Biophysical analysis of the structural evolution of substrate specificity in RuBisCO

Saroj Poudel[a,1], Douglas H. Pike[b,1], Hagai Raanan[a,2], Joshua A. Mancini[a], Vikas Nanda[b], Rosalind E. M. Rickaby[c], and Paul G. Falkowski[a,3]

[a]Environmental Biophysics and Molecular Ecology Program, Department of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901; [b]Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854; and [c]Department of Earth Sciences, University of Oxford, OX1 3AN Oxford, United Kingdom

Contributed by Paul G. Falkowski, October 16, 2020 (sent for review September 8, 2020; reviewed by George H. Lorimer and John Raven)

Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is the most abundant enzyme on Earth. However, its catalytic rate per molecule of protein is extremely slow and the binding of the primary substrate, CO_2, is competitively displaced by O_2. Hence, carbon fixation by RuBisCO is highly inefficient; indeed, in higher C3 plants, about 30% of the time the enzyme mistakes CO_2 for O_2. Using genomic and structural analysis, we identify regions around the catalytic site that play key roles in discriminating between CO_2 and O_2. Our analysis identified positively charged cavities directly around the active site, which are expanded as the enzyme evolved with higher substrate specificity. The residues that extend these cavities have recently been under selective pressure, indicating that larger charged pockets are a feature of modern RuBisCOs, enabling greater specificity for CO_2. This paper identifies a key structural feature that enabled the enzyme to evolve improved CO_2 sequestration in an oxygen-rich atmosphere and may guide the engineering of more efficient RuBisCOs.

RuBisCO | protein engineering | binding selectivity | protein structural evolution

With an estimated mass of approximately 0.7×10^{15} g, ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO; EC 4.1.1.39) is almost certainly the most abundant enzyme on Earth (1). The enzyme catalyzes the addition of carbon dioxide (CO_2) from the environment to ribulose 1,5-bisphosphate (RuBP) to form two molecules of 3-phosphoglycerate (i.e., carboxylation), each of which subsequently are reduced to an aldehyde in the Calvin–Benson–Bassham cycle (2). In the contemporary world, the enzyme is, by far, the most important carbon fixing enzyme and, hence, the source of organic carbon for all animals and many microorganisms (3) on this planet. However, RuBisCO is an ancient enzyme; phylogenetic analyses suggest it evolved from a noncarbon fixing ancestral enzyme (4, 5) and is found in all domains of life (4, 6).

Unlike many other carboxylating enzymes, RuBisCO only reacts with CO_2 rather than bicarbonate (HCO_3^-). Curiously, the carboxylation reaction of RuBisCO is competitively inhibited by molecular oxygen (7). The two gases have very different structures. Nevertheless, the competitive binding between O_2 and CO_2 to RuBisCO severely hinders the efficiency of the enzyme in the contemporary atmosphere where CO_2 concentrations are relatively low and O_2 concentrations are relatively high (8, 9). How the enzyme "mistakes" O_2 for CO_2 is poorly understood.

The discrimination between O_2 and CO_2 can be quantitatively described by the specificity ratio, S, which is defined as:

$$S = V_c K_o / V_o K_c,$$

where V_c and V_o are maximal velocities for carboxylation and oxygenation, respectively, and K_c and K_o are the relative Michaelis constants for CO_2 and O_2, respectively (2, 9). Due to the poor selectivity of RuBisCO and low turnover rate, the enzyme's performance limits global carbon fixation (10, 11). Consequently, RuBisCO has become a major target for artificial engineering to enhance both catalytic rate and substrate specificity in higher plants to boost crop and water use efficiency (12).

To date, four distinct groups of the enzyme have been identified (6, 13). All have a homodimeric functional form composed of two large (L) subunits (~50–55 kDa) and some have attached small (S) subunits (~10–12 kDa) (14). Depending on the type of RuBisCO, the holoenzyme could be anywhere from a simple homodimer (e.g., group II [L2] and group III [L2]) to a hexadecamer (the most widely distributed form) (e.g., group I [L8S8] and group III [L8]) and octadecamer (e.g., group III [L10]) (4, 6). The catalytic site of the enzyme resides in the large subunit in an α/β domain which contains Mg^{2+}, which acts to stabilize and polarize the substrates during catalysis.

The first step of catalysis is activation by carbamylating the catalytic lysine in the active site that binds to Mg^{2+} (14). Upon activation, the enzyme accepts its substrate, RuBP, to be enolized to a 2,3-enediolate intermediate, where a second CO_2 eventually binds (15). This six-carbon intermediate is further, sequentially hydrated, cleaved, and protonated to yield two molecules of phosphoglyceric acid (9, 14–16). The 2,3-enediolate is sensitive to the presence of O_2, which presents challenges to

Significance

RuBisCO, the most abundant enzyme on Earth, catalyzes the fixation of CO_2 to form an organic acid. It does not clearly discriminate between CO_2 and O_2. Reaction with the former leads to the productive formation of organic carbon; reaction with the latter leads to a metabolically futile, but energetically costly pathway. To elucidate how the enzyme discriminates between CO_2 and O_2, we used computational approaches to identify regions around the active site that play key roles in differentiating the substrates. Our research reveals that the specificity of the enzyme is strongly correlated with the structure of the binding channel in the active site. This work poses a potential pathway to genetically engineer more efficient RuBisCOs.

