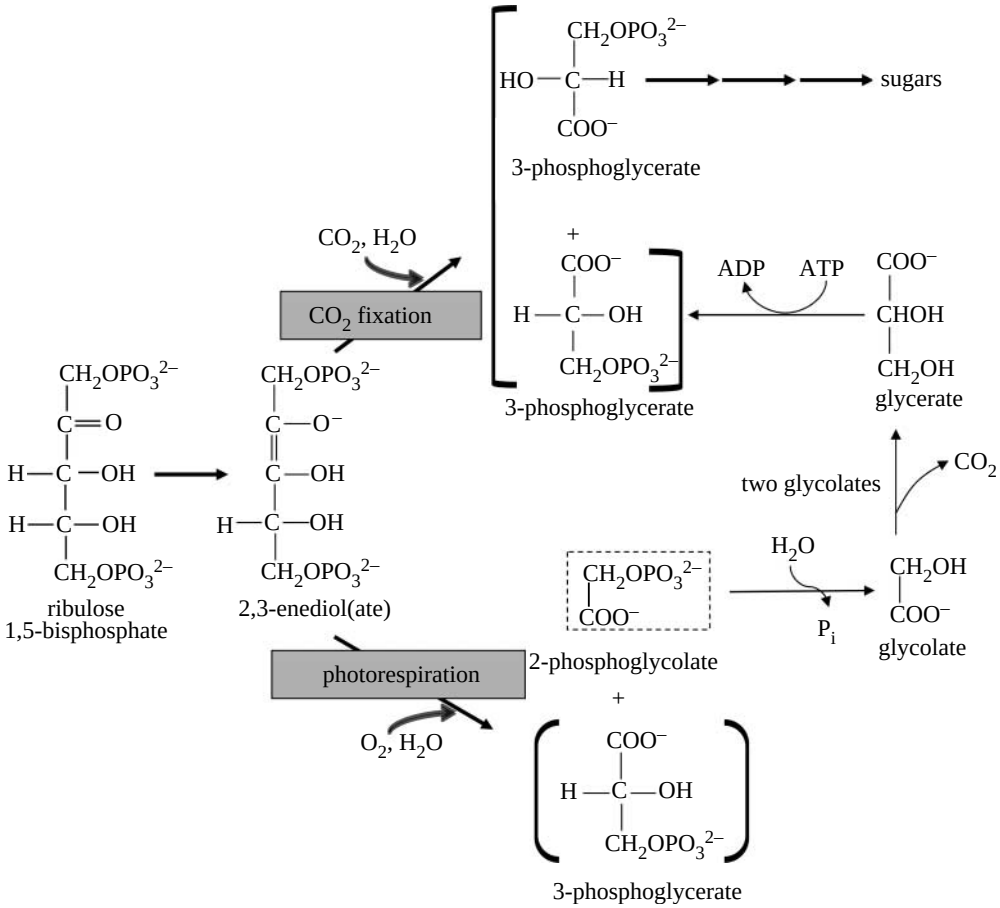

Figure 1. Reactions catalysed by RubisCO showing the inherent problem that aerobic organisms face as a result of the O_2 fixation reaction catalysed by the enzyme. CO_2/O_2 substrate discrimination may be determined experimentally as described in the text.

RubisCO may be released as a result of 2-PG metabolism (figure 1). Clearly, a major key to solving the CO_2 sequestration issue and perhaps improving photosynthetic crop yields might be to somehow enhance RubisCO's ability to discriminate between CO_2 and O_2 (Spreitzer & Salvucci 2002; Long *et al.* 2006*a*,*b*; Tcherkez *et al.* 2006). In this article we review probable 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 molecular adaptations.

2. 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; Tabita *et al.* 2007). In most instances RubisCO, as expected, catalyses the typical carboxylation reaction so necessary for CO_2 reduction via the Calvin–Benson–Bassham (CBB) reductive pentose phosphate pathway (figure 1). However, as the different molecular forms were discovered and analysed, 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 by providing a salvage pathway for intermediates of purine/pyrimidine metabolism (reviewed by Tabita *et al.* 2007). Representative archaea have also 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 analogue has evolved which does not even catalyse RuBP-dependent CO_2 or O_2 fixation.

Table 1. 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

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 comprises large approximately 50 KDa (catalytic) subunits encoded by either the *rbcL* or the *cbbL* genes. In addition, the form I holoenzyme comprises 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 among higher plants, eukaryotic algae, cyanobacteria and proteobacteria. 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 non-green 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 comprises only large subunits that share approximately 30% 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* (Watson *et al.* 1999; Finn & Tabita 2003) 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 catalyse RubisCO activity (Hanson & Tabita 2001). Yet, there are similarities in the primary (Hanson & Tabita 2001) and tertiary (Li *et al.* 2005; Imker *et al.* 2007) structures of these proteins that clearly indicate that RLPs are homologues of RubisCO and are derived from some common ancestor (Tabita *et al.* 2007). RLPs cannot catalyse RubisCO activity because they have non-identical residues at positions where key conserved active-site residues of RubisCO are normally found. The *Bacillus subtilis* (Ashida *et al.* 2003) and *Geobacillus kaustophilus* (Imker *et al.* 2007) RLPs (or YkrW/MtnW), the cyanobacterial RLP from *Microcystis aeruginosa* (Carre-Mlouka *et al.* 2006) and RLPs from the photosynthetic bacteria *Rhodospirillum rubrum* and *Rhodopseudomonas palustris* (J. Singh & F. R. Tabita 2008, unpublished observations) participate in a methionine salvage pathway and catalyse the enolization of the RuBP analogue, 2,3-diketo-5-methylthiopentyl-1-P in a reaction much akin to the enolase reaction catalysed by RubisCO (Ashida *et al.* 2003; Imker *et al.* 2007). Physiological results indicate that RLP from the green sulphur bacterium *Chlorobium tepidum* is involved in some aspect of thiosulphate oxidation (Hanson & Tabita 2001, 2003); however, the precise reaction catalysed 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 (J. Singh & F. R. Tabita 2008, unpublished observations). Only one archaeon, *A. fulgidus*, and one eukaryote, the alga *Ostreococcus tauri*, have thus far been shown to contain an RLP gene (Tabita *et al.* 2007). Each organism contains a functional RubisCO, a form III enzyme in *A. fulgidus* (Watson *et al.* 1999) and a form I enzyme in *O. tauri* with the typical chloroplast localization of *rbcL* and a nuclear *rbcS* gene (Robbens *et al.* 2007).

