
Article:
Studer, R.A., Christin, P-A., Williams, M.A. et al. (1 more author) (2014) Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. Proceedings of the National Academy of Sciences of the United States of America, 111 (6). pp. 2223-2228. ISSN 1091-6490

https://doi.org/10.1073/pnas.1310811111

Stability-activity trade-offs constrain the adaptive evolution of RubisCO

Romain A. Studer[a], Pascal-Antoine Christin[b], Mark A. Williams[c] and Christine A. Orengo[a]

a. Institute of Structural and Molecular Biology, Division of Biosciences, University College London, Gower Street, London, WC1E 6BT, UK.
b. Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK.
c. Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK.

Corresponding author:
Romain A. Studer
Institute of Structural and Molecular Biology
Division of Biosciences
University College London
Gower Street, London
WC1E 6BT, UK
Tel: +44 (0)20 7679 3890
Email: r.studer@ucl.ac.uk

Classification: Biological Sciences (Evolution).

Running title: Structural constraints on the evolution of RubisCO.

Keywords: RubisCO, evolution, stability, protein structure, C_4 photosynthesis.

Abstract

A well-known case of evolutionary adaptation is that of ribulose-1,5-bisphosphate carboxylase (RubisCO), the enzyme responsible for fixation of CO_2 during photosynthesis. While the majority of plants use the ancestral C_3 photosynthetic pathway, many flowering plants have evolved a derived pathway named C_4 photosynthesis. The latter concentrates CO_2, and C_4 RubisCOs consequently have lower specificity for, and faster turnover of CO_2. The C_4 forms result from convergent evolution in multiple clades, with substitutions at a small number of sites under positive selection.

In order to understand the physical constraints upon these evolutionary changes, we reconstructed *in silico* ancestral sequences and 3D structures of RubisCO from a large group of related C_3 and C_4 species. We were able to track precisely their past evolutionary trajectories, identify mutations on each branch of the phylogeny and evaluate their stability effect. We show that RubisCO evolution has been constrained by stability-activity trade-offs similar in character to those previously identified in laboratory-based experiments. The C_4 properties require a subset of several ancestral destabilising mutations, which from their location in the structure are inferred to mainly be involved in enhancing conformational flexibility of the open-closed transition in the catalytic cycle. These mutations are near, but not in, the active site, or at inter-subunit interfaces. The C_3 to C_4 transition is preceded by a sustained period in which stability of the enzyme is increased, creating capacity to accept the functionally necessary destabilising mutations, and is immediately followed by compensatory mutations which restore global stability.

Introduction

The adaptive diversification of organisms often requires the evolution of novel enzymatic properties. The evolutionary shift from one enzymatic function to another involves crossing an energetic barrier in a fitness landscape (1). The number of mutations that confer advantageous function during such a shift is consequently limited. While some residues are critical for maintaining the stability of the protein fold, others are important for the catalytic activity itself. Due to the multiple roles of amino acids in proteins, the adaptation of one physical parameter of an enzyme is likely to affect other properties (2). As proteins usually form thermodynamically stable structures, their evolutionary trajectories are constrained to a narrow range of stability (3). In particular, the stability and activity are likely to be negatively correlated. Most possible amino-acid changes in native proteins are destabilising and, consequently, mutations that lead to a more favourable enzyme activity are likely to decrease the stability of the protein (2, 4). Compensatory mutations are then needed to restore global stability. These processes are referred to as stability-activity trade-offs (5-7). Furthermore, proteins with higher stability confer greater evolvability, since there is more scope to accept destabilising yet functionally beneficial changes (8). Whereas such stability activity trade-offs are well attested in laboratory experiments, it remains unclear as to how strong a signal these particular physical constraints would leave in a naturally, and slowly, evolving population where there are many potentially competing evolutionary pressures and considerable neutral drift (9).

The probability that a new mutation becomes fixed in a species is determined by the relative strengths of genetic drift and natural selection. While the rate of fixation is assumed to be constant under neutral evolution, it is decelerated by negative selection, which tends to remove deleterious mutations, or accelerated by positive selection, under which favourable mutations, e.g., those enabling adaptation of the protein following environmental changes, tend to be retained. A well-known case of adaptation under positive selection is the ribulose-1,5-bisphosphate carboxylase (RubisCO, EC: 4.1.1.39), the enzyme responsible for fixation of CO_2 to ribulose-1,5-bisphosphate in the Calvin-Benson cycle. It is the most abundant protein on earth and represents up to 30% of all soluble proteins in plants. However, this abundant enzyme also has a very low turnover of <10 per second. RubisCO can catalyse reactions with both CO_2 and O_2, and the catalytic rate for CO_2 fixation is negatively correlated with CO_2/O_2 specificity (10). The fixation of O_2 initiates the photorespiratory cycle, which uses ATP to regenerate CO_2, resulting in both energy loss and a net loss of fixed CO_2. Because these losses are disadvantageous, there is selection for increased affinity for CO_2, compared to O_2 and thus for low catalytic rates (10). The dual affinity seems inevitable, as both CO_2 and O_2 can attack the carbanion form of ribulose-1,5-bisphosphate produced during the reaction (11).

Several lineages of flowering plants (angiosperms) have evolved mechanisms that diminish photorespiration by concentrating CO_2 before its fixation by RubisCO. These mechanisms operate in various pathways such as crassulacean acid metabolism (CAM) and C_4 photosynthesis. While CAM is mainly an adaptation to water stress, C_4 photosynthesis is advantageous in all conditions that promote photorespiration, such as warm, open, dry, saline or some aquatic environments. In C_4 plants, atmospheric CO_2 is initially incorporated into small organic compounds by a series of enzymes beginning

with carbonic anhydrase and phosphoenolpyruvate carboxylase, a system without affinity for O_2. These compounds are transported to the specialised compartments (most often distinct cells) where RubisCO is located. The various pathways lead to the formation of malate or oxaloacetate, which are decarboxylated to yield CO_2 and pyruvate or phosphoenolpyruvate (12); producing an up to 10-fold increase of CO_2 concentration in the proximity of RubisCO. Despite its relative complexity, the C_4 trait has evolved more than 62 times in different groups of flowering plants (13), including up to 24 times in grasses alone (14).