Author contributions: S.P., D.H.P., H.R., V.N., R.E.M.R., and P.G.F. designed research; S.P., D.H.P., J.A.M., and R.E.M.R. performed research; S.P. and D.H.P. contributed new reagents/analytic tools; S.P., D.H.P., H.R., J.A.M., V.N., R.E.M.R., and P.G.F. analyzed data; and S.P., D.H.P., H.R., V.N., R.E.M.R., and P.G.F. wrote the paper.

Reviewers: G.H.L., University of Maryland, College Park; and J.R., University of Dundee.

The authors declare no competing interest.

Published under the PNAS license.

[1]S.P. and D.H.P. contributed equally to this work.

[2]Present address: Department of Plant Pathology and Weed Research, Institute of Plant Protection, The Agricultural Research Organization, Gilat Research Center, M.P. Negev 85280, Israel.

[3]To whom correspondence may be addressed. Email: falko@marine.rutgers.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2018939117/-/DCSupplemental.

Downloaded by Paul G. Falkowski on November 16, 2020

the enzyme's bifunctional property to either take the carboxylation or oxygenation route.

What controls this bifunctional property?

The evolutionary history of RuBisCO can potentially reveal the key amino acids that led to increased catalytic selectivity (17). Indeed, the composition of the atmosphere has changed dramatically over the course of Earth's history, in part due to RuBisCO activity (8, 17). RuBisCO evolved under anaerobic conditions (5, 6), in the Archean Eon, prior to the Great Oxidation Event around 2.33 billion years ago (18). From the Archean to the mid-Phanerozoic, atmospheric CO_2 levels dropped by four orders of magnitude, while O_2 levels rose by three orders of magnitude (17). This evolution, from an anaerobic to an aerobic environment, challenged RuBisCO's ability to maintain net carbon fixation rates amid diminishing substrate and increasing inhibitor availability.

Photosynthetic organisms adopted at least three main strategies to improve net photosynthetic carbon fixation in the face of exacerbating O_2/CO_2: 1) improve RuBisCO specificity, 2) increase the intracellular CO_2 concentration through the expression of a carbon concentrating mechanism (CCM), or 3) inhabit an ecological niche which maintains a low O_2/CO_2 (19). These three strategies, which are not mutually exclusive, led to a spectrum of specificity in the extant enzyme. Especially the CCM in algae and its analog of C4 photosynthesis in higher plants allowed the enzyme to be virtually saturated with inorganic carbon, which, in some cases, led to a relaxation in specificity (17, 20, 21).

Here, we examine the correlations between the genetics, amino acid sequences, and the structures of available RuBisCO to better understand what features conferred selectivity between the two substrates. Our results provide a mechanistic interpretation of substrate selectivity and suggest how RuBisCO can be genetically improved.

Results

Phylogenetic Analysis. A maximum-likelihood phylogeny was constructed using 55 RuBisCOs for which specificity ratios were experimentally determined (*SI Appendix*, Table S1). The tree recapitulated the four distinct monophyletic groups G-I–G-IV, consistent with previous analyses (13) (Fig. 1). Of these, G-IV is not known to catalyze O_2/CO_2 and, hence, is considered a paralog enzyme. Therefore, we rooted the tree to G-IV homologs found in the phylum Proteobacteria: *Rhodospirillum rubrum* (ABC22798) and *Rhodopseudomonas palustris* (CAE27610). The root (i.e., the most ancestral form) of the enzyme is found in extant methanogens (e.g., *Methanococcoides burtonii* [WP_011500311]). The remaining G-II enzymes were all derived from bacteria belonging to the phylum Proteobacteria (*SI Appendix*, Table S1).

The overall phylogenetic trajectory supports an increase in specificity for CO_2 over time. Followed by G-II, G-III RuBisCO evolved from the archaeal domain, belonging to the extant organism *Thermococcus kodakarensis* (WP_011251240). G-I RuBisCO evolved late and represents the most modern RuBisCOs found in both Bacteria and most eukaryotic algae and higher plants (*SI Appendix*, Table S1). We also identified four subclasses of G-I, which formed distinct clades. While all of the homologs of G-II and G-III that are closer to the ancestral RuBisCO had low specificity for CO_2, subgroups of recently evolved G-I homologs exhibited a wider range of CO_2 specificity from <50 to >90 (Fig. 1 and *SI Appendix*, Fig. S1).

How did the specificity for CO_2 evolve? With 11 high-resolution molecular structures of RuBisCOs spanning groups G-I to G-III available in the Protein Data Bank (PDB) at the time of this analysis, it was possible to correlate changes in structure with evolutionary adaptations to an increasingly oxidizing environment. The sequence and structure of RuBisCO active sites are highly conserved, providing little insight into the evolutionary pathway of CO_2 specificity. However, residues proximal to enzyme active sites are not as strictly constrained by function and are likely sites for evolution (22). Analysis of surface-facing residues adjacent to the RuBisCO active site provided unique insight into the evolutionary pathway of CO_2 specificity.