3. PHYLOGENETIC RELATIONSHIPS AND THE EVOLUTION OF RubisCO AND THE RLP: RE-EVALUATING THE GLOBAL OCEAN SURVEY DATASET'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 available at this time and current

phylogenetic models; thus it is certainly possible that additional sequences and models, as they are discovered and formulated, might cause this hypothesis to be re-examined.

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 from those previously published by ourselves and others (Hanson & Tabita 2001, 2003; Ashida *et al.* 2005; Carre-Mlouka *et al.* 2006; 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.

Two slightly different sequence sets of GOS data were used (refer to §9 below for details) with 31 sequences in common, as shown in table 2 of the electronic supplementary material. The first set, from GOS cluster 3734, was based on sequences used by Yooseph *et al.* (2007) and the second was independently collected using BLASTP to identify RubisCO/RLP homologues in the GOS dataset. When each set was used in concert with a previously vetted set of RubisCO/RLPs 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 (Tabita *et al.* 2007; figures 2 and 3). Indeed, the trees for each set reveal that identical sequences were recovered and placed in extremely similar positions within each tree. The minor discrepancies between the trees are probably derived from unique sequences recovered within each set causing slight differences during the CD-Hit clustering of the independently collected sequence groups. Both trees indicate that there is a single monophyletic group of GOS-derived RubisCO/RLP sequences that do not have any correspondence with currently analysed genome sequences. As noted in our recent review (Tabita *et al.* 2007), the active site of this clade (IV-GOS) does not appear to be functional, thus making it an RLP lineage. The branching position of the IV-GOS clade within our trees is slightly different, though both positions are consistent with the group being an RLP lineage. With data from GOS cluster 3734 (figure 2), the IV-GOS group emerges from between the IV-NonPhoto and IV-YkrW/MtnW clades. However, with GOS sequences retrieved by BLASTP (figure 3), the IV-GOS clade is an early branch of the RLPs relative to the bona fide RubisCO lineage. The bona fide RubisCOs still appear to be a monophyletic clade in either analysis and are likely to have originated from an ancestral methanogen RubisCO as per our favoured model for the evolution of RubisCO/RLP clades (see fig. 9 of Tabita *et al.* 2007).

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

The phylogenies presented here do not agree well with that presented by Yooseph *et al.* (2007). In particular, the distances between major clades within the RubisCO/RLP superfamily appear to be significantly shorter in the Yooseph *et al.* tree with a larger number of intervening branches, which were interpreted to mean that a large number of additional RubisCO/RLP lineages were represented in the GOS dataset (Yooseph *et al.* 2007). In addition, branch lengths between relatively closely related form I subgroups appear to be unusually long in the Yooseph *et al.* RubisCO tree. These incongruencies are probably due to three major differences in the approaches used for phylogenetic reconstruction.

The first difference relates to ORF prediction and length. Our current, as well as previous analyses (Tabita *et al.* 2007), used sequences that constitute a significant fraction of the length of currently known RubisCO/RLP sequences obtained from genomic sequences. In many of these instances, biochemical evidence existed for the start site and protein size. By contrast, the GOS dataset, retrieved by either BLASTP or from cluster 3734, contains sequences as short as 61 amino acids with 35–45% of the dataset having less than 200 amino acids (figure 4*a*). All protein start sites in the GOS sequences are predicted computationally, with no supporting biochemical evidence. The inclusion of short sequences presents a major difficulty in the reconstruction of phylogenetic relationships as they limit the number of informative positions in an alignment and increase the weight of each informative position towards determining both branching order and branch length. This problem is particularly acute if complete deletion of gaps is used in calculating distances from alignments for use in trees. This was not specified in the details of tree construction methods used by these authors (Yooseph *et al.* 2007). In our reconstructions, gaps are deleted in pairwise fashion, not globally, so that a larger number of informative positions were used to calculate pairwise distances.