The turnover rate of RubisCO is positively correlated with its CO_2 affinity [$Km(CO_2)$] and negatively correlated with the CO_2/O_2 specificity ratio of the enzyme (10, 15, 16). The high concentration of CO_2 at the site of RubisCO in C_4 plants allows a lower specificity ratio of CO_2/O_2 and therefore an increase in turnover rate and, thus, efficiency (17, 18). Experimental studies of RubisCOs from very closely related C_3 and C_4 species within the *Flaveria*, *Atriplex* and *Neurachne* genera showed that very few changes may be necessary to modify enzymatic properties in response to the modification of the metabolic context (19, 20). Indeed, in the *Flaveria* context, a single mutation (M309I) has been identified as key in modifying specificity and increasing turnover (21); it remains unclear as to how this observation applies to a wider range of plants and what are the contributions of other observed mutations to adaptation. Comparative sequence analysis of a broader range of plant species does suggest that in general adaptation of RubisCO to C_4 metabolism involves a larger number of amino acid changes found to be under positive selection (19, 20). Here, we investigate the role of mutations in the adaptation of a large group of plants, focusing in particular on the constraints imposed by stability requirements, which have been previously shown to be important in the directed-evolution of enzymes.

In this study, we focused on the RubisCO of the monocot lineage, which is one of the major groups of flowering plants, and contains both C_3 and C_4 species. Its diversification probably started 120 Mya and the emergence of distinct C_4 species has occurred over the past 40 million years. We took advantage of the convergent nature of the evolution of C_4 photosynthetic pathways and the resulting common changes in the selective pressures on RubisCO, to investigate the structural factors influencing the evolvability of novel enzymatic properties. Our combined phylogenetic framework and structural analyses, allowed an *in silico* reconstruction of the ancestral sequences and 3D structures of the large subunit within the RubisCO complex. Our investigations have enabled the inference of the mutational paths linked to the adaptation to C_4 photosynthesis in the monocots.

This work shows that the evolutionary adaptation of the RubisCO enzyme is mediated by stability-activity trade-offs with many stabilising mutations apparently being fixed simply to allow functionally necessary destabilising mutations to be tolerated. The enzyme has used multiple paths to adapt to new environmental conditions with no single mutation present in more than two-thirds of C_4 species. The paths are structurally diverse, including the mutation of residues close to and remote from the active site. The location of many of the positively selected mutations implies that allosteric modulation of structure at the active site and (possibly cooperative) dynamics of domain and subunit movements are keys to adaptation.

Results

OVERVIEW. The RubisCO of plants, as exemplified by the enzyme from the rice *Oryza sativa*, is a hexadecamer composed of eight large subunits [encoded by *rbcL*] and eight small subunits [encoded by *rbcS*] (L_8S_8) (Fig. 1). The following analysis is necessarily limited to the catalytic *rbcL*, since insufficient sequences of monocot *rbcS* genes are available to reliably reconstruct ancestral sequences (see SI text for additional remarks). We divided our analysis of *rbcL* into two parts. First, the stability landscape was investigated by computationally scanning all possible mutations of the *O. sativa* RubisCO, which is a C_3 form (no structure of a RubisCO from a C_4 plant has been determined). Second, ancestral mutations that occurred during the adaptation of RubisCO in monocots were identified, selective pressures were estimated, and the effect of the positively selected mutations on stability, and their locations in the 3D structure, were examined.

STABILITY LANDSCAPE OF ALL POSSIBLE MUTATIONS. The stability effect of all possible mutations of each residue of the quaternary complex was estimated using FoldX. The wild-type amino acid at each position of the *O. sativa rbcL* was mutated (in all eight chains) to each of the nineteen other possibilities. This energetic landscape highlights positions that are mutation-tolerant (Fig. 2A). For convenience, if we categorise the calculated effects of mutations in proportion to the known accuracy of FoldX predictions (see Methods), then most possible mutations (5,007 of 8,436 = 59.4%) are found to be highly destabilising ($\Delta\Delta G_{fold}$ per chain > +1.84 kcal mol^{-1}) and 3,335 mutations (39.5%) have a moderate effect (-1.84<$\Delta\Delta G_{fold}$<+1.84 kcal mol^{-1}). Ninety-four mutations (1.1%) can strongly stabilise the structure, but only in a smaller number of positions (35/444) most of which are in the active site. It has previously been observed that residues close to an active site are often intrinsically destabilising, because their great functional utility is traded against stability (22, 23). Finally, less than one quarter of the positions (103/444) were found to be actually mutated in our monocot sequence dataset with only 2 to 4 alternative residues observed at each position (Fig. 2B).

ANALYSIS OF MUTATIONS OCCURRING DURING EVOLUTION AND THEIR EFFECT ON STABILITY. The monocot dataset exhibits a >95% pairwise sequence identity at the protein sequence level and no alignment gaps. This high-level of conservation, together with the previously determined, highly-resolved, phylogenetic tree (24), allowed the reconstruction, with high confidence, of the ancestral sequences (each comprising 444 mutable amino acids) for each of the 239 ancestral (internal) nodes of the monocot tree. The average posterior probability (PP) for the reconstruction of all 106,116 residue positions in these sequences is 99.9% and only 16 of these predictions have a PP <80%. The reconstructed sequences were used to infer 3D models of each of the ancestral octomers (L_8) by homology, with high confidence. The stability effect of ancestral mutations was then estimated, using FoldX to make mutations in the homology model of the appropriate ancestral octomer.