It was previously proposed that the local electrostatic field formed by positively charged cavities adjacent to the *G. sulphuraria* RuBisCO active site interact strongly with CO_2, resulting in strong substrate selectivity (14). Solvent accessible surface analysis of all available RuBisCO atomic-resolution structures revealed a significantly higher surface charge potential proximal to the active site of the enzyme (*SI Appendix*, Fig. S2), consistent with this electrostatic mechanism. These positively charged surfaces also form solvent accessible channels of various sizes that connect to the active site (*SI Appendix*, Fig. S3). Those with the lowest specificity for CO_2 have small, positively charged cavities that barely extend past the catalytic site (Fig. 2), while those with the highest specificity have large, cationic channels for solvent and gas along the functionally active dimer interface connected to both active sites. Moderately specific RuBisCOs have intermediate-sized and mostly disconnected active site cavities across the dimer surface. The correlation between the surface area of the positively charged residues in the 11 PDB structures and the substrate specificity is linear (r^2 = 0.59, $P < 0.05$) and explains ~60% of the variance in specificity (Fig. 3). It is important to note that of the 11 structures, 6 were in the active conformation, and the remaining 5 in an inactive form (*SI Appendix*, Table S1). Hence, the correlation of surface charge and specificity is not due to the conformational state of the structure, but rather a structural feature that is coupled to specificity (*SI Appendix*, Fig. S4). These positively charged cavities likely enable greater CO_2 sequestration to the protein surface proximal to the active site via electrostatic steering with the CO_2 quadrupole coming from a partial negative charge on the two oxygen atoms (*SI Appendix*, Fig. S5). This suggests that selection for RuBisCO specificity resulted in expanding these cavities in favor of the polar carbon-oxygen bond of CO_2 compared to the nonpolar oxygen-oxygen bond in the O_2 molecule.

If selective pressure for expansion of these cavities enabled higher specificity, then we would expect their electrostatic nature to be conserved among homologs. Although our analysis is limited to only 11 structures, we can infer the biochemical mechanism of selectivity for the 55 RuBisCOs using sequence homology within each group. The residues that form these positively charged cavities were identified in homologs within the distinct phylogenetic groups using positions obtained from representative structures for each group to investigate their sequence diversity (*SI Appendix*, Table S2). G-III was excluded because it has only one sequence. G-IA was also excluded because no high-resolution structures are available for this group. However, excluding these groups should not significantly affect our results because both groups have RuBisCOs with low specificity represented by the G-II homologs. For simplicity, we focused on two groups: G-II, which represents RuBisCOs with low specificity, and G-ID, which represents modern RuBisCOs with high specificity.

Sequence diversity of positions in the positive cavity from the representative sequence in G-II (PDB ID code: 5MAC) revealed that they are highly conserved in G-ID homologs (Fig. 4). In fact, they were conserved across all of the additional groups: G-IB and G-IC (*SI Appendix*, Fig. S6). This is expected because G-II homologs are the direct descendent of ancestral RuBisCOs (Fig. 1), and their cavities are most proximal to the highly

Downloaded by Paul G. Falkowski on November 16, 2020

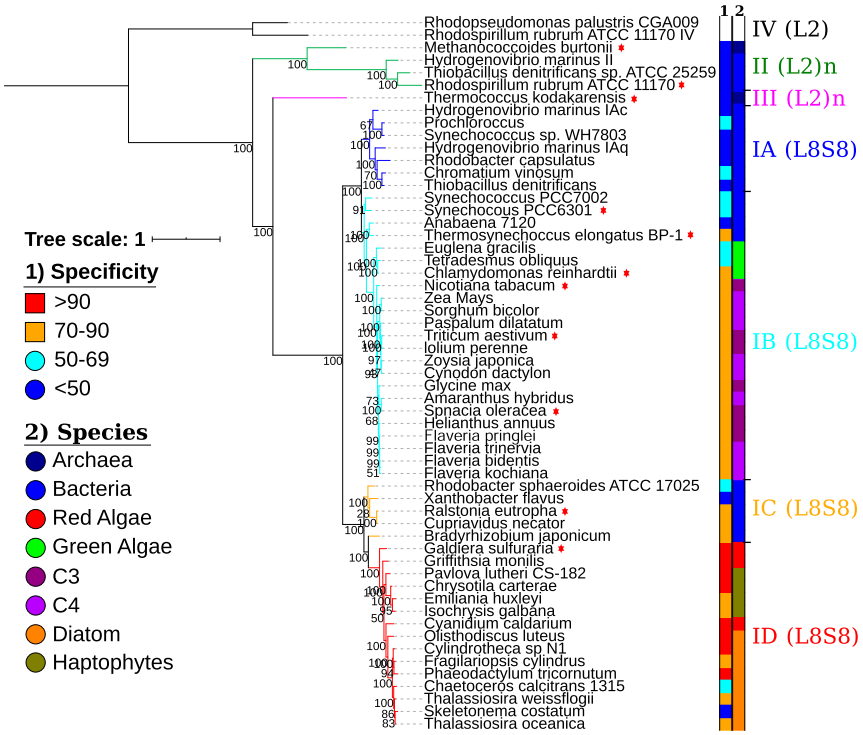

Fig. 1. Maximum-likelihood phylogenetic tree of 55 RuBisCO homologs. The reconstructed tree is rooted to the RuBisCO paralog (i.e., RuBisCO-like protein) from *R. palustris* (CAE27610) and *R. rubrum* (ABC22798). In the color bar on the right, the specificity ratio for CO_2 compared to O_2 (*SI Appendix*, Table S1) in column 1, and the type of organism in column 2, as shown in the legend. The branches are colored based on the groups previously classified (*SI Appendix*, Table S1). The "*" next to the organisms indicates availability of atomic-resolution structures of the specific RuBisCO. The most common subunit composition (i.e., large [L] and small [S]) of each group of RuBisCO is also shown. Only group II and group III RuBisCO have varied numbers of L2 subunit compositions, so we generalized the annotation with "n" that stands for number of subunits. Bootstrap values for each node are also indicated.

conserved active site (Fig. 2). Conversely, recently evolved RuBisCO belonging to G-ID contain cavities that are primarily selected in homologs of its group or groups with similar specificity (e.g., G-IB or G-IC). The fact that these cavities are absent in G-II indicates that there was no selective pressure on those residues that make the cavities in G-ID homologs. In particular, G-ID (using the reference structure 4F0K) contains highly conserved positions that make up its cavities that are not conserved in G-II (Fig. 4). Although not as diverse as G-II homologs, we see significant diversity in both G-IB and G-IC homologs, indicating that specificity-related evolution of extended charged cavities occurred through time (*SI Appendix*, Fig. S6).