Given the uncertainties in start site prediction, it is not clear whether or not short sequences within the GOS dataset represent authentic short variants of RubisCO/RLP or fragmentary sequences of longer genes where the remainder of the sequence had not been determined. One indication of fragmentation in the GOS datasets is that the initiation residue for the amino acid sequences was frequently not the canonical methionine (figure 4*b*). While alternative translation initiation sites do occur in prokaryotes (Tech & Meinicke 2006; Makita *et al.* 2007), this is usually a rare occurrence where leucine or valine specified by the codons T/CTG and GTG replace methionine. Over half (59%) of the ORFs in the cluster 3734 dataset and just under half (47%) of GOS sequences retrieved by

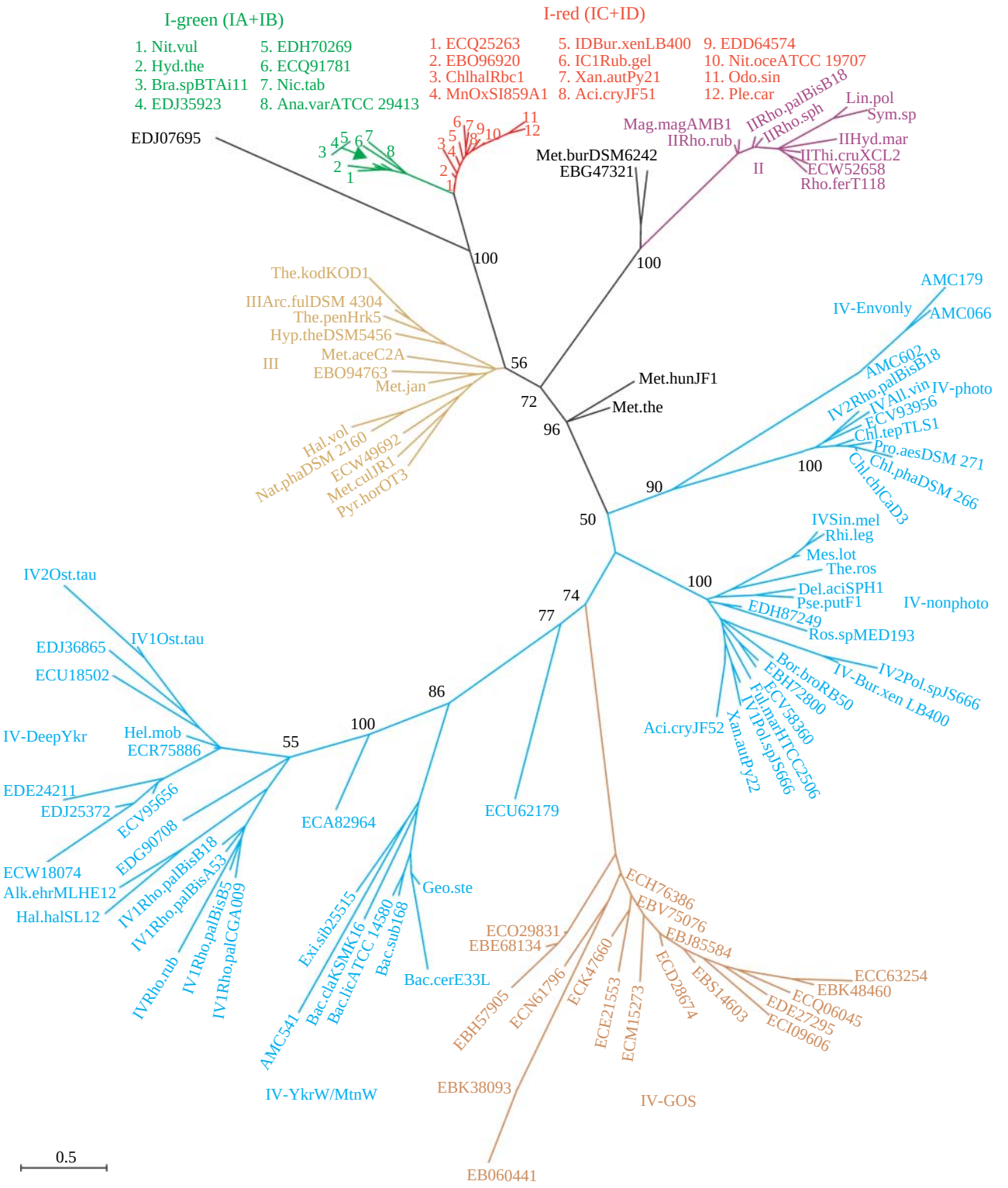

Figure 2. Minimum evolution phylogenetic trees of the RubisCO and RLP superfamilies. Tree built with sequences derived from GOS cluster 3734. Major lineages are labelled and colour coded identically in each panel. Bona fide RubisCO lineages were given individual colours (I, green and red; II, purple; and III, gold), while all RLPs are coloured cyan except for the unique GOS RLP lineage, IV-GOS, which is coloured red-brown. Abbreviations for non-GOS sequences are listed in table 2 of the electronic supplementary material. GOS sequences are identified by their NCBI accession number.

BLASTP did not initiate with M, L or V while over 97% of the RubisCO/RLP sequences in our prior dataset did (figure 4*b*). Therefore, it seems probable 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 that may have driven spurious localization in phylogenetic tree reconstruction.