Global analysis of the stability impact of ancestral mutations. The distribution of the $\Delta\Delta G_{fold}$ values of all possible mutations of *Oryza sativa* RubisCO (Fig. 3) is unimodal and strongly skewed toward positive values, and most possible mutations would be destabilising. In contrast, the global distribution of $\Delta\Delta G_{fold}$ values of the ancestral mutations follows a bimodal distribution with a high peak near zero and a smaller peak

at +0.88 kcal mol^{-1} (Fig. 3). Ancestral mutations are rarely strongly stabilising or destabilising (of the 751 in total, 6 are lower than -1.84 kcal mol^{-1} and 58 are higher than +1.84). The vast majority of ancestral mutations (91.5%) are rather evenly distributed about zero in the -1.84 to +1.84 kcal mol^{-1} range, consistent with the hypothesis that maintenance of the stability of the protein is a strong constraint on evolution.

Stability effects and selective pressures. Among the sites which underwent mutation according to the ancestral reconstruction, two groups can be distinguished: those sites evolving under neutral evolution or negative selection and those sites under positive selection between C_3 and C_4 forms. Previous analyses have identified sets of 1, 2, 3, 7, 11 or 12 positively selected sites with discrepancies and overlap between the sets (Table S1). The 18 sites identified here encompass nearly all of those previously identified and three new sites. The sensitivity of the current analysis resolves many earlier discrepancies [Table S1 and (24-26)]. Ancestral mutations were classified according to their evolutionary pressures (Fig. 4A), as defined by the TDG09 algorithm (27). Independently from the distinction between types of selection, mutations were also classified into three groups following the photosynthetic types of their ancestor and descendant as $C_3 \rightarrow C_3$, $C_3 \rightarrow C_4$ and $C_4 \rightarrow C_4$ (the change $C_4 \rightarrow C_3$ has not been seen and detailed comparative analyses show that, if it has occurred, it must be very rare (28)).

On $C_3 \rightarrow C_3$ branches, the distribution of stability effects follows a normal distribution, with a peak of stability-neutral mutations (Fig. 4B, left) that preserve the global stability of the structure. In contrast, the C3→C4 branches present significantly more destabilising mutations (permutation test, $p=0.0080$) (Fig. 4B, centre), which correspond to the second peak (+0.88 kcal mol^{-1}) in the global distribution (Fig. 3). This tendency for destabilising mutations to occur at the $C_3 \rightarrow C_4$ transition is also apparent in a 'timeline' of cumulative mutational stability changes in the ancestral sequences (Fig. 5). In $C_4 \rightarrow C_4$ branches, a large fraction of destabilising mutations is still observed, but there is a significantly greater proportion of mutations with a stabilising effect compared to other branches ($p<0.0001$) (Fig. 4B, right). The timeline also shows that there is a large proportion of stabilising mutations immediately following the $C_3 \rightarrow C_4$ transition (Fig. 5A) and that the preponderance of stabilising over destabilising mutations means that the loss of stability at the transition is largely recovered within the subsequent three branches (Fig. 5B). Furthermore, considering the cumulative contributions to stability of all stabilising mutations and all destabilising mutations separately, shows that stabilisation due to all stabilising mutations is accumulated more quickly in the branch following the $C_3 \rightarrow C_4$ transition than at any other time (Fig. 5C).

Stability effects and location on the 3D structure. Ancestral mutations were grouped according to their position in the 3D structure of the hexadecamer (Fig. 4C and 4D) following the interface definitions in (29). The stability effects of mutations within the core of the large subunit or the LL1 interfaces within dimers (e.g. L_AL_B) follow an approximately normal distribution. In contrast, although small in number, mutations of residues at the LL2 (e.g. L_AL_H) interface between dimers and at the LS interface between large and small subunits have some tendency to be highly destabilising ($p=0.0318$ and $p=0.0053$ respectively). The proportion of mutations at interfaces between large subunits is significantly greater in the $C_4 \rightarrow C_4$ branches ($p<0.0002$), suggesting that the modification of subunit interactions is important for C_4 optimisation (Fig. 4D).

POSITIVELY SELECTED SITES IN THE TRANSITION TO C_4. At the $C_3 \rightarrow C_4$ transitions, three positively selected mutations with a destabilising effect are especially frequent: A328S, A281S and L270I (Table S1). The A328S mutation and a positively selected, but less frequent, V326I mutation lie either side of H327, which coordinates the P5 phosphate of the substrate in the closed state of the enzyme. Furthermore these two residues are at the base of the active site loop (loop 6 in residues 328-337) that carries the catalytic lysine K334 and undergoes a disorder-order transition upon the binding of both substrates. The replacement of hydrophobic A328 in the C_3 form with a polar serine in C_4 forms is destabilising as it disrupts the packing of the base of loop 6 against α-helix 6 (running from residues 338-350). This destabilisation could directly alter the catalytic parameters by allowing more flexibility in loop 6 thus affecting the opening and closing of the active site (16). Extensive studies of this loop region in algal and cyanobacterial RubisCOs have shown that catalytic parameters are sensitive to its modification even if the mutated residues have no direct interaction with substrates (reviewed in Parry*, et al.* (30)). L270I is located directly beneath H298 which interacts with the P5 phosphate in the pre-activated state. Replacement of V326 and L270 will also lead to packing changes that could alter the spatial disposition of the phosphate binding histidines. Site 281 is in the core of the C-terminal domain and its potential to affect activity is not obvious. However, A281 packs against S321 and G322 at the end of the strand which leads to loop 6, and destabilisation of this interaction may have a long-range effect on the dynamics at the active site.