Discussion

Our results reveal that RuBisCO, an enzyme that evolved early in the Archean Eon, which was an anaerobic environment (23), has undergone limited evolution in the active site for carbon fixation in a world where oxygen became the second most abundant gas on this planet. Regardless, as oxygen rose in the late Paleoproterozoic, the selectivity of the enzyme for CO_2 increased (Fig. 1 and ref. 4). Our analysis strongly suggests that the selectivity was driven by incremental changes in the structure near the active sites of the dimeric functional large subunit.

RuBisCO is hypothesized to have evolved from a noncarbon fixing ancestral enzyme long before the emergence of the Calvin–Benson–Bassham cycle (5). Several studies suggest G-IV or a RuBisCO-like protein (RLP) to be its ancestral enzyme (4, 5). RLP variants in some organisms have been shown to be involved in the methionine salvage pathway, oxidative metabolism of thiosulfate in green sulfur bacteria, and, in many organisms, its function still remains unknown (4). Therefore, the true ancestor of RuBisCO remains enigmatic. Regardless, all of the above studies are in agreement with our finding that RuBisCO emerged in a methanogen which are obligate anaerobes (13). All early RuBisCOs have low specificity for CO_2 over O_2 compared to more recently evolved enzymes, but in some instances, the specificity has relaxed in recent evolution possibly due to carbon concentrating mechanisms (Fig. 1, e.g., in the G-ID forms).

The minimal, functionally active oligomer of RuBisCO is dimeric (14). In most cases, the active site is buried by the N terminus of the counter subunit. It undergoes structural changes while first activated by CO_2 and then catalyzes the reaction with subsequent CO_2 molecules (14, 15, 24–26). Allosteric regulation between residues in and proximal to the active site have been proposed (27, 28). Coevolution analysis conducted using the program CASP (29) on the G-II representative structure (PDB ID code: 5MAC) identifies suites of residues which are coevolving (*SI Appendix*, Fig. S7*A*). Not surprisingly, most of these residues reside along the positively charged cavities that route to the catalytic site (*SI Appendix*, Fig. S7*B*). It is after activation that susceptibility to O_2 competitive inhibition occurs (14).

Although both O_2 and CO_2 are neutral molecules and lack a dipole moment, the oxygen atoms in CO_2 are more electronegative than the carbon, leading to a significant quadrupole (*SI Appendix*, Fig. S5). Positively charged residues on the surface of the protein can interact with the CO_2 quadrupole, drawing it into the active site via electrostatic steering (Fig. 2 and *SI Appendix*, Fig. S5) (14). In fact, most of the residues in the active site and first shell environment (i.e., <10 Å from Mg) are positively charged (Fig. 2) and highly conserved (*SI Appendix*, Fig. S8). These charged cavities can provide a sequestration path and