A large percentage of sequences in the cluster 3734 dataset that initiated with a non-canonical amino acid were found in the BLASTP dataset, which was 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 (figure 4*b*). Within the 29 sequences common between the final sets used to construct the trees, this

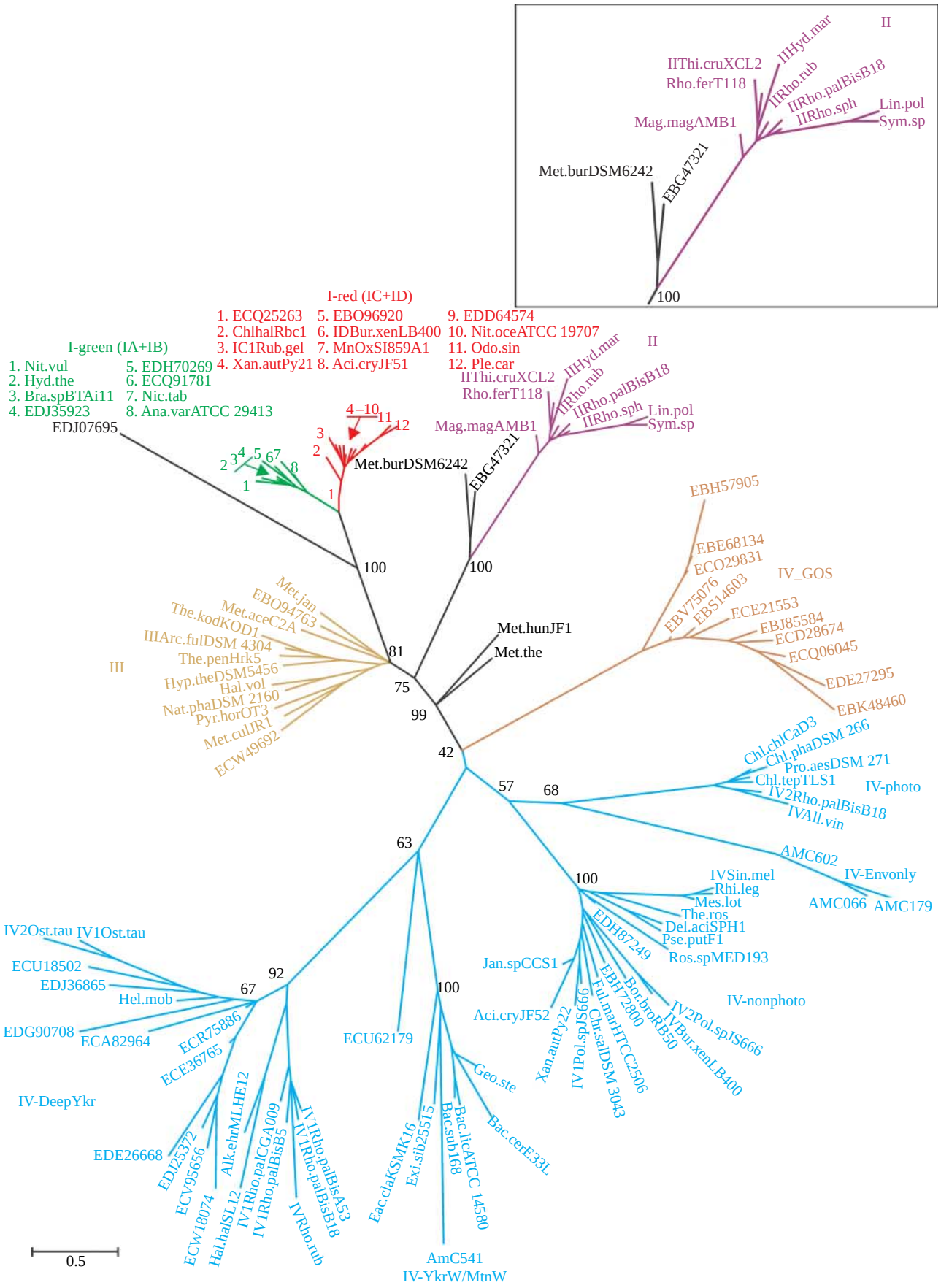

Figure 3. Minimum evolution phylogenetic trees of the RubisCO and RLP superfamilies. Tree built with GOS sequences derived from BLASTP searches of the GenBank env_nr database. Inset: magnification of the form II lineage to highlight the position of the *Methanococcoides burtonii* (Met.burDSM6242) sequence. Colours and abbreviations are the same as indicated in figure 2.