Another frequently positively selected mutation, M309I, also identified in some previous phylogenetic studies (20, 24), lies at the interface of the two C-terminal domains within a dimer, and also close to the junction between N and C-terminal domains within each subunit. This mutation has been demonstrated to act as a catalytic switch between C_3-like and C_4-like properties (i.e. decreasing specificity for CO_2 over O_2 and increasing the turnover) in *Flaveria* species and in chimeric enzymes consisting of large subunits from *Flaveria* and tobacco small subunits (21). However, isoleucine is present in only half of all the C_4 forms. Interestingly, sites 309 and 328 are evolutionary coupled (Table S2). Also under strong positive selection in $C_3 \rightarrow C_4$ (and $C_4 \rightarrow C_4$ branches) is the mutation V101I. The addition of one carbon to this side chain is anticipated to shift the 2nd α-helix in the N-terminal domain toward the active site. Directly on the opposite side of this helix is glutamate-60, which forms a salt-bridge with the catalytic K334 in the closed activated state of the enzyme. Any movement of the α-helix could affect the geometry of the CO_2-bound and transition states of the reaction.

Several of the positively selected mutations found in $C_3 \rightarrow C_4$ branches are also present in $C_4 \rightarrow C_4$ branches, (i.e. V101I, L270I, M309I, A328S). Additionally, three mutations on α-helix 8, the final element of secondary structure of the N-terminal domain of the large subunit, are positively selected in this type of branch: P142A/T, T143A (also strongly selected in $C_3 \rightarrow C_4$ branches) and S145A. This helix forms the symmetric interface between the N-terminal domains of large subunits on neighbouring dimers (at the LL2 interfaces e.g. L_AL_H). At each interface, the threonine and proline from each helix are intercalated (Fig. S3). Structural superposition of the open and closed forms of rice RubisCO suggests that an asymmetric movement of this helix between open and closed states of the upper active site, such as might occur on ligand binding or product release, will be transmitted to the neighbouring active site at its lower left, potentially leading to a preference for the lower site to be closed while the top is open and vice versa (Fig. S3).

Discussion

Diversification of RubisCO on an island of stability. Throughout their evolutionary histories, RubisCO genes have faced significant changes, both internal and external to the organism, which have altered the physiologically optimal properties of RubisCO and thus the selective pressures on its evolution (10). In our example of rice RubisCO, residues at nearly all sites contribute favourably to stability and most putative mutations would lead to destabilisation (Fig 2A). The change in stability that RubisCO can withstand without dysfunction has yet to be established experimentally, but the computed stability effects of mutations that have become fixed in some species are largely confined to a narrow range near zero (Fig. 3). This small amplitude of the effects of mutation observed in nature suggests that RubisCO evolves within a small island of stability (3, 5).

The adaptations of RubisCO to C_4 photosynthesis in numerous plant lineages can be regarded as the result of natural experiments in evolution with a common (or at least similar) outcome of reduced CO_2/O_2 specificity and an increase in turnover (10). The availability of many sequences of closely related C_3 and C_4 species has enabled those branches of the phylogenetic tree associated with gain of C_4 function, and thus increased activity, to be identified reliably by parsimony. Ancestral reconstructions of these lineages allow the mutational pathways of evolution to be recovered with high confidence. Because structures of representative proteins are available, the stability effects of mutations at each point on these pathways can also be estimated. We have found that despite the positively selected mutations forming only a small proportion of the total, overall the changes in stability during evolution display features strongly reminiscent of those previously identified as significant in laboratory experiments by site-directed mutagenesis and directed evolution (9).

Destabilising mutations are more frequently fixed in C_4 lineages. In those evolutionary branches which undergo a functional change ($C_3 \rightarrow C_4$), adaptation is preceded by a long mutational sequence in which neutral to slightly stabilising capacitive mutations dominate, i.e. which create the capacity for the protein to tolerate the destabilisation required for new function (Fig. 5B). A variety of often destabilising mutations occur precisely at the transition to C_4, and these are immediately followed by compensatory stabilising mutations (Fig. 5 and Fig. S2).

Except for cases in which folding is coupled to substrate binding, there is no *a priori* expectation of a direct physical connection between stability and activity. That similar trade-offs between activity and stability are consistently found in both directed and natural evolution, argues that an indirect connection necessarily arises from the tension between selection for optimal stability and selection for activity from a shared pool of possible mutations.

Modulation of conformational change appears to be key to the adaptation of RubisCO. Adaptive mutations occur in several distinct parts of the RubisCO structure. None are in direct contact with the substrates, however, a small number of "second shell" mutations (i.e. residues in contact with active site residues) are strongly positively selected. These tend to be destabilising and, on the basis of structural context and earlier mutational studies of algal RubisCOs, are inferred to modify the active site loop dynamics or position of residues at the P5 and O_2/CO_2 binding sites. Whereas adaptive

mutations 10-20Å from active sites have occasionally been identified in other enzymes (31), in RubisCO these form the majority of positively selected sites that distinguish C_3 and C_4 species. Experiments with RubisCO from the green alga *Chlamydomonas reinhardii* previously suggested an implication of the interfaces between large and small subunits in the modulation of catalytic rates (32). The analysis here increases the number of known inter-subunit sites (Table S2) and demonstrates a link with the C_3-C_4 transitions in flowering plants. Those mutations near the dimer or N-and C-terminal domain interfaces within each large subunit likely affect the substantial relative movements of the domains upon substrate binding. While one of these residue changes (M309I) has previously been shown to switch the enzyme to C_4-like properties in plants (21), it is clear that this change is not essential and there are other mutational routes to equivalent functional changes.