Downloaded by Paul G. Falkowski on November 16, 2020

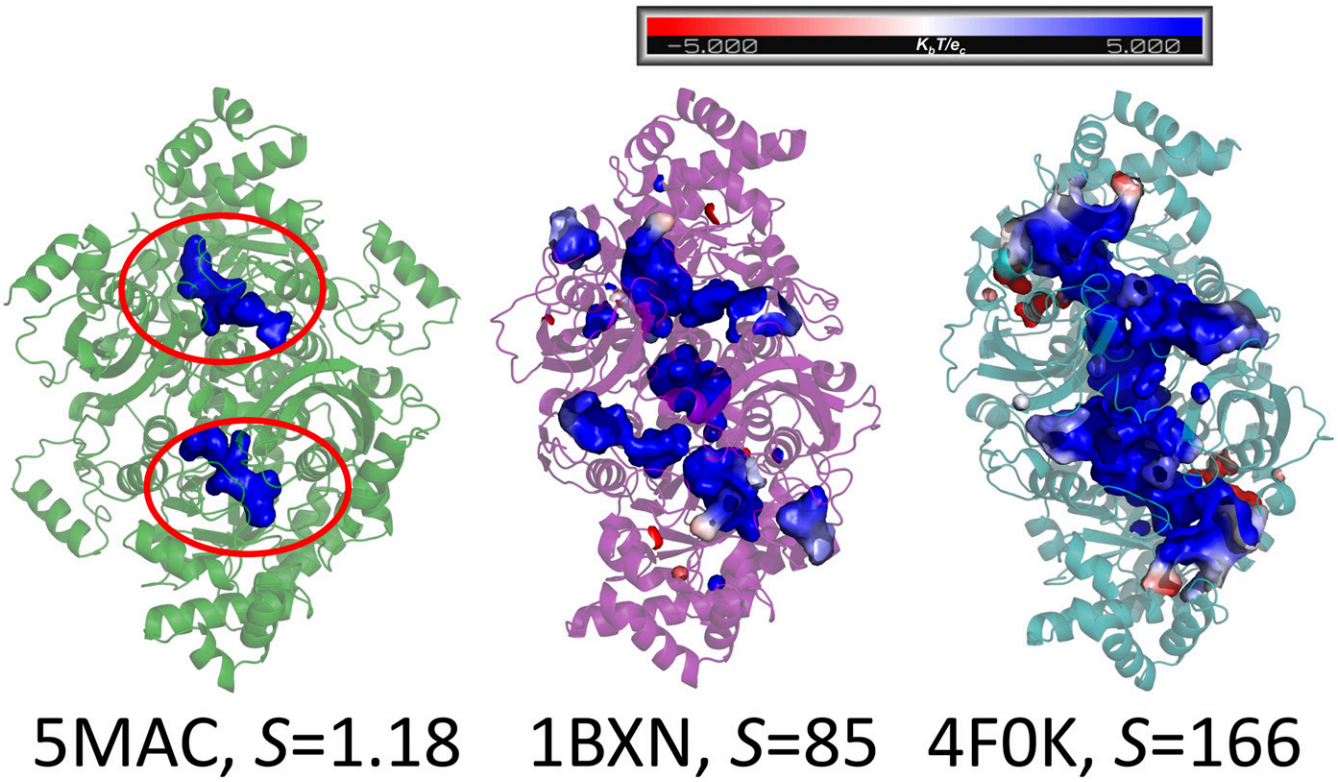

Fig. 2. Graphical surface representation of the positively charged cavities in dimeric RuBisCO. Here, the enzymes with a low, mid, and high specificity (*S*) were selected. The color represents the charge potential (ranging from −5 to +5 K_bT/e_c) based on the results obtained from APBS plugin present in PyMOL. Here, blue represents a positively charged surface potential, red represents a negatively charged surface potential, and white is neutral. The red circle in 5MAC is highlighting the active pocket of RuBisCO. The active site in the remaining structures are also found in the same position as in 5MAC.

potentially act as a reservoir for CO_2, akin to a proximal active site carbon concentrating mechanism. Once the gas is drawn to the surface, the pockets and tunnels that connect the large and small subunits can act as CO_2 reservoirs, stabilizing bound gas through hydrophobic interactions (26). This would suggest that positive cavities and tunnels enhance the channeling of a second and subsequent CO_2 for the carboxylation reaction. However, it may also increase the probability of activation by channeling the first CO_2 to the active site, thereby making both activation and carboxylation processes more efficient.

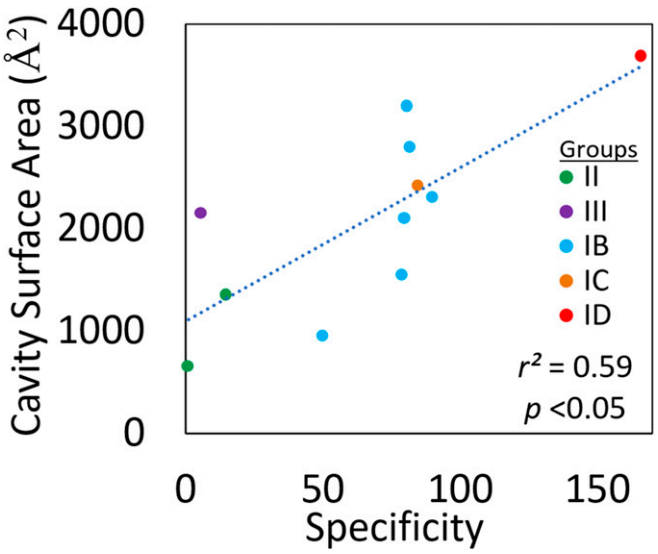

Fig. 3. Scatterplot of positive cavity surface area against the specificity in the known 11 atomic-resolution structure (*SI Appendix*, Table S1). Each point is colored by their groups that were classified in Fig. 1. Here, statistical significance ($r^2 = 0.59$, $P < 0.05$) value is also provided.

The correlation between specificity and accessible, charged cavity surface area supports a mechanism by which expansion of cavities enabled the more recently evolved RuBisCO to preferentially select CO_2 over O_2 (Fig. 3). The fact that each group forms a distinct cluster in the scatterplot further suggests that these positive cavities are a feature of the enzymes with high specificity. Specifically, the structure (PDB ID code: 4F0K) with the highest specificity (i.e., 166) has charged cavities, which form a continuous solvent accessible channel across the dimer surface which leads to each active site (Fig. 2). This may enable the enzyme to store and regulate a continuous supply of CO_2 to the active site, thereby preventing inhibition from oxygen. Despite the prevalence of carbon-concentrating mechanisms in many oxygenic photoautotrophs (30), the diffusion rate of CO_2 in aqueous solution is 10,000 times lower than in air and highlights the additional need for local sequestration of the CO_2 supply (31). The nature of the cavities as a carbon reservoir may also underpin the increasing magnitude of carbon isotopic fractionation with RuBisCO specificity (9) due to an inevitable mass-dependent diffusion of the light isotope into the cavities. In fact, a bigger CO_2 reservoir might limit the number of water molecules creating an almost "dry" site, which would favor CO_2 channeling, minimize Rayleigh fractionation during carboxylation, and a greater isotopic fractionation associated with diffusion of CO_2, similar to that in a gas phase (32).