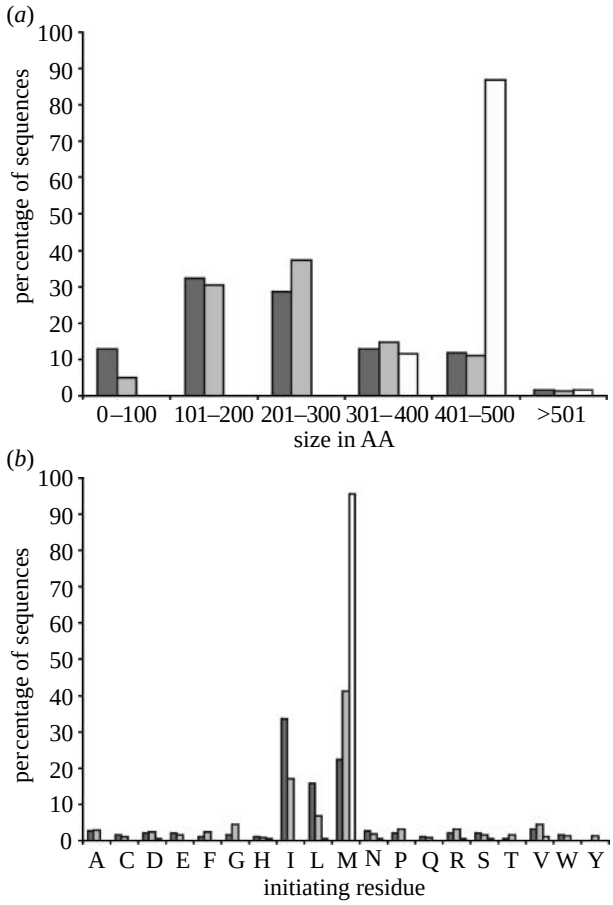

Figure 4. Comparison of starting datasets used to construct phylogenetic trees. (*a*) The percentages of total sequences falling into a given length category are shown for sequences in GOS cluster 3734 (dark grey bars, $n=195$), GOS RubisCO/RLP sequences retrieved by BLASTP versus env_nr from GenBank (light grey bars, $n=377$) and the set previously used for phylogenetic reconstruction (white bars, RubisCO and RLP; $n=191$). (*b*) Distribution of initiating residues for RubisCO and RLP sequences within datasets used to construct phylogenetic trees. The initiating residue is indicated on the horizontal axis. The bars for (*b*) denote the same datasets as in (*a*).

substitution of M in place of I at the initiating residue was observed in 10 sequences, or 34%. It is not clear how this substitution came about, but when sequences were aligned, the initiation site alteration to M was the only difference observed. Thus, users of the GOS data should be sceptical about the identity of an initiating M residue when these sequences are retrieved from GenBank, particularly when these sequences are a fragment of the gene under study.

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 (heterotachy) 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 neighbour-joining phylogeny presented by Yooseph *et al.* nor was the dataset evaluated for appropriate substitution models via a tool like ProtTest (Abascal *et al.* 2005). Model parameters for trees presented here were originally estimated with ProtTest (Abascal *et al.* 2005), and technical details are presented at the end of this report.

The third and final difference has to do with the 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 probably much too inclusive for the identification of clearly defined families or subfamilies. Within all the RubisCO clades, average in-group pairwise amino acid sequence identity range from a low of 39% within the IV-DeepYkr clade to a high of 85% within the form IB subfamily (Tabita *et al.* 2007). This is far below the 100% identity criterion used to cluster the GOS data, probably 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 subgroups. 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.

4. 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 into how all RubisCO molecules function.

5. 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 forms II and III. This sequence is part of a two-member clade with a GOS-derived sequence that originates deeply on the branch leading to recognized form II RubisCO sequences (figure 3, inset). The position of this node is very strongly supported as it is present in 100% of all trees examined during bootstrap analysis in this and prior analyses (Tabita *et al.* 2007). 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 jannaschii* 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 that 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. Second, there is no evidence of other CBB cycle enzymes, especially phosphoribulokinase, the other enzyme unique to the CBB cycle. Finally, homologues of DeoA and E2B2, key enzymes of an AMP-recycling pathway in which form III RubisCO is hypothesized to participate (Sato *et al.* 2007), are present in the genome and are expressed under normal growth conditions (Goodchild *et al.* 2004*a*,*b*, 2005; Saunders *et al.* 2005; Sato *et al.* 2007).

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, Lundqvist & Schneider 1991) or form III (1GEH; Kitano *et al.* 2001) as a template, suggest the presence of a loop (figure 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.

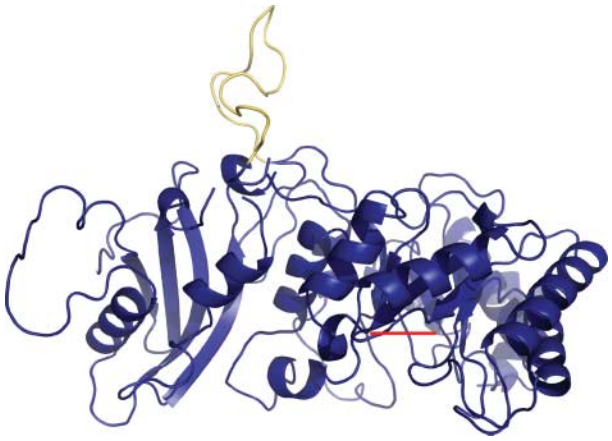

Figure 5. Modelled structure of the *M. burtonii* RubisCO with the loop structure shown in gold. The red bar represents the approximate position of RuBP at the active site.

6. 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 *A. 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 2007), some two orders of magnitude lower than the values obtained for form I or II enzymes (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 threefold 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 & Tabita 2007). The M295D/S363I enzyme has a K_o of approximately 450 μM (N. E. Kreel & F. R. Tabita 2008, unpublished observations), a value that approaches that of form I (from aerobic microbes and plants) 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.

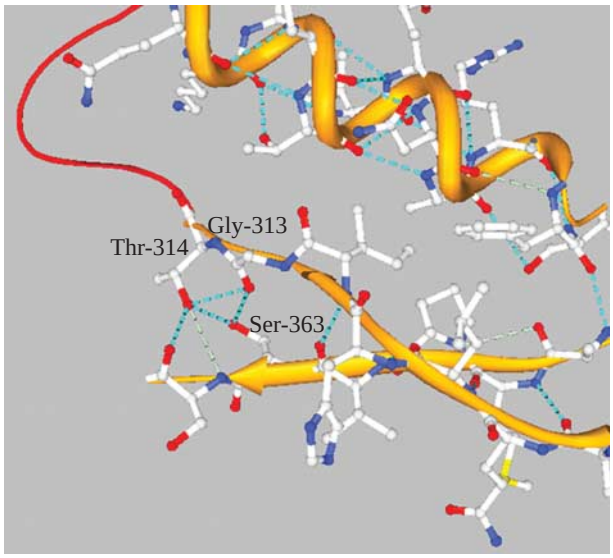

Figure 6. The hydrophobic pocket of the *A. fulgidus* form III RubisCO showing interactions of Ser-363 with conserved residues Gly-313 and Thr-314.