Altered cooperativity may have an adaptive role in some species. Negative cooperativity has been reported for the binding of the transition state analogue 2-carboxyarabinitol bisphosphate to the active site of the C_3 RubisCO from spinach (33). Kinetic data fit a model of rapid binding to one half of the active sites accompanied by the slower binding to the remainder (34). While it has proved difficult to generalise these observations to other species (possibly because of the stringent demands for pure and active protein in such experiments and because weak negative cooperativity is also intrinsically difficult to unambiguously identify in standard turnover kinetics) they naturally led to a postulated enzymatic mechanism whereby binding of substrates to one site of each dimer reduces binding at the other (34). Crystallographic studies have not been able to directly address this issue as they produce symmetric structures, either apo or fully saturated (16). The observation of positive selection on mutations in the interface between the N-terminal domains of neighbouring dimers suggests a different mechanism of cooperativity. Comparison of hybrid structures of apo and holo forms of RubisCO suggests that conformational changes at an active site in the ring of active sites at the top of the oligomer are coupled to the lower site in the dimer to its left. The mutations occurring during the C_3 to C_4 transitions diminish this coupling and would relieve any negative cooperativity between the upper and lower sites thus enhancing turnover. The identified positive selection suggests that these mutations play a role in the adaptation of some C_4 species. Consequently, these mutations and the possibility of a role for cooperativity in RubisCO warrant renewed experimental investigation.

Conclusions

The mutational landscape of RubisCO is strongly constrained by the need to maintain overall stability. This limits the adaptation of RubisCO to novel environmental contexts, to those amino acid changes that can modify the catalytic efficiency without dramatic effect on the overall folding stability. Following the repeated origins of C_4 photosynthesis in flowering plants, a number of amino acid mutations of RubisCO were preferentially kept by natural selection. These include residues that might modify the geometry of the active site, as well as a substantial number of sites at the interface between domains and subunits, which probably alter the properties of the enzyme via modification of the dynamics of conformational change or alteration of the cooperativity between catalytic subunits. It is clear that a substantial proportion of the mutations necessary for C_4 adaptation are themselves destabilising. Evolution accommodates such destabilising

functional adaptations thanks to the previous accumulation of stabilising capacitive mutations and by subsequently fixing stabilising compensating mutations.

Methods

The multiple sequence alignment of genes for RubisCO large subunit (*rbcL*) and its associated phylogenetic tree are from Christin*, et al.* (24). The highest resolution (1.35Å) structure of RubisCO currently available, from the C_3 grass *Oryza sativa* (35), was used as the basis for structural analyses. The complete biological unit (L8S8) was directly downloaded from the PDBePISA website (36). The PDB structure file for the large subunit contains coordinates for residues 11 to 475 (465 residues). This structure was used as a template for the homology modelling of 3D octomeric structures (L8) of each ancestral *rbcL* sequence. The modelling was done with Modeller 9.9 (37). For each sequence, 100 models were built and the model with the lowest energy (based on its DOPE score) was used in further analyses. Using FoldX 3b5.1 (38), the energies for the wild-type ($\Delta G_{fold,wt}$) and mutant ($\Delta G_{fold,mut}$) protein were computed to give the stability change $\Delta\Delta G_{fold} = \Delta G_{fold,mut} - \Delta G_{fold,wt}$. The standard deviation in FoldX is 0.46 kcal mol^{-1} (38) and we used this value to bin the $\Delta\Delta G_{fold}$ values into seven categories. Additional FoldX restraints were applied to the conserved active site to avoid the potential for artefacts arising from unparameterised ligands. The inference of ancestral sequences was performed under maximum likelihood as implemented in CodeML (39). Sites under positive selection between C_3 and C_4 forms were identified by the TDG09 algorithm (27), which performs a likelihood ratio test to assess if the evolutionary rate at a particular position is similar or different between C_3 and C_4 lineages. The $\Delta\Delta G_{fold}$ due to each mutation on each branch was then mapped onto the phylogenetic tree (Fig. S2). Detailed methods are given in SI.

Acknowledgements

This study benefited from use of the UCL *Legion* High Performance Computing Facility (Legion@UCL). RAS acknowledges funding from the Fondation du 450ème anniversaire de l'Université de Lausanne and Swiss National Science Foundation grants 132476 and 136477. PAC is funded by the Marie Curie International Outgoing Fellowship 252568.

References

1. Romero PA & Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 10(12):866-876.
2. DePristo MA, Weinreich DM, & Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 6(9):678-687.
3. Taverna DM & Goldstein RA (2002) Why are proteins marginally stable? *Proteins* 46(1):105-109.
4. Tokuriki N & Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 19(5):596-604.
5. Tokuriki N, Stricher F, Serrano L, & Tawfik DS (2008) How protein stability and new functions trade off. *PLoS Comput Biol* 4(2):e1000002.
6. Soskine M & Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11(8):572-582.
7. Wang X, Minasov G, & Shoichet BK (2002) Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol* 320(1):85-95.
8. Bloom JD, Labthavikul ST, Otey CR, & Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* 103(15):5869-5874.
9. Bloom JD & Arnold FH (2009) In the light of directed evolution: pathways of adaptive protein evolution. *Proc Natl Acad Sci U S A* 106 Suppl 1:9995-10000.
10. Tcherkez GG, Farquhar GD, & Andrews TJ (2006) Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proc Natl Acad Sci U S A* 103(19):7246-7251.
11. Lorimer GH & Andrews TJ (1973) Plant Photorespiration[mdash]An Inevitable Consequence of the Existence of Atmospheric Oxygen. *Nature* 243(5406):359-360.
12. Sage RF, Sage TL, & Kocacinar F (2012) Photorespiration and the evolution of C4 photosynthesis. *Annu Rev Plant Biol* 63:19-47.
13. Sage RF, Christin PA, & Edwards EJ (2011) The C-4 plant lineages of planet Earth. *J Exp Bot* 62(9):3155-3169.
14. Grass Phylogeny Working Group II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytol* 193(2):304-312.
15. Savir Y, Noor E, Milo R, & Tlusty T (2010) Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc Natl Acad Sci U S A* 107(8):3475-3480.
16. Andersson I & Backlund A (2008) Structure and function of Rubisco. *Plant Physiol Biochem* 46(3):275-291.
17. Young JN, Rickaby RE, Kapralov MV, & Filatov DA (2012) Adaptive signals in algal Rubisco reveal a history of ancient atmospheric carbon dioxide. *Philos Trans R Soc Lond B Biol Sci* 367(1588):483-492.
18. Sage RF (2002) Variation in the k(cat) of Rubisco in C-3 and C-4 plants and some implications for photosynthetic performance at high and low temperature. *J Exp Bot* 53(369):609-620.
19. Hudson GS*, et al.* (1990) Comparisons of rbcL genes for the large subunit of ribulose-bisphosphate carboxylase from closely related C3 and C4 plant species. *J Biol Chem* 265(2):808-814.