Downloaded by Paul G. Falkowski on November 16, 2020

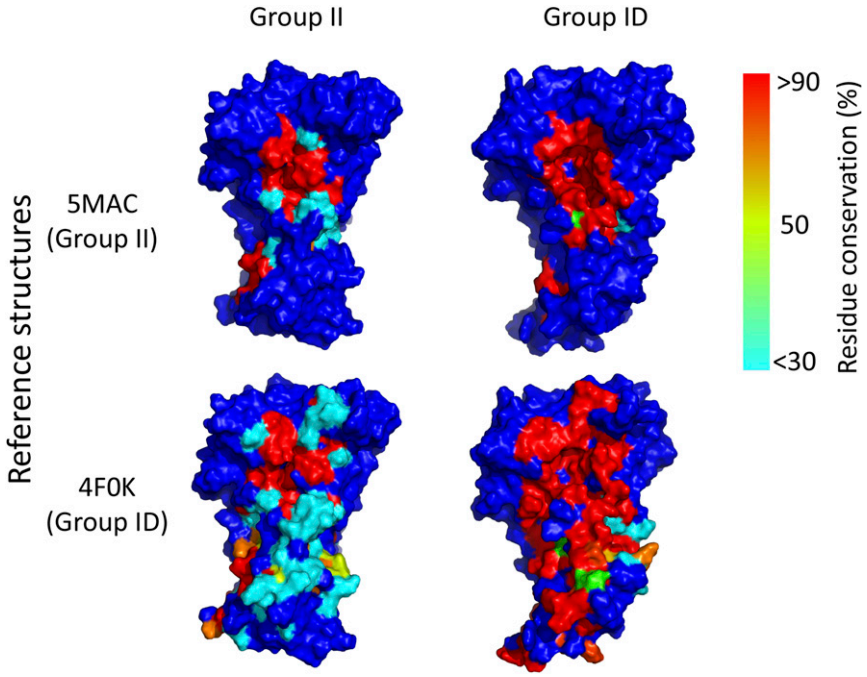

Fig. 4. Conservation of residues that make positive cavities in the monomeric subunit of the dimer. Here, positions from the representative structure (in *y* axis) from both G-II (i.e., 5MAC) and G-ID (i.e., 4F0K) were considered to extract all of the residues in the corresponding positions found in both of the groups (*x* axis). Blue in the structure indicates residues that were not involved in making those positive cavities. Residue conservation (in percent) for each position are highlighted in the structure as shown.

Analysis of sequence diversity shows the extension of cavities is conserved across groups with higher specificity for CO_2 and even in groups with specificity variance (see group 1D in Fig. 1 and ref. 17) and, therefore, is likely a significant component of selection pressure for specificity (Fig. 4 and *SI Appendix*, Fig. S6). G-II homologs are closer to the root and, hence, are the direct descendent of ancestral RuBisCOs. The representative structure of G-II (PDB ID code: 5MAC) is from a methanogen. Methanogens are strictly anaerobic and cannot survive in the presence of oxygen; hence, their RuBisCOs are not under selection pressure to mitigate competitive binding of O_2 over CO_2. They have the smallest positively charged cavities which span the active site region but do not extend beyond it. Our results indicate that all of the homologs independent of group exhibit conservation across this more ancestral cavity (*SI Appendix*, Fig. S6).

Subgroups of G-I evolved from the immediate ancestor of G-III and G-I. Our tree does not include a molecular clock and, therefore, we cannot definitively conclude which group emerged first between IB and ID. Nevertheless, all three groups (i.e., G-IB, IC, and ID) primarily have homologs with high CO_2 specificity and have evolved more positive cavities than early evolving RuBisCO (e.g., G-II). The residues that make up these cavities are evolutionarily selected in all of the homologs of G-ID (i.e., 4F0K, Fig. 4) and across the additional groups as observed in our diversity boxplot (*SI Appendix*, Fig. S6). Interestingly, all G-ID RuBisCOs diverged from that encoded in *Galdieria sulfuraria* (i.e., PDB ID: 4F0K), suggesting all RuBisCOs evolved vertically from the enzyme encoded by *G. sulfuraria*. The diversity of cavity-facing residues is low in RuBisCOs with high specificity and high in RuBisCOs with low specificity.

Our analysis primarily focused on the minimal functional form of RuBisCO, which is dimeric. We cannot ignore the influence of additional subunits found in varied oligomeric forms of RuBisCO that are distinct across all of the groups of the enzyme (6). The simplest form of RuBisCO belongs to G-II, which solely exists as a dimer (i.e., L_2), but the most abundant RuBisCO exists as a hexadecamer (L_8S_8) belonging to G-I (4, 6, 14). In a recent molecular dynamics study (26), the authors speculated that variation in polar contacts, identified across the oligomer, could be one of the contributing factors dictating specificity in the enzyme. We observed a large number of positively charged pockets that extend beyond the catalytic site to the surface near the interface of the subunits. This suggests that the more complex forms of RuBisCO with a large number of subunits (such as L_8S_8) may have even more interconnected cavities bridging the catalytic sites. This could help to explain why G-I RuBisCO on average has higher specificity than other groups.

It is important to note that neither the active nor the inactive conformational state of the enzyme showed differences in the size and the number of positively charged pockets (*SI Appendix*, Fig. S4). This is largely due to the fact that most conformational change between inactive and active states is proximal to the active site center, which is highly conserved across all groups. Positions more distally interacting with the active sites via cavities are less conserved across groups (Fig. 4 and *SI Appendix*, Fig. S8) and, therefore, comprise most of the variance resulting in the correlation between surface area and specificity. Such positions are likely tolerant to mutation, making them accessible to evolution and attractive for engineering RuBisCOs with higher specificity by extending positively charged cavities.