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 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 (figure 6). Gly-313 and Thr-314, found in all forms of RubisCO, show no ionic interactions with the amino acid residues 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 (figure 6), but these substitutions may also cause conformational changes that probably affect the folding of the enzyme either before or during catalysis.

7. APPLICATIONS TO FORM I ENZYMES

The aforementioned studies with the oxygen-sensitive archaeal enzyme clearly implicated residues in hydrophobic regions near the active site to be important for interactions with oxygen. Further examination indicated that such hydrophobic regions are found associated with the active sites of all forms of RubisCO

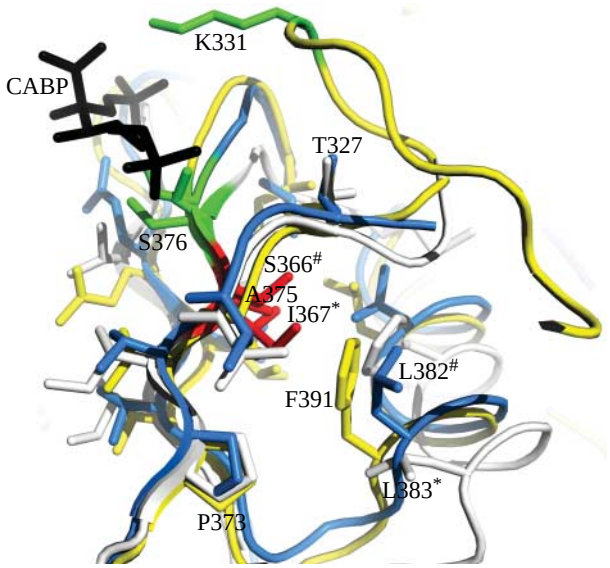

Figure 7. Superimposition of the hydrophobic pocket regions surrounding equivalent residues (red) implicated, or potentially implicated, in oxygen interactions of forms I, II and III RubisCO, i.e. Ala-375 of the form I *Synechococcus* PCC6301 RubisCO (yellow; 1RBL, Newman *et al.* 1993), Ile-367 of the form II *R. rubrum* enzyme (grey; 9RUB, Lundqvist & Schneider 1991) and Ser-366 of the archaeal *T. kodakaraensis* form III enzyme (blue; 1GEH, Kitano *et al.* 2001). Ser-368 of the *T. kodakaraensis* enzyme is equivalent to Ser-363 of the *A. fulgidus* enzyme, previously shown to be involved with oxygen interactions (Kreel & Tabita 2007). Active-site residues, e.g. the adjacent Ser and the specificity-determining loop-6 Lys (present only in the *Synechococcus* RubisCO crystal structure) are both coloured green in all three structures. The transition state analogue, 2-carboxyarabinitol 1,5-bisphosphate (CABP), which is present only in the *Synechococcus* RubisCO crystal structure, is shown in black. Residues shown in each structure are within 4 Å of Ala-375/Ile-367/Ser-366 of the form I/II/III proteins, respectively. Residues of the *R. rubrum* and *T. kodakaraensis* enzymes are indicated by 'asterisks' or 'hashes', respectively, following the residue numbers.

(figure 7). Previous work showed that directed enzyme evolution procedures might be applied to bioselect useful mutant forms of prokaryotic RubisCO after random mutagenesis using a *R. capsulatus* host with its endogenous RubisCO genes deleted (Smith & Tabita 2003, 2004). This system (with *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 with respect to the interactions with oxygen (S. Satagopan, S. S. Scott & F. R. Tabita 2008, unpublished observations). 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

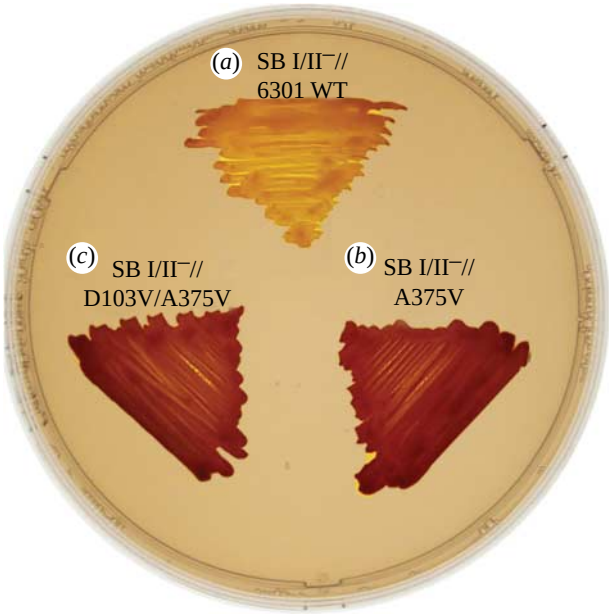

Figure 8. Phenotypic response of mutant form I *Synechococcus* PCC6301 RubisCO constructs in *R. capsulatus* strain SBI/II$^-$. All strains were grown under aerobic conditions in a complex peptone–yeast extract media. Sectors contained (*a*) *R. capsulatus* strain SBI/II$^-$ complemented with the wild-type *Synechococcus* PCC 6301 *rbcLS* genes, (*b*) *R. capsulatus* strain SBI/II$^-$ complemented with the A375V mutant *Synechococcus* PCC 6301 *rbcL* gene and wild-type *rbcS* and (*c*) *R. capsulatus* strain SBI/II$^-$ complemented with the D103V/A375V mutant *Synechococcus* PCC 6301 *rbcL* gene and wild-type *rbcS*.