20. Kapralov MV, Kubien DS, Andersson I, & Filatov DA (2011) Changes in Rubisco kinetics during the evolution of C4 photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme. *Mol Biol Evol* 28(4):1491-1503.
21. Whitney SM*, et al.* (2011) Isoleucine 309 acts as a C4 catalytic switch that increases ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in Flaveria. *Proc Natl Acad Sci U S A* 108(35):14688-14693.
22. Dessailly BH, Lensink MF, & Wodak SJ (2007) Relating destabilizing regions to known functional sites in proteins. *BMC Bioinformatics* 8:141.
23. Beadle BM & Shoichet BK (2002) Structural bases of stability-function tradeoffs in enzymes. *J Mol Biol* 321(2):285-296.
24. Christin PA*, et al.* (2008) Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. *Mol Biol Evol* 25(11):2361-2368.
25. Wang M, Kapralov MV, & Anisimova M (2011) Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco. *BMC Evol Biol* 11:266.
26. Kapralov MV & Filatov DA (2007) Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol Biol* 7:73.
27. Tamuri AU, Dos Reis M, Hay AJ, & Goldstein RA (2009) Identifying changes in selective constraints: host shifts in influenza. *PLoS Comput Biol* 5(11):e1000564.
28. Christin PA, Freckleton RP, & Osborne CP (2010) Can phylogenetics identify C(4) origins and reversals? *Trends Ecol Evol* 25(7):403-409.
29. van Lun M, van der Spoel D, & Andersson I (2011) Subunit interface dynamics in hexadecameric rubisco. *J Mol Biol* 411(5):1083-1098.
30. Parry MA, Andralojc PJ, Mitchell RA, Madgwick PJ, & Keys AJ (2003) Manipulation of Rubisco: the amount, activity, function and regulation. *J Exp Bot* 54(386):1321-1333.
31. Thomas VL, McReynolds AC, & Shoichet BK (2010) Structural bases for stability-function tradeoffs in antibiotic resistance. *J Mol Biol* 396(1):47-59.
32. Spreitzer RJ, Peddi SR, & Satagopan S (2005) Phylogenetic engineering at an interface between large and small subunits imparts land-plant kinetic properties to algal Rubisco. *Proc Natl Acad Sci U S A* 102(47):17225-17230.
33. Johal S, Partridge BE, & Chollet R (1985) Structural characterization and the determination of negative cooperativity in the tight binding of 2-carboxyarabinitol bisphosphate to higher plant ribulose bisphosphate carboxylase. *J Biol Chem* 260(17):9894-9904.
34. Zhu G & Jensen RG (1990) Status of the substrate binding sites of ribulose bisphosphate carboxylase as determined with 2-C-carboxyarabinitol 1,5-bisphosphate. *Plant Physiol* 93(1):244-249.
35. Matsumura H*, et al.* (2012) Crystal Structure of Rice Rubisco and Implications for Activation Induced by Positive Effectors NADPH and 6-Phosphogluconate. *J Mol Biol*.
36. Krissinel E & Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372(3):774-797.
37. Sali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779-815.
38. Schymkowitz J*, et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res* 33(Web Server issue):W382-388.
39. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586-1591.