Concluding Remarks

Our genomic and structural analyses of 55 available RuBisCO variants reveal that the ancestral RuBisCO emerged in an anaerobic environment and had a low specificity for CO_2. With the rise of oxygen, the enzyme's selectivity was driven by incremental changes, near the active site, of the large subunit of the dimeric

Downloaded by Paul G. Falkowski on November 16, 2020

structure. Positively charged pockets were identified around the active site that expanded in recently evolved modern RuBisCO with high specificity. We propose this expansion favored the polar carbon-oxygen bond of CO_2 compared to the nonpolar O-O bond in O_2, thereby enhancing specificity of the enzyme for CO_2. Future, experimental work will need to be conducted to test this bioinformatically informed hypothesis. Our analysis suggests it is possible to genetically engineer enhanced specificity for CO_2 in RuBisCO by expansion of positively charged cavities across the surface of the functionally active dimer.

Methods

Generation of Characterized RuBisCO Homologs. The protein sequence of all of the available RuBisCOs that have been experimentally tested for CO_2/O_2 specificity by us in our previous work and other groups in the field were extracted, which accounted to a total of 55 homologs (*SI Appendix*, Table S1). Of the 55 sequences, only 11 had their corresponding structures solved and deposited in the PDB. Six of the structures were solved in an active state and the remaining in an inactive state (*SI Appendix*, Table S2). Since the large subunit of RuBisCO is involved in catalytic activity, the small subunit although involved likely has indirect effect on the activity. Therefore, we only focused on the large subunit for downstream analysis.

Phylogenetic Analysis. Extracted sequences were subjected to multiple sequence alignment using Clustal Omega (33) with default settings. The maximum-likelihood phylogenetic tree was reconstructed using the multiple sequence alignment specifying LG substitution matrix in RAxML (34). Tree was visualized in iTOL (35).

Sequence Conservation. Diversity of sequence was calculated using hamming distance (36), which compares two strings with identical length and gives a dissimilarity score between them. The calculated score was then normalized by the total length of the sequence. This was further normalized by multiplying the total number of mismatches, which we refer to as "Weighted Dissimilarity Score" in the figures.

We also measured sequence diversity present at a specific distance (in angstroms) from Mg that is present in the catalytic site. The distance between each residue and the Mg in the structure were calculated using the script protGetdist (deposited in GitHub: https://github.com/spoudel1/protcad/tree/dev/projects) present in our in-house program protCAD (37).

Solvent-Accessible Surface Area. The Adaptive Possion-Boltzmann Solver (38, 39) (APBS) tool in PyMOL (PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC) with default parameters was used to calculate the charged surface potential using pdb2pqr (38) with the AMBER forcefield (40) to assign protonation states. This solvent-accessible surface area (SASA) is assumed to also be accessible by gas molecules (i.e., CO_2 and O_2). The visual representation of SASA cavities in RuBisCO proteins were generated in PyMOL using a 7 Å cavity detection radius and a 3 Å solvent radius cavity-detection cutoff (default parameters in PyMOL).

Identifying Solvent Accessible Cavities. All solvent accessible cavities were identified using the CASTp program (41). The default probe radius of one solvent radii (i.e., 1.4 Å) was used. For our analysis, we only included the positively charged cavities that directly overlapped with the active site or indirectly overlapped through adjacent cavities within van der Waals contact of one another. Surface area of all of the cavities were also calculated using the CASTp program. The area of all of the included cavities were summed to report the net available surface area of the positively charged cavities.

Statistical Analysis. The significance of the distribution was analyzed using the Welch's *t* test in R (42). Box plot was generated in R.

Data Availability. All study data are included in the article and supporting information.

ACKNOWLEDGMENTS. The work was supported by National Aeronautics and Space Administration (NASA) Astrobiology Grant 80NSSC18M0093 (to P.G.F. and V.N.). R.R. acknowledges support from European Research Council Starting Grant S.P.2-GA-2008–200915. S.P. acknowledges funding from the NASA postdoctoral fellowship.