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

Since mutations of both Ser-363 of the *A. fulgidus* form III (archaeal) enzyme and Ala-375 of the *Synechococcus* form I enzyme affected the K_o, it was surmised that the A375V mutant enzyme might show a phenotypic response, especially under aerobic growth conditions. Thus, the effect of the A375V mutation was determined using the *R. capsulatus* system, i.e. *R. capsulatus* strain SBI/II$^-$ was complemented with the gene containing the A375V change (figure 8*b*). In this experiment it is clear that only enzymes containing the A375V mutation (figure 8*b*,*c*) are able to support luxuriant growth in strain SBI/II$^-$ (equivalent to wild-type *R. capsulatus* strain SB1003, data not shown) in a complex medium under aerobic conditions. Neither the wild-type (figure 8*a*) nor the D103V mutant *Synechococcus rbcL* gene (data not shown) supported luxuriant growth in strain SBI/II$^-$ using complex media. Furthermore, under aerobic chemoautotrophic conditions, with CO_2 as the sole source of carbon, only the A375V mutant supported substantial growth (S. Satagopan, S. S. Scott & F. R. Tabita 2008, unpublished observations). Kinetic analyses indicated that the K_o/K_c ratio, and not any particular relationship of the substrate specificity factor and k_{cat} (Tcherkez *et al.* 2006), was the major factor in allowing the A375V mutant enzymes to support aerobic CO_2-dependent growth (S. Satagopan, S. S. Scott & F. R. Tabita 2008, unpublished observations). These studies indicate that residue Ala-375 influences the ability of RubisCO to interact with oxygen and its effect is manifested in a physiologically significant way.

8. CONCLUSIONS

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

9. NOTE: SEQUENCE SELECTION AND PHYLOGENETIC RECONSTRUCTION

GOS sequences for analysis were collected in two batches, which were treated independently. First, in constructing RubisCO/RLP phylogeny, all sequences within cluster 3734 (Yooseph *et al.* 2007) were retrieved (H. Li 2007, 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 subgroups observed within major RubisCO/RLP lineages (i.e. form IA versus IB). The longest sequence seeding each cluster was retained, and all sequences representing clusters of less than 200 amino acid residues in length (less than 50% of the average 448 ± 40 residue length of well-characterized RubisCO/RLPs) were discarded.

Remaining sequences were aligned by ClustalW (Thompson *et al.* 1994) and phylogenetic trees constructed in MEGA v. 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. These parameters are based on the criteria described by Tabita *et al.* (2007) where four different reconstruction methods were employed after evaluating a much larger aligned dataset with ProtTest for model selection. The tree topologies obtained here are essentially identical to those obtained with the larger dataset. Additional details are provided by Tabita *et al.* (2007).

F.R.T. 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. T.E.H. was supported by NSF CAREER Award MCB-0447649.

REFERENCES

Abascal, F., Zardoya, R. & Posada, D. 2005 ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105. (doi:10.1093/bioinformatics/bti263)

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. (doi:10.1126/science.1086997)

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. (doi:10.1016/j.resmic.2005.01.014)

Carre-Mlouka, A. *et al*. 2006 A new RubisCO-like protein coexists with a photosynthetic RubisCO in the planktonic cyanobacteria *Microcystis*. *J. Biol. Chem.* **281**, 24 462–24 471. (doi:10.1074/jbc.M602973200)

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 CO_2 assimilation, nitrogen fixation, hydrogen metabolism and energy generation. *FEMS Microbiol. Rev.* **28**, 353–376. (doi:10.1016/j.femsre.2004.01.002)

Ezaki, S., Maeda, N., Kishimoto, T., Atomi, H. & Imanaka, T. 1999 Presence of a structurally novel type ribulose-bisphosphate carboxylase/oxygenase in the hyperthermophilic archaeo, *Pyrococcus kodakaraensis* KOD1. *J. Biol. Chem.* **274**, 5078–5082. (doi:10.1074/jbc.274.8.5078)

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. (doi:10.1128/JB.185.10.3049-3059.2003)

Finn, M. W. & Tabita, F. R. 2004 Modified pathway to synthesize ribulose 1,5-bisphosphate in methanogenic archaea. *J. Bacteriol.* **186**, 6360–6366. (doi:10.1128/JB.186.19.6360-6366.2004)

Franzmann, P. D., Springer, N., Ludwig, W., Demacario, E. C. & Rohde, M. 1992 A methanogenic archaeon from Ace Lake, Antarctica—*Methanococcoides-Burtonii* Sp-Nov. *Syst. Appl. Microbiol.* **15**, 573–581.