Figure Legends

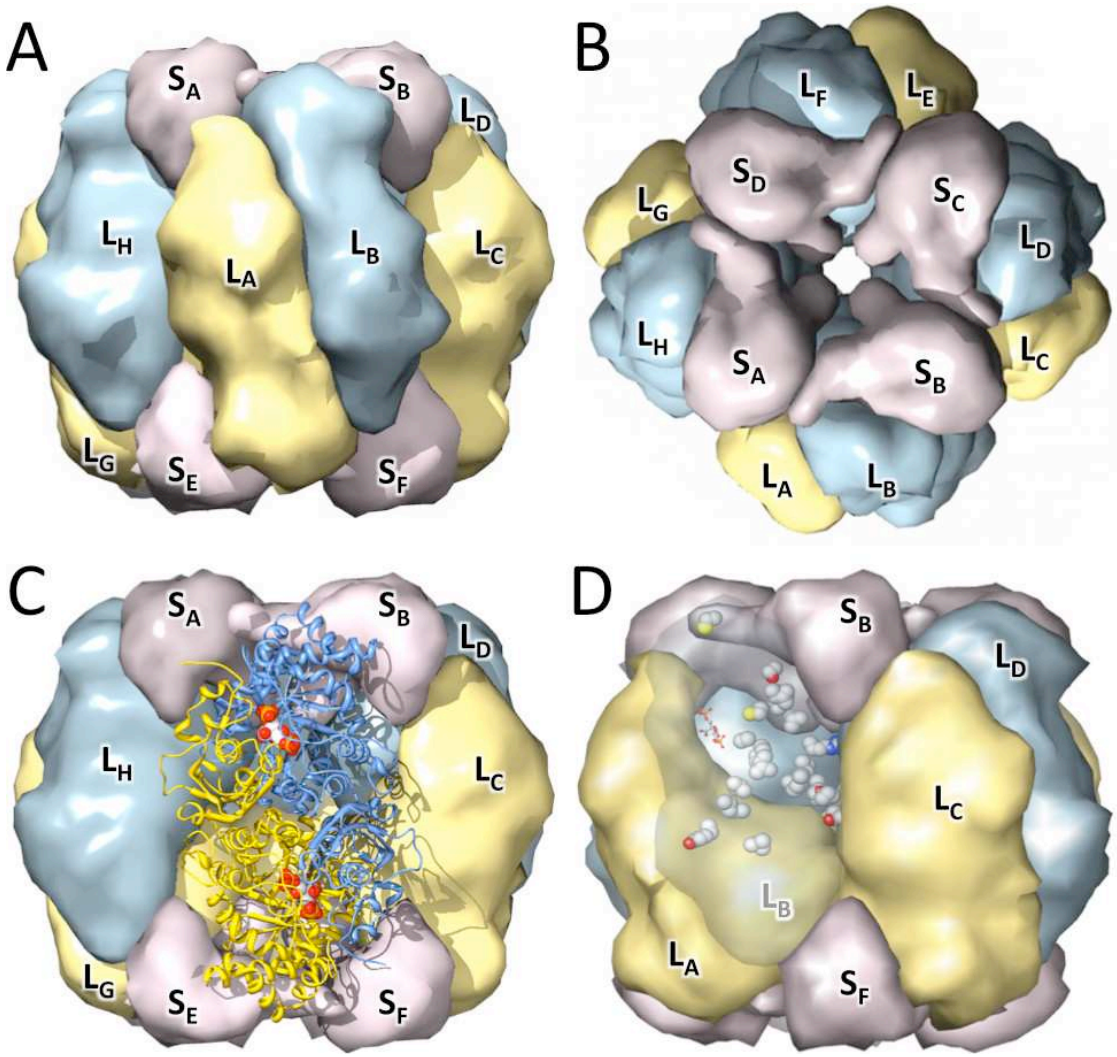

Fig. 1. The RubisCO hexadecamer structure. Pairs of large subunits (blue and yellow) form dimers with an extensive interface, four of these dimers form an octomeric ring. The inter-dimer interfaces are comparatively small and the overall structure is stabilised by the binding of eight small subunits (lavender) that bridge dimers. (A) and (B) present surface views from side and top, respectively. (C) The two chains forming the L_AL_B dimer are shown in ribbon form. Each dimer forms two active sites, the upper site here being between the N-terminal domain of L_A and the C-terminal domain of L_B. Each site undergoes an open to closed structural transition upon substrate binding. The reaction intermediate analogue 2-carboxyarabinitol-1,5-bisphosphate is shown bound at each site in this structure (PDB:1WDD). The larger C-terminal domain contributes most residues to each active site, but the N-terminal domain is critical for positioning the CO_2 or O_2 molecule. (D) Atoms of residues under positive selection in the large subunit (L_B) are shown as spheres. These residues are frequently close to subunit interfaces.

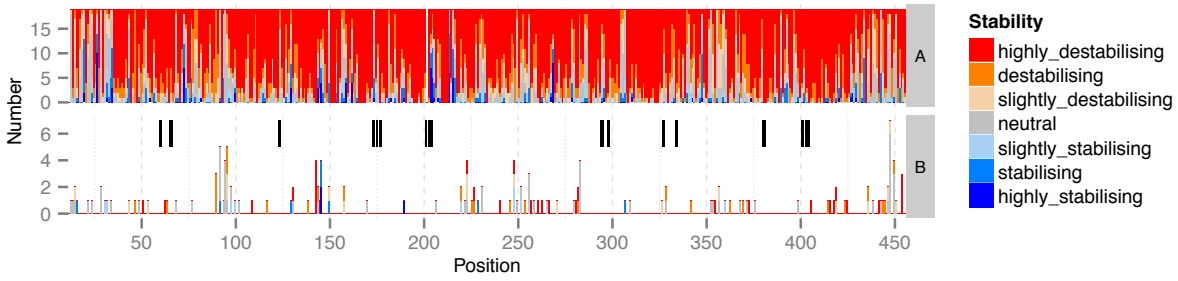

Fig. 2. Effect of mutations on protein stability. (A) Stability landscape of the large subunit (*rbcL*). All 19 possible mutations at each position observed in the *O. sativa* structure [positions 12 to 456] are coloured on a vertical bar in terms of their stability relative to the native residue. Residues that are part of the active site are indicated by a black bar. The thresholds for $\Delta\Delta G_{fold}$ in kcal mol^{-1} are: highly stabilising (< -1.84), stabilising (-1.84 to -0.92), slightly stabilising (-0.92 to -0.46), neutral (-0.46 to +0.46), slightly destabilising (+0.46 to +0.92), destabilising (+0.92 to +1.84), highly destabilising (> +1.84). Positions where the vertical bar is substantially grey or blue are predicted to be tolerant of mutation, and where largely red are intolerant. Highly destabilising mutations are very unlikely to occur in nature. (B) Stability effect of observed mutations at each position, relative to the O. sativa rbcL sequence. Within the monocot species, 105 positions of the 444 aligned residues of the peptide chain have alternate amino-acids. The overwhelming majority of observed mutations (79.5%) have modest stability changes in the range -1.84 to +1.84 kcal mol^{-1}.