1. Y. M. Bar-On, R. Milo, The global mass and average rate of rubisco. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 4738–4743 (2019).
2. P. G. Falkowski, J. A. Raven, *Aquatic Photosynthesis* (Princeton University Press, Princeton, NJ, 2007).
3. J. A. Raven, J. Beardall, *Respiration in Aquatic Photolithotrophs* (Oxford University Press, Oxford University Press Inc., New York, 2005).
4. F. R. Tabita *et al*., Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol. Mol. Biol. Rev.* **71**, 576–599 (2007).
5. T. J. Erb, J. Zarzycki, A short history of RubisCO: The rise and fall (?) of nature's predominant CO_2 fixing enzyme. *Curr. Opin. Biotechnol.* **49**, 100–107 (2018).
6. F. R. Tabita, S. Satagopan, T. E. Hanson, N. E. Kreel, S. S. Scott, Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J. Exp. Bot.* **59**, 1515–1524 (2008).
7. T. J. Andrews, G. H. Lorimer, "3–Rubisco: Structure, mechanisms, and prospects for improvement" in *Photosynthesis*, M. D. Hatch, N. K. Boardman, Eds. (Academic Press, 1987), pp. 131–218.
8. R. A. Berner *et al*., Isotope fractionation and atmospheric oxygen: Implications for phanerozoic O(2) evolution. *Science* **287**, 1630–1633 (2000).
9. G. G. B. Tcherkez, G. D. Farquhar, T. J. Andrews, Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7246–7251 (2006).
10. J. N. Young, J. A. L. Goldman, S. A. Kranz, P. D. Tortell, F. M. M. Morel, Slow carboxylation of Rubisco constrains the rate of carbon fixation during Antarctic phytoplankton blooms. *New Phytol.* **205**, 172–181 (2015).
11. A. Sukenik, J. Bennett, P. Falkowski, Light-saturated photosynthesis — Limitation by electron transport or carbon fixation? *Biochim. Biophys. Acta BBA–Bioenerg.* **891**, 205–215 (1987).
12. M. A. J. Parry *et al*., Rubisco activity and regulation as targets for crop improvement. *J. Exp. Bot.* **64**, 717–730 (2013).
13. B. Kacar, V. Hanson-Smith, Z. R. Adam, N. Boekelheide, Constraining the timing of the Great oxidation event within the Rubisco phylogenetic tree. *Geobiology* **15**, 628–640 (2017).
14. B. Stec, Structural mechanism of RuBisCO activation by carbamylation of the active site lysine. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18785–18790 (2012).
15. J. Pierce, G. H. Lorimer, G. S. Reddy, Kinetic mechanism of ribulosebisphosphate carboxylase: Evidence for an ordered, sequential reaction. *Biochemistry* **25**, 1636–1644 (1986).
16. G. Tcherkez, Modelling the reaction mechanism of ribulose-1,5-bisphosphate carboxylase/oxygenase and consequences for kinetic parameters. *Plant Cell Environ.* **36**, 1586–1596 (2013).
17. J. N. Young, R. E. M. Rickaby, M. V. Kapralov, D. A. Filatov, Adaptive signals in algal Rubisco reveal a history of ancient atmospheric carbon dioxide. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 483–492 (2012).
18. G. Luo *et al*., Rapid oxygenation of Earth's atmosphere 2.33 billion years ago. *Sci. Adv.* **2**, e1600134 (2016).
19. J. N. Young *et al*., Large variation in the Rubisco kinetics of diatoms reveals diversity among their carbon-concentrating mechanisms. *J. Exp. Bot.* **67**, 3445–3456 (2016).
20. M. V. Kapralov, D. A. Filatov, Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol. Biol.* **7**, 73 (2007).
21. R. E. M. Rickaby, M. R. E. Hubbard, Upper ocean oxygenation, evolution of RuBisCO and the Phanerozoic succession of phytoplankton. *Free Radical Biology and Medicine* **140**, 295–304, https://doi.org/10.1016/j.freeradbiomed.2019.05.006 (2019).
22. A. J. M. Ribeiro, J. D. Tyzack, N. Borkakoti, G. L. Holliday, J. M. Thornton, A global analysis of function and conservation of catalytic residues in enzymes. *J. Biol. Chem.* **295**, 314–324 (2019).
23. T. W. Lyons, C. T. Reinhard, N. J. Planavsky, The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
24. P. L. Cummins, B. Kannappan, J. E. Gready, Revised mechanism of carboxylation of ribulose-1,5-biphosphate by rubisco from large scale quantum chemical calculations. *J. Comput. Chem.* **39**, 1656–1665 (2018).
25. M. van Lun, D. van der Spoel, I. Andersson, Subunit interface dynamics in hexadecameric rubisco. *J. Mol. Biol.* **411**, 1083–1098 (2011).
26. M. van Lun, J. S. Hub, D. van der Spoel, I. Andersson, CO2 and O2 distribution in Rubisco suggests the small subunit functions as a CO2 reservoir. *J. Am. Chem. Soc.* **136**, 3165–3171 (2014).
27. M. A. J. Parry, A. J. Keys, P. J. Madgwick, A. E. Carmo-Silva, P. J. Andralojc, Rubisco regulation: A role for inhibitors. *J. Exp. Bot.* **59**, 1569–1580 (2008).
28. Y. Marcus, H. Altman-Gueta, A. Finkler, M. Gurevitz, Mutagenesis at two distinct phosphate-binding sites unravels their differential roles in regulation of Rubisco activation and catalysis. *J. Bacteriol.* **187**, 4222–4228 (2005).

Downloaded by Paul G. Falkowski on November 16, 2020

29. M. A. Fares, D. McNally, CAPS: Coevolution analysis using protein sequences. *Bioinformatics* **22**, 2821–2822 (2006).
30. J. A. Raven, C. S. Cockell, C. L. De La Rocha, The evolution of inorganic carbon concentrating mechanisms in photosynthesis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 2641–2650 (2008).
31. J. V. Moroney, R. A. Ynalvez, Proposed carbon dioxide concentrating mechanism in Chlamydomonas reinhardtii. *Eukaryot. Cell* **6**, 1251–1259 (2007).
32. M. H. O'Leary, Carbon isotopes in photosynthesis. *Bioscience* **38**, 328–336 (1988).
33. F. Sievers, D. G. Higgins, Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
34. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
35. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
36. H. P. Pinheiro, A. de Souza Pinheiro, P. K. Sen, Comparison of genomic sequences using the Hamming distance. *J. Stat. Plan. Inference* **130**, 325–339 (2005).
37. D. H. Pike, V. Nanda, Empirical estimation of local dielectric constants: Toward atomistic design of collagen mimetic peptides. *Biopolymers* **104**, 360–370 (2015).
38. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10037–10041 (2001).
39. M.G. Lerner, H. A. Carlson, *APBS plugin for PyMOL - Version 2.4* (University of Michigan, Ann Arbor, MI, 2006).
40. J. W. Ponder, D. A. Case, Force fields for protein simulations. *Adv. Protein Chem.* **66**, 27–85 (2003).
41. W. Tian, C. Chen, X. Lei, J. Zhao, J. Liang, CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.* **46**, W363–W367 (2018).
42. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2014).

Downloaded by Paul G. Falkowski on November 16, 2020