Goodchild, A., Raftery, M., Saunders, N. F. W., Guilhaus, M. & Cavicchioli, R. 2004*a* Biology of the cold adapted archaeon, *Methanococcoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **3**, 1164–1176. (doi:10.1021/pr0498988)

Goodchild, A., Saunders, N. F. W., Ertan, H., Raftery, M., Guilhaus, M., Curmi, P. M. G. & Cavicchioli, R. 2004*b* A proteomic determination of cold adaptation in the Antarctic archaeon, *Methanococcoides burtonii*. *Mol. Microbiol.* **53**, 309–321. (doi:10.1111/j.1365-2958.2004.04130.x)

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. (doi:10.1021/pr049760p)

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. USA* **98**, 4397–4402. (doi:10.1073/pnas.081610398)

Hanson, T. E. & Tabita, F. R. 2003 Insights into the stress response and sulfur metabolism revealed by proteome analysis of a *Chlorobium tepidum* mutant lacking the RubisCO-like protein. *Photosynth. Res.* **78**, 231–248. (doi:10.1023/B:PRES.0000006829.41444.3d)

Imker, H. J., Fedorov, A. A., Fedorov, E. V., Almo, S. C. & Gerlt, J. A. 2007 Mechanistic diversity in the RubisCO superfamily: the 'enolase' in the methionine salvage pathway in Geobacillus kaustophilus. *Biochemistry* **46**, 4077–4089. (doi:10.1021/bi7000483)

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**, 473–481. (doi:10.1016/S0969-2126(01)00608-6)

Kreel, N. E. & Tabita, F. R. 2007 Substitutions at methionine 295 of *Archaeglobus fulgidus* ribulose-1,5-bisphosphate carboxylase/oxygenase affect oxygen binding and CO_2/O_2 specificity. *J. Biol. Chem.* **282**, 1341–1351. (doi:10.1074/jbc.M609399200)

Li, W. & Godzik, A. 2006 CD-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659. (doi:10.1093/bioinformatics/btl158)

Li, H., Sawaya, M. R., Tabita, F. R. & Eisenberg, D. 2005 Crystal structure of a RubisCO-like protein from the green sulfur bacterium *Chlorobium tepidum*. *Structure* **13**, 779–789. (doi:10.1016/j.str.2005.02.017)

Long, S. P., Ainsworth, E. A., Leakey, A. D. B., Nosberger, J. & Ort, D. R. 2006*a* Food for thought: lower-than-expected crop yield stimulation with rising CO_2 concentrations. *Science* **312**, 1918–1921. (doi:10.1126/science.1114722)

Long, S. P., Zhu, X.-G., Naidu, S. L. & Ort, D. R. 2006*b* Can improvement in photosynthesis increase crop yields? *Plant Cell Environ.* **29**, 315–330. (doi:10.1111/j.1365-3040.2005.01493.x)

Lundqvist, T. & Schneider, G. 1991 Crystal structure of activated ribulose-1,5-bisphosphate carboxylase complexed with its substrate, ribulose-1,5-bisphosphate. *J. Biol. Chem.* **266**, 12 604–12 611.

Makita, Y., de Hoon, M. J. & Danchin, A. 2007 Hon-yaku: a biology-driven Bayesian methodology for identifying translation initiation sites in prokaryotes. *BMC Bioinform.* **8**, 47. (doi:10.1186/1471-2105-8-47)

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. Sect. D* **49**, 548–560. (doi:10.1107/S090744499300530X)

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & Van de Peer, Y. 2007 The complete chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. *Mol. Biol. Evol.* **24**, 956–968. (doi:10.1093/molbev/msm012)

Sato, T., Atomi, H. & Imanaka, T. 2007 Archaeal type III RubisCOs function in a pathway for AMP metabolism. *Science* **315**, 1003–1006. (doi:10.1126/science.1135999)

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. (doi:10.1021/pr049797+)

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. (doi:10.1016/S0022-2836(03)00786-1)

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**, 25 632–25 637. (doi:10.1074/jbc.M401360200)

Spreitzer, R. J. & Salvucci, M. E. 2002 RubisCO: structure, regulatory interactions, and possibilities for a better enzyme. *Annu. Rev. Plant Biol.* **53**, 449–475. (doi:10.1146/annurev.arplant.53.100301.135233)

Tabita, F. R. 1999 Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: a different perspective. *Photosynth. Res.* **60**, 1–28. (doi:10.1023/A:1006211417981)

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**, 576–599. (doi:10.1128/MMBR.00015-07)

Tamura, K., Dudley, J., Nei, M. & Kumar, S. 2007 MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599. (doi:10.1093/molbev/msm092)

Tcherkez, G. G. B., Farquhar, G. & Andrews, T. J. 2006 Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases my be perfectly optimized. *Proc. Natl Acad. Sci. USA* **103**, 7246–7251. (doi:10.1073/pnas.0600605103)

Tech, M. & Meinicke, P. 2006 An unsupervised classification scheme for improving predictions of prokaryotic TIS. *BMC Bioinform.* **7**, 121. (doi:10.1186/1471-2105-7-121)

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**, 4673–4680. (doi:10.1093/nar/22.22.4673)

Tyson, G. W. *et al.* 2004 Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43. (doi:10.1038/nature02340)

Watson, G. M., Yu, J. P. & Tabita, F. R. 1999 Unusual ribulose 1,5-bisphosphate carboxylase/oxygenase of anoxic archaea. *J. Bacteriol.* **181**, 1569–1575.

Yooseph, S. *et al.* 2007 The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5**, e16. (doi:10.1371/journal.pbio.0050016)

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**, 64–69.