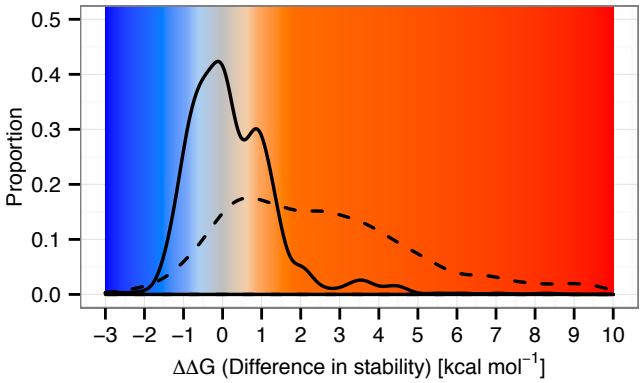

Fig. 3. Distribution of stability effects of possible mutations and those occurring during evolution. The distribution of stability changes arising from mutations observed in the evolutionary history of the reconstructed ancestral sequences (solid line) stands in contrast to that of all possible simulated mutations (dashed line). Both distributions have their largest peak close to a ΔΔG of zero. The observed mutations have an excess of slightly stabilising observed mutations, and also a distinct peak of slightly destabilising and destabilising values centred at +0.88 kcal mol^{-1}. The majority of possible mutations are highly destabilising, and rarely occur during evolution. The probability distributions shown here are obtained by kernel smoothing of the original data (Fig. S4).

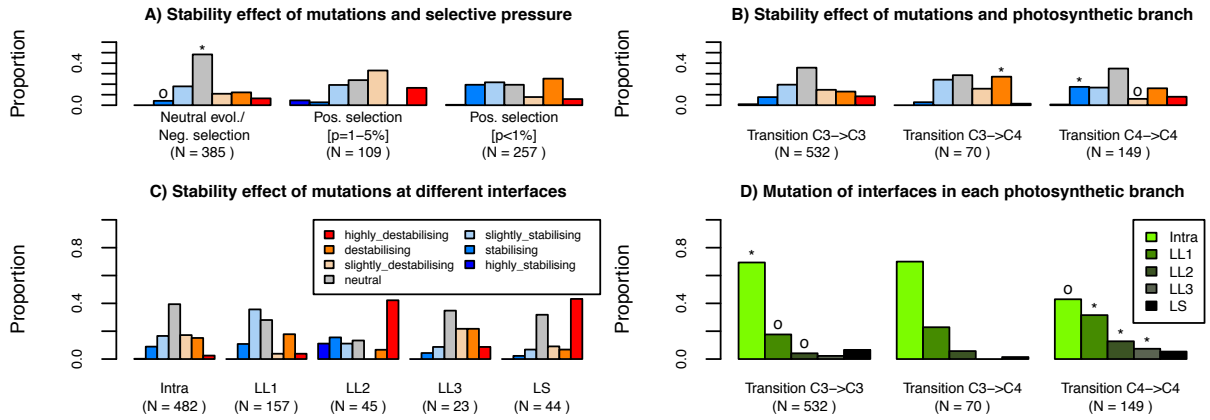

Fig. 4. Stability effect and location of ancestral mutations. The 751 mutations occurring during evolution are separated in A) by their selection constraints: negative selection or neutral evolution ($p>0.05$ from TDG09 after false-discovery rate correction), positive selection ($0.01<p<0.05$) and strong evidence of positive selection ($p<0.01$) and binned according to their stability effect. B) Mutations are separated into their branch type ($C_3 \rightarrow C_3$, $C_3 \rightarrow C_4$ or $C_4 \rightarrow C_4$) and binned by their stability effect. C) Mutations are classified following the subunit interface definitions in (29) and Fig. S1: "Intra" are in contact only with other residues of the same large subunit, LL1 residues are in contact with the other large subunit of the same dimer (e.g. the L_AL_B interface), LL2 and LL3 contact a large subunit of another dimer (e.g., respectively, L_BL_C and L_BL_D), and LS are all residues in contact only with any of the small subunits. D) Mutations are separated into their branch type and binned into their contact interfaces. Categories are highlighted by a (*) when enriched or a (o) when depleted.

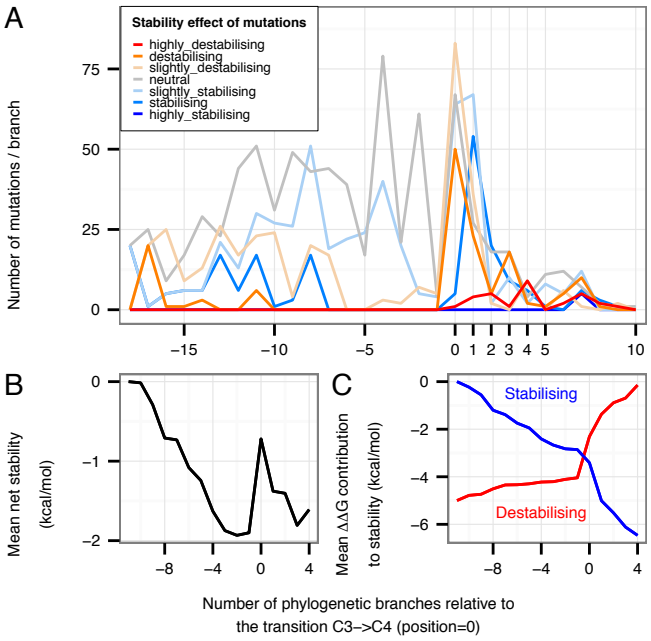

Fig. 5: Changes in stability through evolution. (A) Frequency of mutations in each category of stability against their evolutionary branch positions relative to the $C_3 \rightarrow C_4$ transition. There is a long period in which slightly stabilising mutations are accumulated prior to the transition in which a substantial number of destabilising and slightly destabilising mutations occur. In the branch following the transition there is a peak of apparently compensatory stabilising or slightly stabilising mutations. Stability categories as in Fig. 2. (B) Cumulative mean net change in stability in the neighbourhood of the $C_3 \rightarrow C_4$ transition. (C) The corresponding cumulative mean contributions to stability of all stabilising and all destabilising mutations (the latter is offset by -5 kcal/mol to aid comparison).